This report is automatically generated with the R package **knitr** (version `1.24`) .

~~~
---
title: "Final Project; Categorical Data Analysis"
author: "Abby Smith, Miruna Barnoschi, Martha Eichlersmith"
date: "2019-12-05"
output:
  pdf_document:
    fig_caption: yes
    number_sections: yes
header-includes:
- \usepackage{color}
- \usepackage{mathtools}
- \usepackage{bbm} #for mathbb for numbers
- \usepackage{amsbsy}
- \usepackage{caption}
- \usepackage{booktabs}
- \usepackage{geometry}
- \usepackage{float} #to hold things in place
- \usepackage{lastpage}
- \usepackage{fancyhdr}
- \usepackage{graphicx}
- \newcommand{\indep}{\rotatebox[origin=c]{90}{$\models$}}
- \pagestyle{fancy}
- \floatplacement{figure}{H}
- \fancyhf{}
- \fancyhead[L]{STAT 455 Fall 2019 \\ Final Project}
- \fancyhead[R]{Abby, Miruna, Martha \\ Page \thepage\ of\ \pageref*{LastPage}}
- \setlength{\headheight}{22.5pt} #to remove \fancyhead error for head height
geometry: left=0.75in,right=0.75in,top=1.1in,bottom=1in
---
```{r setup, include=F}
knitr::opts_chunk$set(message=F, warning=F)

require(tidyverse)
require(janitor)
require(rgdal)
require(sp)
require(sf)
require(raster)
require(Rcapture)
require(dga)
require(rgeos)
require(MASS)
library(bookdown) #for fig captions
library(kableExtra)

knitr::opts_chunk$set(fig.width = 10, fig.height = 4)
knitr::opts_chunk$set(echo=FALSE)
knitr::opts_chunk$set(fig.pos='H')

~~~

# Estimates of Killing in Casanare

Casanare is a large, rural state in Colombia that includes 19 municipalities and a population of almost 300,000 inhabitants. Located in the foothills of

So how many people have been killed or disappeared? We review the Human Rights Data Analysis Group (HRDAG)’s
reporting on this issue of population estimation (Guzman et al., 2007). In this study, the authors used information
about victims of killings and disappearances provided by 13 datasets. We have 15 datasets of deaths and disappearances
in Casanare from 1998-2007. Any accounting of lethal violence will be incorrect if we assume that any one dataset or
combination of datasets contains a comprehensive count of violent acts and disappearances. Registries of violent acts
kept by governmental and non-governmental institutions contain some, but not all, of the records of lethal violence.
Organizations collecting this data may only have access to certain subsets of a population or geographic areas.

The datasets come from state agencies – including government, security, forensic and judicial bodies – and from civil society organizations. Across thes

\newpage
# Multiple Systems Estimation (MSE), Capture Re-Capture

The goal is to estimate the overall population of victims by first estimating the victims who are not captured by any of the datasets. MSE estimates th

```{r img.MSE, fig.height=3, fig.cap="Multiple System Estimation^[Green, Amelia Hoover (2013) Multiple Systems Estimation: Stratification and Estimation
library(png)
library(grid)
img <- readPNG("mse_review.png")
 grid.raster(img)
#![](mse_review.png) need to cite this

```

## MSE Assumptions

There are several MSE assumptions that are need for when there are two "systems" (lists), but not more than that. The assumptions are:

1. *Closed system*: The population of interest does not change during the measurement period. This means that the object of measurement, whether that i

2. *Perfect matching (record linkage!)*: The overlap between systems (i.e., the group of cases recorded in more than one list) is perfectly identified.

3. *Equal probability of capture*: For every data system, each individual has an equal probability of being captured. For example, every death has proba
```
```

~~~
4. *Independence of lists*: Capture in one list does not affect probability of capture in another list. For example, being reported to one NGO does not

Like differences in capture probability, dependencies between systems are impossible to account for in the the two-system setting. A common example here

# Overview of Data

```{r data-sources}
casanare<- read_delim('summary-table.csv', delim= "|")

casanare_capture_recapture <- casanare %>%
  dplyr::select(-c(vln, muni, year)) #just want a matrix of 0s and 1s

cols_lists <- casanare_capture_recapture %>% dplyr::select(starts_with("d_"))

#this function returns any instance of a "capture" for each org
return_1 <-function(x) {
  casanare_capture_recapture %>%
   filter(x ==1) %>%
  summarize(n=sum(Freq)) %>% dplyr::select(n)

}

#unlist(map(cols_lists, return_1))

unique_records<- casanare_capture_recapture %>%
  mutate(sum_cols = reduce(dplyr::select(., starts_with("d_")), `+`)) %>%
  filter(sum_cols == 1)

unique_records<-unique_records %>%
  group_by(d_CCJ, d_EQU, d_FON, d_IMLD ,d_PN0 ,d_CIN, d_FAM ,
           d_FSR, d_IMLM,  d_VP, d_CCE, d_CTI, d_FDC, d_GAU , d_PL) %>%
  summarise(n = sum(Freq))


# print table with each data source and total records "captured"
org_names <- c("Colombian Commission of Jurists", "Equitas", "Fondelibertad", "National Institute of Forensic Medicine Disappearances", "Policía Naciona


contigency_table<-data.frame(org = org_names,
           total_captures= unlist(map(cols_lists, return_1)),
           only_captured_in_this = c(48,0, 67,9,221,91,1,0,1219,284,30,0,376,1,0),
           type = c("judicial", "civil", "security", "forensic", "security",
                "civil","civil", "security","forensic", "judicial", "civil", "judicial",
                "forensic", "security", "civil")) #just hard coded whoops

colnames(contigency_table) <- c("Organization", "Total Captures", "Unique", "Type")

knitr::kable(contigency_table, format="latex", booktabs=T, caption="Contingency Table") %>%
  kableExtra::kable_styling(latex_options="scale_down") %>%
  kableExtra::kable_styling(latex_options = "HOLD_position")

```

## Overall Trends

The following graph shows that violence has been concentrated in Yopal.  We can see that the *reported* violence (i.e. the counts in the 15 datasets) ap

```{r mapping-whole-country}
#group deaths by municipality
muni <- casanare %>% mutate(name_2= str_to_title(muni)) %>%group_by(name_2)  %>% summarise(count = sum(Freq))

colombia_dat <- getData('GADM', country='COL', level = 2) %>%
  st_as_sf() %>%
  clean_names()

casanare_dat <-  colombia_dat %>%
  filter(name_1 == "Casanare") %>%
  mutate(name_2 =  stringi::stri_trans_general(name_2, "Latin-ASCII"))%>% #removes accents for joining purposes
  mutate(name_2 = str_to_title(name_2)) %>%
  left_join( muni, by =c('name_2'))

casanare_overall_plot <- ggplot(casanare_dat, aes(fill = count)) +scale_fill_gradient(low='white', high= 'red')+
  geom_sf() +
  theme_void() +
  ggtitle('Total Counts in 15 Datasets by Municipality 1998-2007')

```

```{r mapping-by-muni}
#partition by years
#group deaths by municipality
muni_by_year<- casanare %>% mutate(name_2 = str_to_title(muni)) %>%
  group_by(name_2, year)  %>%
  summarise(count = sum(Freq))

casanare_dat_with_years <-  colombia_dat %>%
  filter(name_1 == "Casanare") %>%
  mutate(name_2 =  stringi::stri_trans_general(name_2, "Latin-ASCII"))%>%
  mutate(name_2 = str_to_title(name_2)) %>%
  left_join( muni_by_year, by =c('name_2'))



casanare_by_years <- ggplot(casanare_dat_with_years, aes(fill = count)) +scale_fill_gradient(low='white', high= 'red')+
  geom_sf() +
  theme_void() + facet_wrap( ~ year)+
  ggtitle('Yearly Counts in 15 Datasets by Municipality 1998-2007')
```
~~~

~~~
```{r gisplots, fig.cap="Overall Trends in 15 Datasets by Municipality"}
gridExtra::grid.arrange(casanare_overall_plot, casanare_by_years, nrow=1)
~~~

Looking just at the totals of the 15 datasets, that data suggest the violence peaked in 2003-2004 and the total number of victims is 3501. However, jus

## Heterogeneity Issues

One of the requirements for MSE is for each individual has equal probability of capture for a given list. However, we can see from the figure below thi

```{r threeorg2004, fig.height=2.5, fig.cap="Counts from Three Different Datasets in 2004"}
gather_cols <-colnames(casanare %>% dplyr::select(starts_with("d_")))
for_2004<- pivot_longer(casanare, cols= gather_cols, names_to='data_lists', values_to='indicator') %>% filter(indicator == 1) %>%
  filter(data_lists == "d_PN0" | data_lists == "d_IMLM" |  data_lists == "d_CIN")  %>% filter(year == 2004) %>%
  mutate(name_2 = str_to_title(muni))

test_this <- colombia_dat %>% filter(name_1 == "Casanare") %>%
  mutate(name_2 =  stringi::stri_trans_general(name_2, "Latin-ASCII"))%>%
  mutate(name_2 = str_to_title(name_2)) %>%
  full_join(for_2004, by =c('name_2'))

hetromap <- ggplot(test_this, aes(fill= Freq)) + scale_fill_gradient(low='white', high= 'red')+
  geom_sf() +
  theme_void() + facet_wrap( ~ data_lists)+ ggtitle('Disappearances/Kidnappings Across Municipalities by Different Organizations, in 2004')

hetromap
```

## Different Datasets - Different Stories

We motivate our need for a more sophisticated analysis by showing the reporting patterns of 3 different organizations across 1998-2007. If we relied on

```{r motivation-plot}
#list by year

casanare_long <-  gather(casanare, key = "data_lists", value= 'Freq', gather_cols) %>%
  filter(data_lists == "d_GAU" | data_lists == "d_IMLM" |  data_lists == "d_FAM") %>%
  group_by(data_lists,year) %>% summarise(n = sum(Freq))


motivation.plot <- ggplot(casanare_long, aes(x= factor(year), y= as.integer(n), fill = as.factor(data_lists)))+
  geom_bar(stat='identity', position='dodge') + xlab("Year") + ylab("Reported Disappearances/Deaths") +
  ggtitle('Reported Disappearances/Deaths in Casanare') +
  guides(fill = guide_legend(title='Lists'))  +
  theme_bw()
```

```{r, fig.cap= 'Count Trends for Three Organizations'}
motivation.plot
```

\newpage
# Loglinear Modelling

In order to do our estimation, we will use loglinear modeling. We will describe the model as if we had two datasets (for simplicity). We know the vict

```{r img.example, fig.height=3,echo=FALSE}
library(png)
library(grid)
img <- readPNG("2datasetexample.png")
 grid.raster(img)
```

The count of victims captured into a dataset or combination of datasets is $n_{11}, n_{10}, n_{01}$, and $n_{00}$. Each of these cells is a count of vi

Our estimates are primarily based on Poisson regression.

## Estimating the Total Count of Victims

We are interested in estimating $n_{00}$, which also allows us to estimate the total number of victims.
The (log of the) expected cell count $n_{00}$ is a function of the other observed cell counts, as shown in the equation below.

$$
\log(n_{00}) = \alpha
+ \beta_1 \cdot \mathbbm{1}(x\in n_{10})
+ \beta_2 \cdot \mathbbm{1}(x\in n_{01})
$$

This is the saturated form of the log-linear models introduced in Bishop, Fienberg and Holland (1975). To quote from Agresti, "the saturated GLM has a s

When estimation of the total “population” of missing people in Casanare is the goal (as it typically is with multiple-systems estimation), the key value

$$
\begin{aligned}
\log(n_{00}) &= \alpha
\underbrace{
+ \beta_1 \cdot \mathbbm{1}(x\in n_{10})
+ \beta_2 \cdot \mathbbm{1}(x\in n_{01})
}_{\text{each }=0}
\\[.5ex]
\log(n_{00}) & = \alpha
\\[.5ex]
\hat{n}_{00} & = \exp \left\{ \hat{\alpha} \right\}
\\[3ex]
```
```

~~~
\hat{N} &= n_{11} + n_{10} + n_{01} + \hat{n}_{00}
\end{aligned}
$$

As with any regression and data mining model, we want to avoid over-fitting. There is a trade off we need to balance between "goodness of fit", and simp

We need to find the best model in order to get an accurate estimate of $\alpha$. Thus, we should determine whether the full (saturated) model above, whi

$$
\log(n_{00}) = \alpha
+ \beta_1 \cdot \mathbbm{1}(x\in n_{10})
+ \beta_2 \cdot \mathbbm{1}(x\in n_{01})
+ \beta_{12} \cdot \mathbbm{1}(x \in n_{11})
$$



In this case, we can clearly write out the possible model.  During model selection, we estimate these models and choose the model that minimizes the Bay


## Challenges with Loglinear models

1.  **Interpretation:** The inclusion of so many variables in loglinear models often makes interpretation very difficult.
2. **Independence Assumption:** The frequency in each cell is independent of frequencies in all other cells, which is not necessarily the case here. We
3. **Sample Size Requirement:** With loglinear models, you need to have at least 5 times the number of cases as cells in your data.  If you do not have


## Choosing our "Systems"

Our dataset encompasses 15 datasets, far too many to model with a loglinear model. We collapse these 15 datasets into 4 systems (i.e. groups) based on t

```{r adding-systems}
#want to count how many records are in just one of these (or both) of these lists vs. these + another system
cap_indicators<- casanare_capture_recapture  %>% #could have used fct_collapse() here
  mutate(security_ind = case_when((d_GAU == 1 | d_FON == 1 | d_PN0==1) ~ 1,
                                TRUE ~ 0),
         forensic_ind = case_when( (d_IMLM == 1 | d_IMLD==1) ~1 ,
                                  TRUE ~ 0),
        judicial_ind = case_when( (d_CCJ == 1 |d_FSR == 1 | d_CTI == 1 |d_VP ==1 | d_FDC ==1) ~1 ,
                                  TRUE ~ 0),
        civil_ind = case_when( (d_EQU==1 | d_FAM== 1 | d_PL == 1 |d_CCE==1 | d_CIN == 1) ~1 ,
                                  TRUE ~ 0)
         )  %>%
  dplyr::select(-Freq, Freq) # move Freq to last column for closedp() function

table.system <- head(cap_indicators %>% dplyr::select(security_ind, forensic_ind, judicial_ind, civil_ind, Freq))

knitr::kable(table.system, format="latex", booktabs=T
                      , caption="Example of Information from Systems"
                      ,linesep=""
                      ) %>%
  #kableExtra::kable_styling(latex_options="scale_down") %>% for if the table is too large
  kableExtra::kable_styling(latex_options = "HOLD_position")

~~~

```{r systems.venndiagram, fig.height=5, fig.wdith=6, fig.cap="Estimated Venn Diagram of Systems"}
for_venn <- cap_indicators[,c(16:20)] %>%
  group_by(security_ind, forensic_ind, judicial_ind, civil_ind) %>%
   summarise(new_Freq = sum(Freq))
par(mar=c(.1,1,2,1)) #make margins smaller
dga::venn4(c(2000,for_venn$new_Freq),num.test.points = 100000, main='Overlap of Security, Forensic, Judicial and Civil lists') #rough diagram of overlap
```

In the above Venn Diagram, the colored dots represent a specific victim. We are trying to estimate the victims that are not captured in one of the list

## Model Definitions

### Types of Models

Models will be denoted by $M$, and the subscripts will denote the type of model.

* $M_0$: The $M_0$ model is the simplest possible multiple source capture recapture model. It assumes that there is no heterogeneity and that all lists
* $M_t$: This model relaxes the $M_0$ model to allow for lists to have different capture rates.
* $M_h$: This model relaxes the $M_0$ model to allow for individual capture heterogeneity.
* $M_{th}$: This model allows for both list heterogeneity and capture events having different rates.

### Types of Heterogeneity

When heterogeneity in capture probability is present (i.e. the probability of a list capturing a victims differs), there are different forms that this h

* Normal: The log odds of capture follows a Normal distribution.
* Darroch: The log odds of capture among those who were not captured follows a Normal distribution.
* Poisson: The log odds of capture among those who were not captured follows a Poisson distribution.
* Gamma: The log odds of capture among those who were not captured follows a Gamma distribution.

## Non-Hierarchical Models

After collapsing the 15 lists into four systems, we fit several loglinear models. We see that the best fits clearly take into account both system and in

We will need the hierarchical structure to perform model selection. It’s important to note that a model is not chosen if it bears no resemblance to the

The “number of captured units” is the number of observed elements, in this example, the number of people documented as missing/killed, we usually call t

```{r NOT-HIERCH-models}
```

```
test <- cap_indicators[,c(16:20)] %>%
  group_by(security_ind, forensic_ind, judicial_ind, civil_ind) %>%
  summarise(new_Freq = sum(Freq))


loglinear_models <- closedp.t(test, dfreq=T)
NonHierchModels <- loglinear_models$results

#table.logmodels <- as.data.frame(loglinear_models)
knitr::kable(round(NonHierchModels[,1:6], 3)
             , format="latex"
             , booktabs=T
             , caption="Summary of Models (Non-hierarchical models)"
             , linesep="") %>%
  #kableExtra::kable_styling(latex_options="scale_down") %>%
  kableExtra::kable_styling(latex_options = "HOLD_position")

```

Each model controls for a subset of all the possible interactions among the models. In the context of MSE, the two- and three-way interactions estimate

```{r NOT-HIERCH-models-boxplot, fig.cap="Boxplot of Residuals for Models"}
par(mar=c(2,1,2,1)) #make margins smaller
boxplot(loglinear_models) #residuals for heterogen.
```

These boxplots of residuals offer a general assessment of model fit. These boxplots of residuals offer a general assessment of model fit.

Since the individual cell counts:
$$n_i \sim Pois(m_i)$$

$$E[n_i] = Var(n_i) = m_i $$

it follows that the Pearson residuals:

$$r_i = \frac{n_i - \hat{m_i}}{\sqrt(m_i)} $$

are approximately mean 0, variance 1. This is why they are sensible residuals to use.

The light dotted line represents zero, and ideally you want the residuals centered around zero. We see that there is significantly less variation in th

## Specific Model Results

```{r modelProfileLike, fig.cap='Profile Likelihood of Specific Model'}
par(mar=c(4,4,2,1)) #make margins smaller
profilelike <- profileCI(test,dfreq=T, m = "Mth", h = "Darroch", a=2) #profile likelihood CI
```

```{r modelProfileLikeCI}
knitr::kable(round(profilelike$results, 2), format="latex", booktabs=T, caption="Confidence Interval of Specific Model") %>%
  kableExtra::kable_styling(latex_options = "HOLD_position")
```

\newpage
## Capture Recapture

We display some basic capture-recapture frequency statistics to explore capture patterns. It displays, for $i= 1,...t$, the number of people captured $i

If the $n_i$ statistics vary among capture occasions, there is a temporal effect-- which we clearly see here. We would expect the top panel of the plot

```{r descriptive-stats}
desc<- descriptive(test, dfreq = T)
descstat <- t(desc$base.freq)
rownames(descstat) <- c(
   "fi: number of units captured i times"
  ,"ui: number of units captured for the first time on occasion i"
  ,"vi: number of units captured for the last time on occasion i"
  ,"ni: number of units captured on occasion i"
)

knitr::kable(descstat, format="latex", booktabs=T, caption="Capture Recapture Statistics") %>%
  kableExtra::kable_styling(latex_options = "HOLD_position")
```

$$\log\left(\frac{f_i}{{t \choose i}}\right) = \log \left(\frac{N \times P(i \text{ captures})}{{t \choose i}}\right) = \log(N(1-p)^{t-i} p^i = \log(N(1

```{r descriptive-plot, fig.height=6, fig.cap="Capture-Recapture Frequency Statistics"}
plot(desc)
```

\newpage
## Hierarchical Models

### More Interactions: Better Fit

We start by fitting three simple models. The first, and simplest, is the model that assumes independence between the four systems. The second model lo

$$
\begin{aligned}
\text{Model 1:} \quad \log(\hat{N}) &= \alpha + \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4
\\
\text{Model 2:} \quad \log(\hat{N}) &= \alpha + \lambda_1 + \lambda_2 + \lambda_{13} + \lambda_{14} +
\lambda_{23} + \lambda_{24} + \lambda_{34}
\\
\text{Model 3:} \quad \log(\hat{N}) &= \alpha + \lambda_1 + \lambda_2 + \lambda_{13} + \lambda_{14} +
```
```

~~~~
\lambda_{23} + \lambda_{24} + \lambda_{34}  + \lambda_{123} + \lambda_{124} + \lambda_{134} + \lambda_{234}
\\
\end{aligned}
$$

 We perform log-likelihood ratio test and see that the higher the order of the interaction, the better the fit (with saturated model signifying a perfec
$$G^{2}=2 \sum_{i} n_{i} \log \left(n_{i} / \hat{m}_{i}\right)=\sum_{i} d_{i} $$

We perform log-likelihood ratio test and see that the higher the order of the interaction, the greater the deviance explained. Each LRT for the nested m

```{r hierch_models1}
independence_model <- loglm(new_Freq ~ security_ind + forensic_ind + judicial_ind + civil_ind, data=test)

no_three <- loglm(new_Freq ~ security_ind*forensic_ind + security_ind*judicial_ind + security_ind*civil_ind + forensic_ind*judicial_ind + forensic_ind*c

no_four_way <- loglm(new_Freq ~ security_ind*forensic_ind*judicial_ind*civil_ind - security_ind:forensic_ind:judicial_ind:civil_ind, data=test)

anovafit <- anova(independence_model, no_three, no_four_way, test="LR") #LRT

anova.table <- anovafit[1:4,]


knitr::kable(anova.table, format="latex", booktabs=T, caption="ANOVA for 3 standard models") %>%
  kableExtra::kable_styling(latex_options = "HOLD_position")
```

### Iterative Proportional Fitting

We will use iterative proportional fitting to estimate $N$.  The iterative proportional fitting process generates maximum likelihood estimates of the ex
Here the marginal subtables would be:

1. Security by Judicial by Forensic systems
2. Security by Forensic by Civil systems
3. Judicial by Forensic by Civil systems
4. Judicial by Civil by by Security  systems


To do this four-dimensional IPF:

1. Proportionally adjust each (three-dimensional) row of cells to equal the pre-determined totals of Marginal 1.
2. Proportionally adjust each column of cells to equal the pre-determined totals of Marginal 2.
3. Proportionally adjust each slice of cells to equal the pre-determined totals of Marginal 3.
4. Proportionally adjust each stack of cells to equal the pre-determined totals of Marginal 4. This is the end of the first ‘Iteration’.
5. Repeat the above steps until the desired level of convergence is reached.

Under mild restrictions, we know that the cell-values at the end of this process that satisfy the fitted marginal totals are the MLEs.

### Results

Using iterative proportional fitting, we fit all possible second-order interaction models (we restrict to second-order to aid interpretation). These wil

We choose he best model according to the BIC criteria, is below: estimates a total for missing/disappeared people as $8817$, which is similar to the  $8

```{r hierch-models2}
hierch_models <- closedpMS.t(test, dfreq=T ,h='Poisson', maxorder = 2) #restrict to second order interactions

Top <- hierch_models$results[1:5,]
Middle <- hierch_models$results[30:34,]
Bottom <- hierch_models$results[60:64,]
HierchModels <- rbind(Top, Middle, Bottom)
rnames <- rownames(HierchModels)
rownames(HierchModels)<- gsub("\\[|\\]", "", rnames)
#get rid fo brackets otherwise booktabs goes all wonky

knitr::kable(HierchModels[,1:6], format="latex"
             , booktabs=T
             , caption="`Top' Five, `Middle' Five, and `Bottom' Five Models (Hierarchical models)") %>%
  #kableExtra::kable_styling(latex_options="scale_down") %>%
  kableExtra::kable_styling(latex_options = "HOLD_position")
```
```{r estimaterange}
MAX.EST <- max(hierch_models$results[,1])
MIN.EST <- min(hierch_models$results[,1])
results <- as.data.frame(cbind("abundance"=hierch_models$results[,1], "BIC"=hierch_models$results[,6] ))
BIC.results <- results[results$BIC < 300,]
MAX.goodBIC.EST <- max(BIC.results[,1])
MIN.goodBIC.EST <- min(BIC.results[,1])
```

We can see that the independence model listed as "1,2,3,4" has the worst fit-- this is to be expected, as we have demonstrated in previous sections the

The shorthand "14,2,3" means that lists 1 and 4 are independent of 2, and also independent of list 3, i.e. $(1,4) \indep 2$ and $(1,4) \indep 3$. It is

$$\log(\hat{N}) = \alpha + \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4  + \lambda_{14} $$

The shorthand identifies both the model and the margins that must be fitted to obtain MLEs using IPF. Not all shorthands have a conditional independence

These the hierarchical models have different estimates, ranging from 5053 to 9268.

We also plot the BIC values for different models and their accompanying estimates of $\hat{N}$.  When we only look at estimates with a corresponding BIC

```{r hierch-modelsBIC, fig.cap="Hiercharical Models BIC Plot"}
plot(hierch_models) #BIC for a bunch of different models
```

This is the fundamental problem of the frequentist approach. We could just pick the “best” model, i.e., the one with the lowest BIC (note the y-axis is

Unfortunately, just picking one model ignores the error that we introduce by the selection itself. It also forces us to decide which dependencies among
~~~~

~~~
# Conclusions

Our findings show that the total number of deaths and disappearances in Casanare is far greater than previously assessed. The initial analysis made by G

Importantly, given estimation techniques used in our analysis, there is strong evidence that the true total of victims is actually much higher than 3,50


As can be seen, there is a lot of uncertainty about the true total of deaths and disappearances. Nevertheless, the number of deaths and disappearances w


# References

Agresti, A. (2003). Introduction to Categorical Data Analysis. 394.

Ballesteros, A., Restrepo, J., Spagat, M., and Vargas, J. (2007) The Work of Amnesty International and Human Rights Watch: Evidence from Colombia, Bogot

Christensen, R. (2006). Log-Linear Models and Logistic Regression. Springer Science & Business Media.

Darroch, J., Fienberg, S., Gary F. V. Glonek, & Junker, B. (1993). A Three-Sample Multiple-Recapture Approach to Census Population Estimation with Heter

Gomez-Suarez, Andre (2007). “Perpetrator blocs, genocidal mentalities and geographies: the destruction of the Union Patriotica in Colombia and its lesso

Guzman, D., Guberek, T., Hoover, A. and Ball, P. (2007) Missing People in Casanare, Palo Alto, CA: Human Rights Data Analysis Group. Palo Alto, CA: Bene

Junker, B. (2007). Carnegie Mellon Lecture Notes: 36-720: Log-Linear Models: Three-Way Tables. Retrieved from <http://www.stat.cmu.edu/~brian/720/week03

Landman, Todd and Carvalho, Edzia (2009). Measuring Human Rights. London: Routledge.

Mitchell, Shira (2014). Capture-recapture Estimation for Conflict Data and Hierarchical Models for Program Impact Evaluation. Retrieved from <https://da


```{r printcode}
#PRINTING THE CODE
#knitr::stitch("HW06.Rmd") to go to latex
#knitr::stitch(   script="Categorical-FinalProject.Rmd"  , system.file("misc", "knitr-template.Rhtml", package="knitr")) #code to HTML
```
~~~

```
## Error: <text>:10:3: unexpected input
## 9: header-includes:
## 10: - \
##       ^
```

The R session information (including the OS info, R version and all packages used):

```
    sessionInfo()
```

```
## R version 3.6.1 (2019-07-05)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 17763)
## 
## Matrix products: default
## 
## locale:
## [1] LC_COLLATE=English_United States.1252  LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252 LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
## 
## attached base packages:
## [1] grid      stats     graphics  grDevices utils     datasets  methods   base
## 
## other attached packages:
##  [1] png_0.1-7        kableExtra_1.1.0 bookdown_0.12    MASS_7.3-51.4    rgeos_0.5-2
##  [6] dga_1.2          Rcapture_1.4-2   raster_3.0-7     sf_0.8-0         rgdal_1.4-8
## [11] sp_1.3-1         janitor_1.2.0    forcats_0.4.0    stringr_1.4.0    dplyr_0.8.3
## [16] purrr_0.3.2      readr_1.3.1      tidyr_1.0.0      tibble_2.1.3     ggplot2_3.2.1
## [21] tidyverse_1.2.1
## 
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.2         lubridate_1.7.4    lattice_0.20-38    class_7.3-15
##  [5] assertthat_0.2.1   zeallot_0.1.0      digest_0.6.20      R6_2.4.0
##  [9] cellranger_1.1.0   chron_2.3-54       backports_1.1.4    evaluate_0.14
## [13] e1071_1.7-2        highr_0.8          httr_1.4.1         pillar_1.4.2
## [17] rlang_0.4.0        lazyeval_0.2.2     readxl_1.3.1       rstudioapi_0.10
## [21] rmarkdown_1.14     webshot_0.5.1      munsell_0.5.0      broom_0.5.2
## [25] compiler_3.6.1     modelr_0.1.5       xfun_0.8           pkgconfig_2.0.2
## [29] htmltools_0.3.6    tidyselect_0.2.5   gridExtra_2.3      codetools_0.2-16
## [33] viridisLite_0.3.0  crayon_1.3.4       withr_2.1.2        nlme_3.1-140
## [37] jsonlite_1.6       gtable_0.3.0       lifecycle_0.1.0    DBI_1.0.0
## [41] magrittr_1.5       units_0.6-5        scales_1.0.0       KernSmooth_2.23-15
## [45] cli_1.1.0          stringi_1.4.3      xml2_1.2.2         generics_0.0.2
## [49] vctrs_0.2.0        tools_3.6.1        glue_1.3.1         hms_0.5.0
## [53] yaml_2.2.0         colorspace_1.4-1   classInt_0.4-2     rvest_0.3.4
## [57] knitr_1.24         haven_2.1.1
```

```
    Sys.time()
```

```
## [1] "2019-12-05 13:11:34 CST"
```