MTH3041 Coursework

2023-01-12

Model

Import the libraries:

```
library(brms)
library(tidyverse)
library(dplyr)
library("gridExtra")
library("tidybayes", lib = "/Users/hpapa/r_packages")
```

Set the number of cores to use when compiling our models.

```
options(mc.cores = parallel::detectCores())
```

Load in the dataset containing the details of the homicides that occurred in London from 2003 to 2022.

```
homicides = read.csv("homicides.csv")
```

All categorical features in the dataset were converted to factors. Furthermore, I created a new feature which indicates whether the homicide was committed before or after 2010. The conservative party came into power in 2010 and I thought it would be interesting to look at the effect of this on the chances of solving homicides. This is all contained in a function so that I can apply this to the hypothetical homicides later.

```
clean_df = function(dataframe){
  if("solved_status"%in%colnames(dataframe)){
    dataframe$solved_status = as.factor(dataframe$solved_status)
    dataframe$solved_status = factor(dataframe$solved_status, levels = c("Unsolved", "Solved"))
  }

  dataframe$domestic_abuse = as.factor(dataframe$domestic_abuse)
  dataframe$domestic_abuse = factor(dataframe$domestic_abuse,
                                    level = c("Not Domestic Abuse", "Domestic Abuse"))
  dataframe$borough = as.factor(dataframe$borough)
  dataframe$method_of_killing = as.factor(dataframe$method_of_killing)
  dataframe$observed_ethnicity = as.factor(dataframe$observed_ethnicity)
  dataframe$sex = as.factor(dataframe$sex)
  dataframe$age_group = as.factor(dataframe$age_group)
  dataframe$month.name = as.factor(dataframe$month.name)
  dataframe$month.name = factor(dataframe$month.name,
                                levels = c("JAN", "FEB", "MAR", "APR", "MAY",
                                           "JUN", "JUL", "AUG", "SEP", "OCT",
                                           "NOV", "DEC"))
  dataframe$season = as.factor(dataframe$season)
  dataframe = dataframe %>%
    mutate(before_after_2010 = ifelse(year > 7, "after2010", "before2010"))
  dataframe$before_after_2010 = as.factor(dataframe$before_after_2010)
  dataframe$before_after_2010 = factor(dataframe$before_after_2010,
                                       levels = c("before2010", "after2010"))

  # Renaming some of the columns of our dataset so that the summary table
  # for our model is not too big.
  dataframe = dataframe %>% rename(
    ethn = observed_ethnicity,
    domAb=domestic_abuse  ,
    methOfK = method_of_killing,
    ba2010= before_after_2010
  )
  dataframe
}
homicides = clean_df(homicides)
```

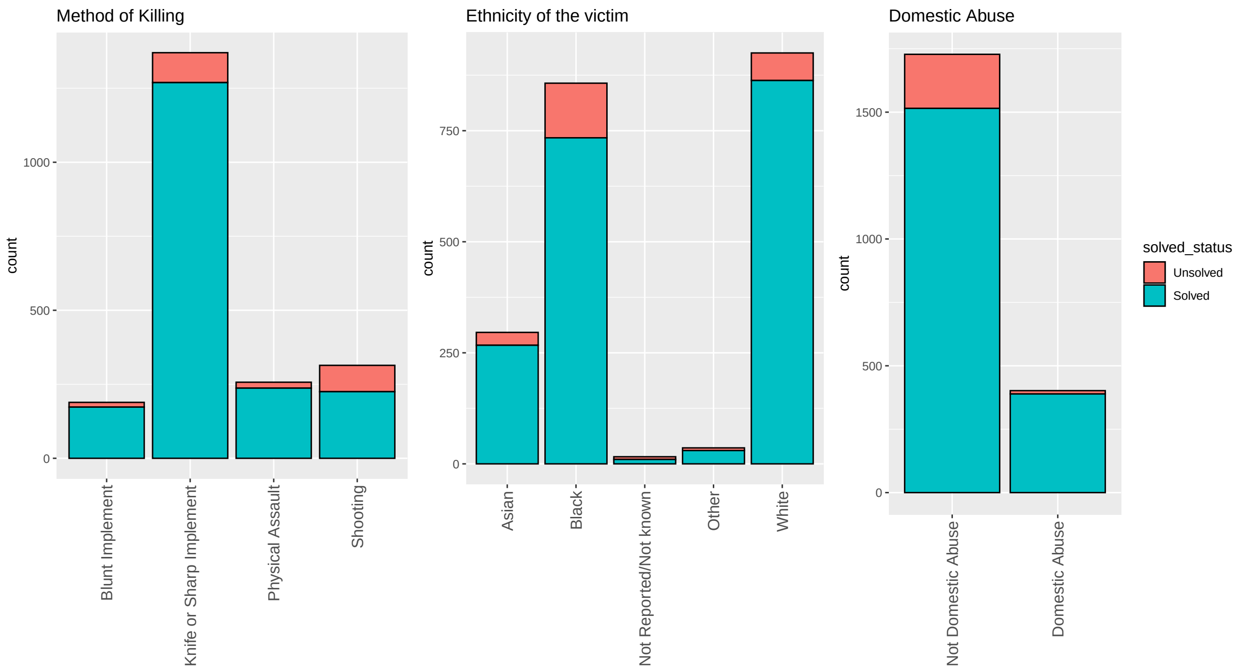
Furthermore, rather than looking at differences in the probability of a crime being solved in a specific borough, I thought it would be more interesting to look at the differences in the solving of homicides with regards to the poverty rate of the borough in which the homicide was committed. To do so, the boroughs were divided into three groups based off the poverty rates which are available at https://www.trustforlondon.org.uk/data/poverty-borough/. If a borough has a poverty rate greater than 28, it was classified as being a "poor" borough, if a borough has a poverty rate inferior to 23 it was classed as being "rich", else it was classed as "middle". The thresholds are arbitrary as I could not find a definition for each group. They were chosen so as to have around the same amount of boroughs per group. The function

```
borough_poverty_group(dataframe)
```

returns a vector corresponding to the poverty group of the borough in which each homicide was committed. The inner workings of the function have been omitted from this report as it is just a big for loop setting different poverty rates for each borough and storing them in a vector and would take up 2 pages of this report.

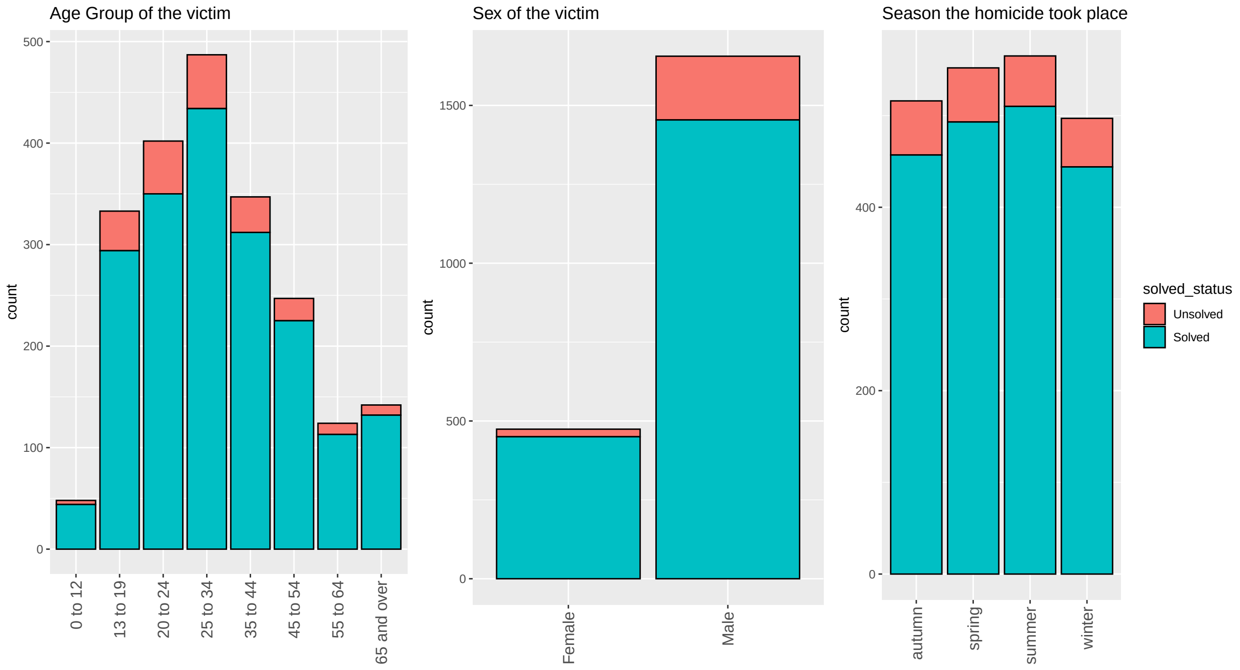
```
homicides$borough_class_income = borough_poverty_group(homicides)
homicides = homicides %>% rename(
    poverty = borough_class_income
  )
```

Exploring the data


Method of Killing

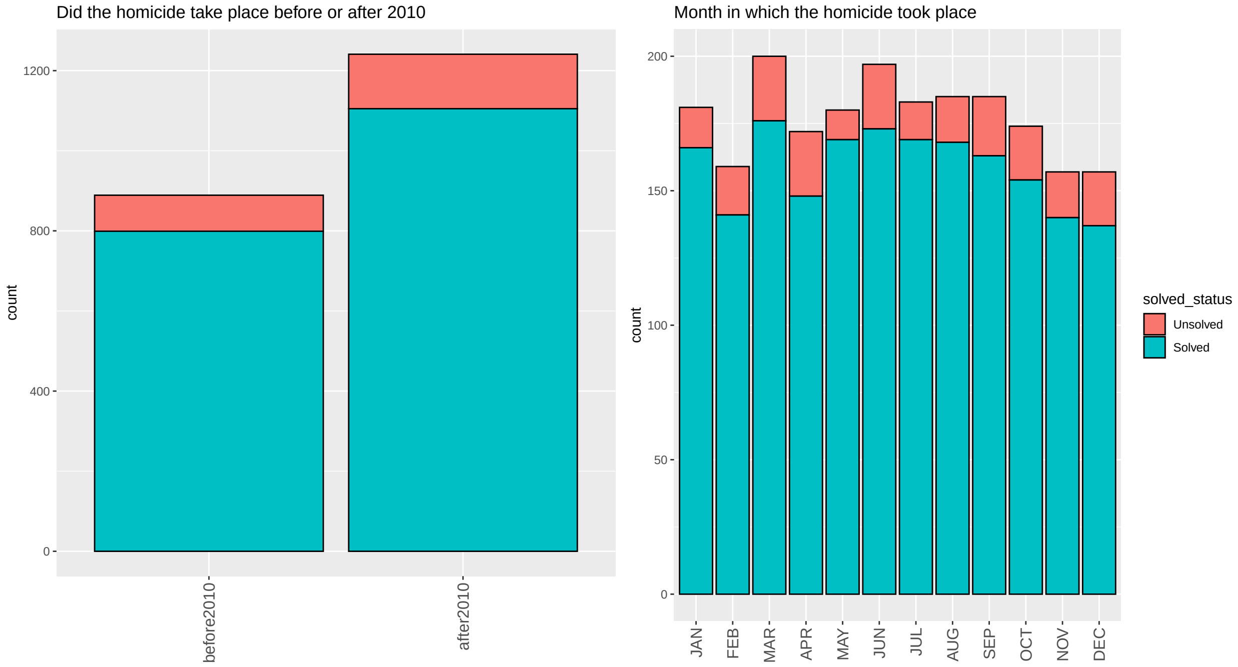
Ethnicity of the victim

Domestic Abuse


Age Group of the victim

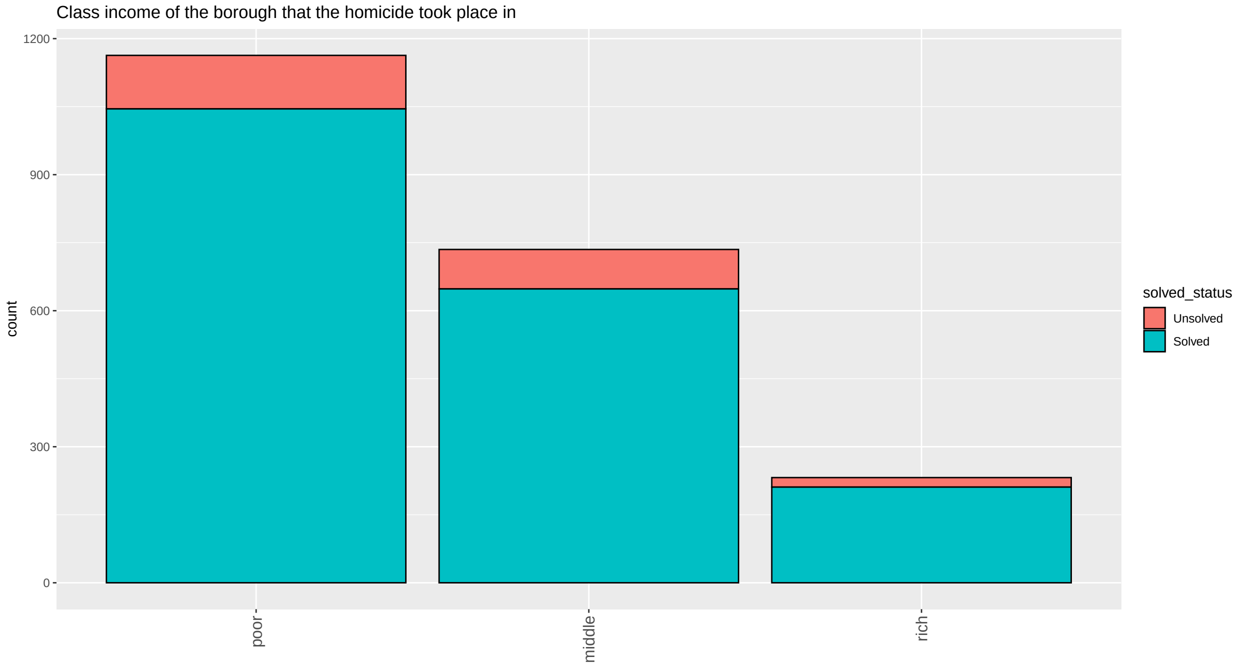
Sex of the victim

Season the homicide took place



Did the homicide take place before or after 2010

Month in which the homicide took place



Class income of the borough that the homicide took place in

There are multiple interesting points that arise when looking at these plots. Firstly, we can see that there are differences in the proportion of homicides that are solved depending on the method of killing. Homicides done using a knife or another sharp object seem to be much more likely to be solved than crimes involving other methods. There are roughly the same amount of unsolved homicides that involved shootings as there are for ones involving knives or sharp objects even even though there have been around 1100 more homicides involving knives/sharp objects.

Additionally, we notice that even though there have been less homicides with a black victim than with a white victim, more homicides involving white victims have been solved.

Moreover, it seems to be that there is a very slim chance of a crime being solved if the case involved domestic abuse. However, this might not be that domestic abuse was not involved, it might be that no one knew that the victim was also victim to domestic abuse as often times domestic abuse remains unreported until it is too late.

It also seems like there is a slimmer chance of a homicide being solved if the victim is female than if the victim is a male.

In addition, the proportion of total homicides being solved seems to be greater since 2010 compared to before 2010, which would seem to indicate that under conservative leadership, there is a greater success rate in solving homicides.

Furthermore, it seems that there is a higher chance of a homicide being solved in a "poorer" borough than in a richer one.

On the other hand, 3 of the features look to be quite uninformative. For the age_group feature the relationship seems to be that the more homicide victims there are per age group, the more homicides are solved, which is not very interesting as that sort of relationship is what we would expect. The same can be said for the month feature. The season in which a homicide was committed also does not seem very informative as there were around the same total number of homicides per season since 2003 as well as around the same amount of homicides solved.

Model

In this report, we will be trying to predict whether a given homicide will be solved or not using the homicides dataset. This is therefore a binary classification problem as there are only two possible values that the outcome variable can take: "Unsolved" and "Solved". A Logistic Regression is therefore an appropriate model that can be used for this task, and in this report we will be fitting a Bayesian Hierarchical Logistic Regression.

Following the analysis of the plots, the model we fit will contain the following features, the sex and ethnicity of the victim, the method of killing, whether or not domestic abuse was involved and the features that we generated about whether the homicide was done before or after 2010 and whether it was in a "poor", "middle" or "rich" borough. Furthermore, in my opinion, the conservative party tend to favor the rich over the poor. Therefore, I thought it would be interesting to group the variable corresponding to the "poverty class" of the borough in which the homicide took place on whether the homicide happened before or after 2010. For example, I thought it might be more likely that a homicide in a richer borough would be more likely to be solved after 2010 than before 2010, when the labour party lost power and the conservatives gained power. Additionally, I also thought that it might be interesting to group the domestic abuse variable by the sex of the victim. Indeed, I thought it might be more or less likely that a homicide case involving domestic abuse may be more or less likely to be solved depending on the sex of the victim.

The model will therefore be of the form

$$y_{ij}|\beta_j, x_{ij} \sim Ber(p = g^{-1}(\beta, x))$$

$$logit(p) = log(\frac{p}{1-p})$$

$$= b_0 + \sum_{k=2}^{5} b_k x_{ki} + b_7 x_{7i} + \sum_{k=9}^{11} b_k x_{ki} + b_{13} x_{13i} + \sum_{k=15}^{16} b_k x_{ki} + b_{18} x_{18i}$$

$$+\beta_{0j} + \sum_{k=15}^{16} \beta_{kj} x_{kij} + \beta_{0z} + \beta_{7z} x_{kiz}$$

where b_0 is the intercept and to which the first level of each of our factors in the model are aliased to. Here, p is the inverse link function defined as $\frac{1}{1+e^{-x}}$

β_{0j} is the intercept for the j^{th} level of the first grouping variable, here ba2010, where $j = 1$ corresponds to before 2010 and $j = 2$ corresponds to after 2010.

β_{0z} is the intercept for the z^{th} level of the second grouping variable, here the sex of the victim, where $z = 1$ corresponds to a female victim and $z = 2$ corresponds to a male victim.

For $k \in [2, 5]$: $x_{ki} = 1$ if x_i belongs to the k^{th} level of the observed ethnicity factor, and 0 otherwise; b_k is the corresponding coefficient. For $k = 2$, the victim was black, for $k = 3$, the victim's ethnicity was not reported, for $k = 4$, the victim was of another ethnicity, and for $k = 5$ the victim was white. The first level, corresponding to an asian victim was aliased to the intercept.

For $k = 7$: $x_{ki} = 1$ if x_i belongs to a homicide where domestic abuse was involved, and 0 otherwise; b_7 is the corresponding coefficient. The first level of this factor, corresponding to a case where domestic abuse was not involved, was aliased to the intercept. Moreover, $x_{kiz} = 1$ if x_i belongs to the z^{th} level of the sex feature and domestic abuse was involved. β_{7z} is the corresponding coefficients.

For $k \in [9, 11]$: $x_{ki} = 1$ if x_i belongs to the $(k - 7)^{th}$ level of the method of killing factor, and 0 otherwise; b_k is the corresponding coefficient. For $k = 9$, the homicide was done using a knife or a sharp object, for $k = 10$, with physical assault, for $k = 11$, using a gun. The first level of this factor, corresponding to a blunt object was aliased to the intercept.

For $k = 13$: $x_{ki} = 1$ if x_i belongs to a homicide where the victim is male, and 0 otherwise; b_k is the corresponding coefficient.

For $k \in [15, 16]$: $x_{ki} = 1$ if x_i belongs to the $(k - 13)^{th}$ level of the borough class income factor, and 0 otherwise; b_k is the corresponding coefficient. Where $k = 15$ corresponds to a borough with a "middle" class income and $k = 16$ to a borough with a "rich" class income. The "poor" class income was aliased to the intercept. Furthermore, $x_{kij} = 1$ if x_i belongs to the j^{th} level of the before or after 2010 feature, and to the $(k - 13)^{th}$ level of the class income factor, as previously defined.

When $k = 18$: $x_{ki} = 1$ if x_i is a homicide that took place after 2010, and 0 otherwise; b_{18} is the corresponding coefficient.

We have $\beta_j \sim N(0, \Sigma)$ and must choose prior $\pi(b, \Sigma)$

Defining and setting the priors

Firstly, I set the prior for b_0 which corresponds to the intercept, more specifically how likely it is for any given homicide to be solved regardless of any extra information. From the data we can see that around 89% of all homicides are solved:

```
length(homicides$solved_status[homicides$solved_status=="Solved"])/nrow(homicides)
```

```
## [1] 0.8938967
```

```
l = round(logit_scaled(0.65),3)
u = round(logit_scaled(0.95),3)
l
```

```
## [1] 0.619
```

```
u
```

```
## [1] 2.944
```

I believe that at worse, any given homicide has a 65% chance of being solved, and a 95% chance of being solved in the best of cases. Translating this to $logit(p) = log(odds)$ returns 0.619 and 2.944 respectively. Now, I want 95% of my distribution to be contained between 0.619 and 2.944. Using a normal prior for b_0 with mean μ and sd σ, $N(\mu, \sigma^2)$, means we want to simultaneously solve

$$0.619 = \mu - 2\sigma$$

$$2.944 = \mu + 2\sigma$$

which yields $\mu = 1.7835$ and $\sigma = 0.58225$ and our prior for b_0 is therefore $N(1.7835, 0.58225^2)$.

For the b_k, I first look at their relationship with $logit(p)$. Following the previously defined mathematical model for our Logistic Regression, we can see that everything else remaining constant, if $x_{ki} = 1$, $logit(p)$ will increase or decrease by b_k. So for example, if the victim of the homicide was male, if everything else is 0, besides b_0, then $logit(p)$ will increase or decrease by b_{13}. This means that the difference that the victim being the male makes to $logit(p)$ is $|logit(p_0) - logit(p_{13})|$, where p_0 is the probability we give for any homicide being solved and p_{13} is the probability we give for a homicide being solved where the victim is a male, at worst. We can then use this difference to set our priors.

Weakly informative priors will be used for b_k, of the shape $N(0, \sigma^2)$ where σ will be $\frac{|logit(p_0)-logit(p_k)|}{2}$. The reasoning behind this is that if we set $\sigma = \frac{|logit(p_0)-logit(p_k)|}{2}$ then we are saying that we believe that 95% of the distribution of b_k is between $-|logit(p_0) - logit(p_k)|$ and $|logit(p_0) - logit(p_k)|$.

So for the example described previously, looking at the data and from my own beliefs, I believe that it is more likely that a homicide with a male victim is, at worse, 15% less likely to be solved than the homicide of a female (this one being aliased to the intercept) so I set the prior for b_{13} to be $N(0, \frac{|logit(0.9)-logit(0.75)|}{2}^2) = N(0, 0.549^2)$

```
abs(logit_scaled(0.9)-logit_scaled(0.75))/2
```

```
## [1] 0.5493061
```

The same type of reasoning will be used to set all the priors for the b_k.

From the plots of the data, and my own beliefs, I believe that a homicide where the victim is black is 20% less likely to be solved than the average homicide. I believe that there are racial biases within police forces and governing bodies in the UK that mean that there might be less “incentive” for them to solve the homicide of a black victim. My prior for b_2 is therefore $N(0, \frac{|logit(0.9)-logit(0.7)|}{2}^2) = N(0, 0.675^2)$. In the same school of thought, I believe that the chances of the homicide of a white person being solved is 5% greater than the average so my prior for b_5 is $N(0, \frac{|logit(0.9)-logit(0.95)|}{2}^2) = N(0, 0.374^2)$. I also believe that victims of other ethnicity are less likely to have their homicide solved although not to the same level as for black victims and therefore I set the prior for b_4 as $N(0, \frac{|logit(0.9)-logit(0.85)|}{2}^2) = N(0, 0.231^2)$. I do not really have any beliefs regarding the effect of the ethnicity of the victim’s ethnicity being unreported on whether the homicide gets solved or not, but I believe that they may be slightly more likely and therefore set $b_3 \sim N(0, \frac{|logit(0.9)-logit(0.93)|}{2}^2) = N(0, 0.195^2)$.

```
abs(logit_scaled(0.9)-logit_scaled(0.7))/2
```

```
## [1] 0.6749634
```

```
abs(logit_scaled(0.9)-logit_scaled(0.95))/2
```

```
## [1] 0.3736072
```

```
abs(logit_scaled(0.9)-logit_scaled(0.85))/2
```

```
## [1] 0.2313118
```

```
abs(logit_scaled(0.9)-logit_scaled(0.93))/2
```

```
## [1] 0.1947324
```

Moreover, I believe that the chances of a homicide being solved are quite higher than the average if domestic abuse is involved, around 8% higher, because I feel like there is an obvious suspect to go after if domestic abuse is involved. This also seems to be backed at the data. My prior for b_7 is then $N(0, \frac{|logit(0.9)-logit(0.98)|}{2}^2) = N(0, 0.847^2)$.

```
abs(logit_scaled(0.9)-logit_scaled(0.98))/2
```

```
## [1] 0.8472979
```

Furthermore, the plots seem to indicate that there is a much greater chance of a homicide being solved if a knife/sharp object was used to kill. I deem this to be around 97% as I also believe what the data is showing, I believe it more likely to find a knife that was used to kill and retrieve DNA traces from it rather than a blunt object or, say, if a victim died of physical abuse. My $\pi(b_9)$ is then $N(0, \frac{|logit(0.9)-logit(0.97)|}{2}^2) = N(0, 0.639^2)$. I would have believed it to be the same case for a gun, however the plot does not seem to tell the same story. Indeed, I would believe the chances of solving a homicide involving a gun to be around 95% and so I set $b_{11} \sim N(0, \frac{|logit(0.9)-logit(0.95)|}{2}^2) = N(0, 0.374^2)$. As mentioned previously, I believe that there is less chance of solving a homicide where physical abuse killed the victim, around 10% less than the average homicide so I set the prior for $b_{10} \sim N(0, \frac{|logit(0.9)-logit(0.8)|}{2}^2) = N(0, 0.405^2)$.

```
abs(logit_scaled(0.9)-logit_scaled(0.97))/2
```

```
## [1] 0.6394371
```

```
abs(logit_scaled(0.9)-logit_scaled(0.8))/2
```

```
## [1] 0.4054651
```

When looking at the effect of the poverty level of a specific borough on the chances of a homicide having taken place in that bough being solved, I believe that a homicide that took place in a richer borough has a higher chance of being solved than the average homicide. I deem this to be around 5% greater because I believe that a homicide in a richer borough would provoke more commotion in the police than a homicide in a poorer borough given that there are more homicides in poorer boroughs.I then set the prior $b_{16} \sim N(0, \frac{|logit(0.9)-logit(0.95)|}{2}^2) = N(0, 0.374^2)$. In continuance with my previous argument, I believe that a homicide done in a borough with a "middle" class income has around 5% lesser chance of being solved than the average homicide, yielding the prior $b_{15} \sim N(0, \frac{|logit(0.9)-logit(0.85)|}{2}^2) = N(0, 0.231^2)$.

Furthermore, I believe that a homicide that took place after 2010, when the conservatives came into power, has 5% less chance to be solved than the average homicide because I believe that the conservatives have defunded the police when in power. I therefore set the prior $b_{18} \sim N(\frac{|logit(0.9)-logit(0.85)|}{2}^2) = N(0, 231^2)$

We now must set priors for the Σ of the β. When using the brm package, the β have a $N(0, \Sigma)$ distribution. Looking back at the definition of our model, we see that the β play the same role as the b_k in how our $logit(p)$ changes. For example, if a victim is male, and domestic abuse was involved then $logit(p)$ will increase/decrease by $\beta_{7,2}$. Following from this train of thought, we can set the priors for the different σ in the same way that we set the priors for b_k. The prior for the β intercepts will remain default as the defaults are wide. However we will set weakly informative priors for the slope β that will be of the same shape as the corresponding priors for the b_k. So we have $\pi(\sigma_{15}) \sim N(0, 0.231^2)$, $\pi(\sigma_{16}) \sim N(0, 0.374^2)$ and $\pi(\sigma_7) \sim N(0, 0.847^2)$.

Setting the priors in R:

```
Intercept_prior = set_prior("normal(1.7835, 0.58225)", class = "Intercept")
BlackEthnicity_prior = set_prior("normal(0, 0.675)", class = "b", coef = "ethnBlack")
WhiteEthnicity_prior = set_prior("normal(0, 0.374)", class = "b", coef = "ethnWhite")
NotReportedEthnicity_prior = set_prior("normal(0, 0.195)", class = "b",
                                       coef = "ethnNotReportedDNotknown")
OtherEthnicity_prior = set_prior("normal(0, 0.231)", class = "b", coef = "ethnOther")
DomesticAbuse_prior = set_prior("normal(0, 0.847)", class = "b", coef = "domAbDomesticAbuse")
KnifeSharpObj_prior = set_prior("normal(0,0.639)", class = "b", coef = "methOfKKnifeorSharpImplement")
PhysAssault_prior = set_prior("normal(0,0.405)", class = "b", coef = "methOfKPhysicalAssault")
Shooting_prior = set_prior("normal(0,0.374)", class = "b", coef = "methOfKShooting")
Sex_prior = set_prior("normal(0,0.549)", class = "b", coef ="sexMale")
MiddleBorough_prior = set_prior("normal(0,0.231)", class = "b", coef = "povertymiddle")
RichBorough_prior = set_prior("normal(0,0.374)", class = "b", coef = "povertyrich")
after2010_prior = set_prior("normal(0,0.232)", class = "b", coef = "ba2010after2010")
sdMiddle_prior = set_prior("normal(0, 0.231)", class = "sd", group = "ba2010",
                          coef = "povertymiddle")
sdRich_prior = set_prior("normal(0, 0.374)", class = "sd", group = "ba2010",
                           coef = "povertyrich")
sdDomAbuse_sex_prior = set_prior("normal(0, 0.847)", class = "sd", group = "sex",
                           coef = "domAbDomesticAbuse")

priors = c(Intercept_prior, BlackEthnicity_prior, WhiteEthnicity_prior, NotReportedEthnicity_prior,
                         OtherEthnicity_prior, DomesticAbuse_prior, KnifeSharpObj_prior, PhysAssault_prior,
                         Shooting_prior, Sex_prior, MiddleBorough_prior, RichBorough_prior, after2010_prior,
                         sdMiddle_prior, sdRich_prior, sdDomAbuse_sex_prior)
```

Checking the convergence of the MCMC

Before fitting the model, split the dataset into a training set (70% of the data) and a testing set (30% of the data)

```
homicides$id <- 1:nrow(homicides)
homicides.tr = homicides %>% dplyr::sample_frac(0.70)
homicides.te = dplyr::anti_join(homicides, homicides.tr, by = 'id')
```

Then, using the priors defined in the previous section, the model is:

```
model = brm(formula = solved_status ~ ethn + domAb + methOfK + sex +
                      poverty + ba2010 + (poverty|ba2010) +
                      (domAb|sex), data = homicides.tr,
           chains = 2, cores = 2, family = bernoulli(),
           prior = priors, iter = 20000)
```

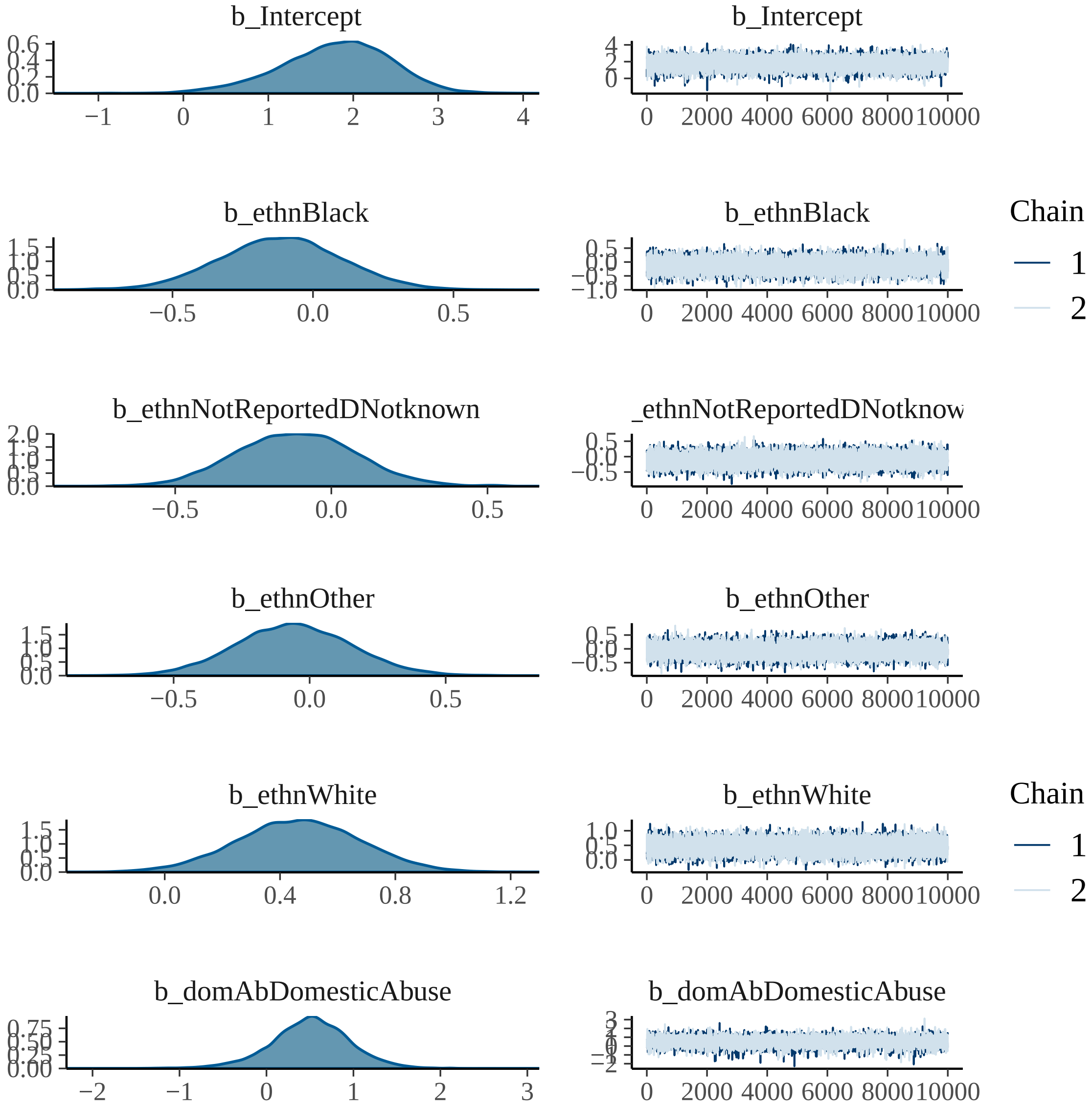
The base number of iterations is set to 2000, however this does not allow enough iterations for the MCMC to converge and therefore I set the iterations to 20000. The summary of the model is as follows:

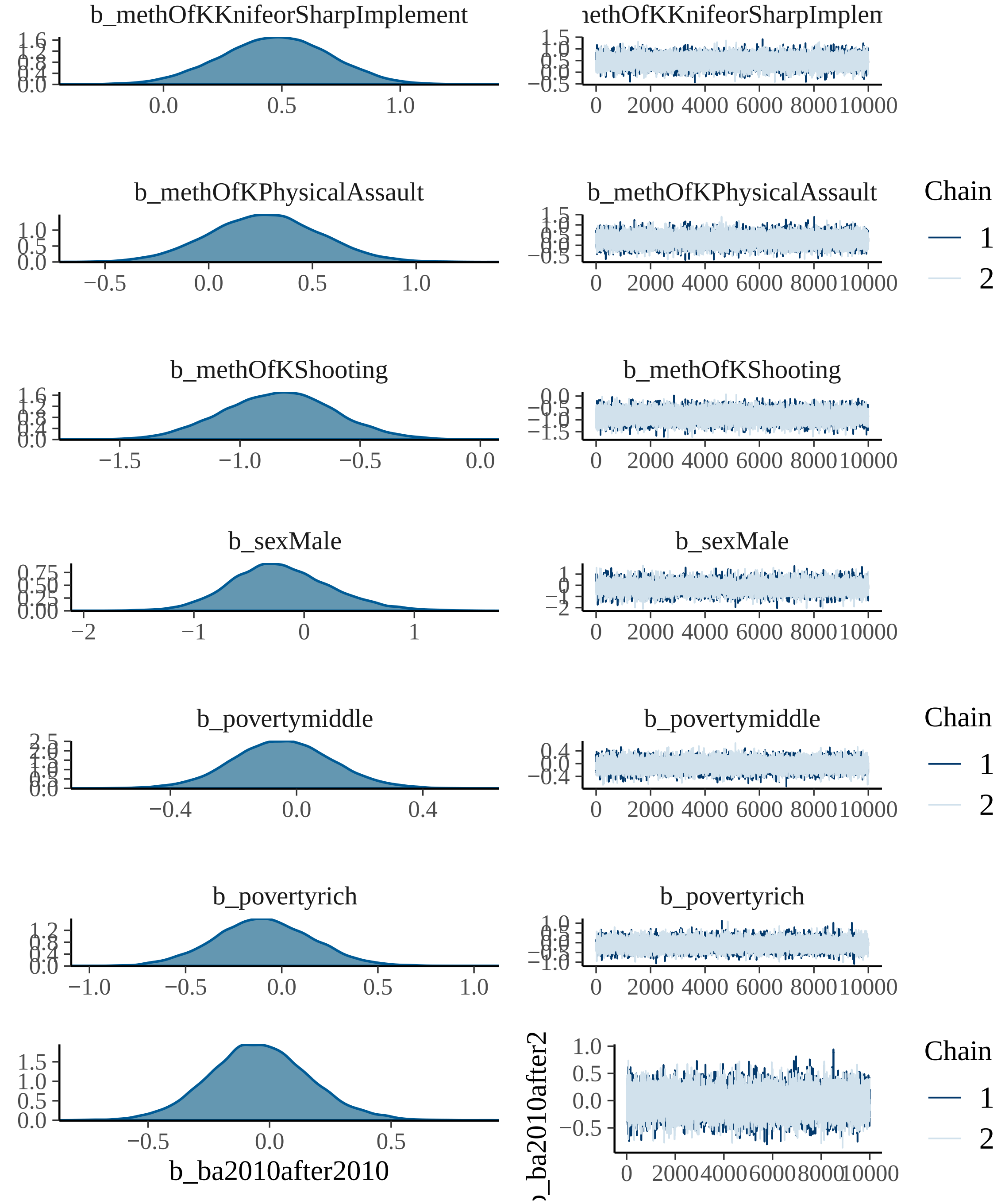
```
summary(model)
```

```
## Warning: There were 311 divergent transitions after warmup. Increasing
## adapt_delta above 0.8 may help. See http://mc-stan.org/misc/
## warnings.html#divergent-transitions-after-warmup
```

```
##  Family: bernoulli
##   Links: mu = logit
## Formula: solved_status ~ ethn + domAb + methOfK + sex + poverty + ba2010 + (poverty | ba2010) + (domAb | sex)
##    Data: homicides.tr (Number of observations: 1491)
##   Draws: 2 chains, each with iter = 20000; warmup = 10000; thin = 1;
##          total post-warmup draws = 20000
##
## Group-Level Effects:
## ~ba2010 (Number of levels: 2)
##                                 Estimate Est.Error l-95% CI u-95% CI Rhat
## sd(Intercept)                       0.74      0.85     0.02     3.13 1.00
## sd(povertymiddle)                   0.15      0.12     0.00     0.43 1.00
## sd(povertyrich)                     0.24      0.19     0.01     0.71 1.00
## cor(Intercept,povertymiddle)       -0.01      0.51    -0.89     0.88 1.00
## cor(Intercept,povertyrich)         -0.01      0.50    -0.88     0.87 1.00
## cor(povertymiddle,povertyrich)      0.01      0.50    -0.87     0.88 1.00
##                                 Bulk_ESS Tail_ESS
## sd(Intercept)                      4829     5222
## sd(povertymiddle)                 11014     7414
## sd(povertyrich)                   10017     7268
## cor(Intercept,povertymiddle)      16110    10175
## cor(Intercept,povertyrich)        18312    12877
## cor(povertymiddle,povertyrich)    14641    13029
##
## ~sex (Number of levels: 2)
##                                   Estimate Est.Error l-95% CI u-95% CI Rhat
## sd(Intercept)                         0.92      0.91     0.03     3.42 1.00
## sd(domAbDomesticAbuse)                0.52      0.43     0.02     1.60 1.00
## cor(Intercept,domAbDomesticAbuse)    -0.02      0.59    -0.95     0.95 1.00
##                                   Bulk_ESS Tail_ESS
## sd(Intercept)                         5485     4935
## sd(domAbDomesticAbuse)                8733     4795
## cor(Intercept,domAbDomesticAbuse)    13718    11464
##
## Population-Level Effects:
##                              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
## Intercept                        1.79      0.65     0.41     2.96 1.00     8813
## ethnBlack                       -0.11      0.22    -0.54     0.31 1.00    12741
## ethnNotReportedDNotknown        -0.11      0.19    -0.47     0.27 1.00    12081
## ethnOther                       -0.06      0.21    -0.47     0.36 1.00    18223
## ethnWhite                        0.47      0.21     0.05     0.88 1.00    13071
## domAbDomesticAbuse               0.50      0.47    -0.47     1.40 1.00    11711
## methOfKKnifeorSharpImplement     0.47      0.23     0.02     0.92 1.00    11709
## methOfKPhysicalAssault           0.27      0.27    -0.25     0.80 1.00    15261
## methOfKShooting                 -0.84      0.23    -1.29    -0.38 1.00    11454
## sexMale                         -0.21      0.46    -1.05     0.76 1.00    11213
## povertymiddle                   -0.04      0.16    -0.35     0.27 1.00    17008
## povertyrich                     -0.09      0.25    -0.59     0.41 1.00    14733
## ba2010after2010                 -0.04      0.21    -0.44     0.38 1.00    13626
##                              Tail_ESS
## Intercept                        8201
## ethnBlack                       12344
## ethnNotReportedDNotknown         6998
## ethnOther                       11719
## ethnWhite                       13002
## domAbDomesticAbuse               9352
## methOfKKnifeorSharpImplement    10041
## methOfKPhysicalAssault          14235
## methOfKShooting                 12010
## sexMale                         12202
## povertymiddle                   13831
## povertyrich                     10973
## ba2010after2010                 12682
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

And the posterior distributions density plots as well as the trace plots are as follows for the population level effects:


b_Intercept
0.6
0.4
0.2
0.0
−1
0
1
2
3
4
b_Intercept
4
2
0
0
2000
4000
6000
8000
10000
b_ethnBlack
1.5
1.0
0.5
0.0
−0.5
0.0
0.5
b_ethnBlack
0.5
0.0
−0.5
−1.0
Chain
1
2
b_ethnNotReportedDNotknown
2.0
1.5
1.0
0.5
0.0
−0.5
0.0
0.5
_ethnNotReportedDNotknow
0.5
0.0
−0.5
b_ethnOther
1.5
1.0
0.5
0.0
−0.5
0.0
0.5
b_ethnOther
0.5
0.0
−0.5
b_ethnWhite
1.5
1.0
0.5
0.0
0.0
0.4
0.8
1.2
b_ethnWhite
1.0
0.5
0.0
Chain
1
2
b_domAbDomesticAbuse
0.75
0.50
0.25
0.00
−2
−1
0
1
2
3
b_domAbDomesticAbuse
3
2
1
0
−1
−2



As we can see, the MCMC seems to have converged, the trace plots are well mixed, there are no discernible patterns in the trace plots , the rhat values are all equal to 1 and the effective sample sizes for both the bulk of the posterior distributions and the tail of the posterior distributions are high.

Model Checking/Validation

We can now validate this model by looking at how it performs at predicting unseen homicides (here at predicting the homicides in the testing set). We first obtain the predictions for the testing dataset:

```
preds = predict(model, homicides.te)
head(preds,1)
```

```
##      Estimate Est.Error Q2.5 Q97.5
## [1,]   0.9541 0.2092735    0     1
```

The estimate returned is the probability that a crime is solved. However, we want to obtain a classification for each homicide in the testing set. The common threshold is to classify everything with a probability greater than 0.5 as being of class 1.

```
classifier <- preds[,"Estimate"]>=0.5
```

We can then use a confusion matrix to visualize how using this threshold would classify our points and calculate the accuracy:

```
ConfusionMatrix <- function(Classifier, Truth){
  if(!(length(Classifier)==length(Truth)))
    stop("Make the length of your vector of predictions the same as the length of the truth")
  if(is.logical(Classifier))
    Classifier <- as.integer(Classifier)
  WhichClass0s <- which(Classifier < 1)
  ZeroCompare <- Truth[WhichClass0s]
  Predicted0 <- c(length(ZeroCompare)-sum(ZeroCompare), sum(ZeroCompare))
  WhichClass1s <- which(Classifier>0)
  OnesCompare <- Truth[WhichClass1s]
  Predicted1 <- c(length(OnesCompare)-sum(OnesCompare), sum(OnesCompare))
  ConMatrix <- cbind(Predicted0,Predicted1)
  row.names(ConMatrix) <- c("Actual 0", "Actual 1")
  colnames(ConMatrix) <- c("Pred 0", "Pred 1")
  ConMatrix
}
confMat = ConfusionMatrix(classifier, as.integer(homicides.te$solved_status == "Solved"))
confMat
```

```
##          Pred 0 Pred 1
## Actual 0      0     74
## Actual 1      0    565
```

```
# Accuracy
accuracy = round(sum(diag(confMat))/sum(confMat)*100,3)
```

This model has 88.419% accuracy, however, this is because all new datapoints are predicted as being of class 1 and because most of the new datapoints are of class 1, this prediction is often accurate, as illustrated by the confusion matrix. The specificity of our model, which is an indicator of how many of the class 0 datapoints our model correctly predicts, can also be calculated to show this:

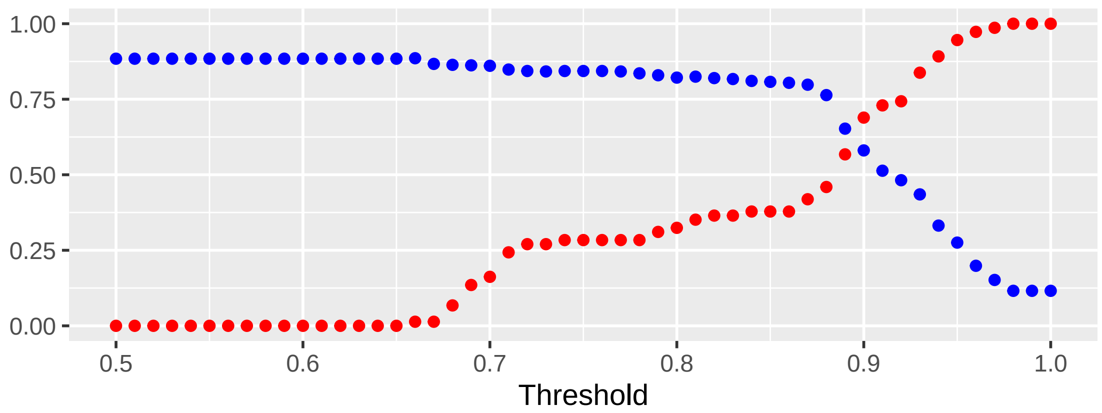
```
specificity = round(confMat[1,1]/(confMat[1,1] + confMat[1,2])*100,3)
```

The specificity of our model when using a threshold of 0.5 to classify each data points is 0, i.e., unsolved homicides are predicted as so 0% of the time. This would seem to indicate that our decision threshold of 0.5 is not adequate.

We can plot the accuracy and specificity of our model against different thresholds to choose a better threshold:

```
accs = numeric(length(seq(0.5, 1, 0.01)))
specificity = numeric(length(seq(0.5, 1, 0.01)))
j = 1

for(i in seq(0.5, 1, 0.01)){
  a_classifier <- preds[,"Estimate"]>=i
  conmat = ConfusionMatrix(a_classifier, as.integer(homicides.te$solved_status == "Solved"))
  accs[j] = sum(diag(conmat))/sum(conmat)
  specificity[j] = conmat[1,1]/(conmat[1,1] + conmat[1,2])
  j = j+1
}
ggplot()+geom_point(aes(x=seq(0.5, 1, 0.01), y = accs), color = "blue", show.legend = TRUE)+
  geom_point(aes(x=seq(0.5, 1, 0.01), y = specificity), color = "red", show.legend = TRUE)+
  xlab("Threshold")+
  ylab("")
```



Choosing a threshold of 0.88 leads to a much better specificity, although coming at the cost of accuracy. The confusion matrix using this threshold is:

```
classifier <- preds[,"Estimate"]>=bestThreshold
confMat = ConfusionMatrix(classifier, as.integer(homicides.te$solved_status == "Solved"))
accuracy = round(sum(diag(confMat))/sum(confMat)*100,3)
specificity = round(confMat[1,1]/(confMat[1,1] + confMat[1,2])*100,3)
confMat
```

```
##          Pred 0 Pred 1
## Actual 0     34     40
## Actual 1    111    454
```

which is much better than using a 0.5 threshold. It has 76.369% accuracy and 45.946% specificity. Therefore, our final model will classify new homicides as being solved if the probability estimate returned by our logistic regression is superior or equal to 0.88 and as being unsolved otherwise.

Critically evaluating the overall performance of the model

```
recall = round(confMat[2,2]/(confMat[2,1] + confMat[2,2])*100,3)
```

The model fit in the previous sections performs quite well. We have shown that the MCMC converges, the trace plots, rhat values and bulk/tail effective sample sizes are good. Furthermore, the accuracy of our final model is high which also indicates that the model performs well. There is however one downside to this model, which is that whereas most of the homicides that were solved were correctly predicted as being solved, demonstrated by a recall score of 80.354%, our model found it a lot harder to correctly classify unsolved homicides as being unsolved, as indicated by the low specificity score.

Therefore, if someone wanted to use this model to correctly predict whether a homicide will be solved or not, if the model predicted that a crime would be solved, they could be confident in this prediction. On the other hand, one should be more skeptical if the model predicts that a homicide will not be solved.

Inference

```
# obtain the raw samples for each of the population level coefficients
raw_samples = as_tibble(model)
# Calculate the corresponding probabilities of solving a homicide for each population level coefficient in the summary table
# as well as the probability that the effect is actually there (probability that the coefficient is greater than or less than 0)
int_prob = round(inv_logit_scaled(1.79)*100, 3)
blackCoef_prob = round(inv_logit_scaled(1.79-0.11)*100, 3)
blackCoef_under0 = round(mean(raw_samples$b_ethnBlack<0)*100, 3)
```

```
otherCoef_prob = round(inv_logit_scaled(1.79-0.06)*100, 3)

whiteCoef_prob = round(inv_logit_scaled(1.79+0.47)*100, 3)
whiteCoef_over0 = round(mean(raw_samples$b_ethnWhite>0)*100, 3)

domAbCoef_prob = round(inv_logit_scaled(1.79+0.5)*100, 3)

knifeCoef_prob =  round(inv_logit_scaled(1.79+0.47)*100, 3)
physCoef_prob = round(inv_logit_scaled(1.79+0.27)*100, 3)
shootCoef_prob = round(inv_logit_scaled(1.79-0.84)*100, 3)

maleCoef_prob = round(inv_logit_scaled(1.79-0.21)*100, 3)
maleCoef_over0 = round(mean(raw_samples$b_sexMale>0)*100, 3)
```

The summary of the model and the density plots of the posterior distributions of the estimates of each coefficient present in the model are displayed in the previous section. The intercept coefficient represents the estimate of the log-odds of a homicide being solved without any additional information. The summary shows that this coefficient is estimated to be 1.79. Translating this to a probabilistic interpretation means that our model predicts an average homicide to be solved 85.693‘% of the time. Furthermore, the posterior distribution of the intercept and the 95% credible interval shows that it is close to certain that the chances of solving any given crime are greater than 50% as close to the entirety of the distribution is after 0.

The estimate of the non-intercept coefficients are estimates of how much the log-odds of a homicide being solved increases/decreases depending on whether that level of the feature is present or not. More specifically, they determine how much the intercept will increase/decrease. For example, the estimate for b_2, corresponding to the coefficient which is activated when the victim is black, is -0.11. Indicating that when everything else remains constant, the intercept, i.e. the log-odds of an average homicide being solved, is decreased by -0.11 when the victim is black. Probabilistically, this means that a crime involving a black victim has an 84.29% chance of being solved. Looking at the density plot of the posterior distribution of b_2, we are quite confident that the victim being black will decrease the log-odds of a homicide being solved because there is a 70.325% probability that the real value of b_2 is lesser than 0. As previously mentioned, I expected it to be less likely for the homicide of a black person to be more likely to go unsolved due to racial biases that are present within law enforcement agencies.

The log-odds is expected to decrease by the same amount for a victim who's ethnicity is unreported or not known, as it is for a black victim. Indicating that the probability of a homicide being solved, where the victim's ethnicity is unknown, is the same as for a homicide with a black victim.

In addition, the victim being of another ethnicity also has a negative effect on the probability of the homicide being solved, although not as strong as for a black victim or for a victim who's ethnicity is not known. The log-odds in this situation is expected to decrease by -0.06, corresponding to an estimated probability of 84.941 of a homicide being solved if the victim is of ethnicity "Other". However, the density plot of the posterior distribution shows that we are not certain if the effect of this on the log-odds is positive or negative and that it is also not very certain that the victim being of another ethnicity has any effect on the chances of a homicide being solved as the mean looks to be quite close to 0.

On the other hand, if the victim is white, the log-odds of a homicide being solved are estimated to increase by 0.47, meaning that a white person's homicide has a probability of around 90.551% of being solved according to the model. As the posterior distribution shows, we are quite certain that the victim being white has a positive effect on a homicide being solved as there is 98.415% probability that the real value of b_5 is greater than 0. The interpretation for this is the same as the one offered for which a black person's homicide is less likely to be solved than the average. Indeed, racial biases present in law enforcement make it so that there is more incentive to solve the homicide of a white person rather than that of another ethnicity.

Moreover, homicides where domestic abuse is involved have a much greater chance of being solved compared to homicides where it was not involved. As shown by the summary table, the log-odds of a homicide being solved are estimated to increase by 0.5 if domestic abuse is involved. This means that there is an estimated 90.805% chance of a homicide being solved if domestic abuse is involved. The posterior distribution's density plot indicates that we are quite confident that there is a positive effect with regards to the involvement of domestic abuse on the chances of solving a homicide as most of the distribution is greater than 0, meaning there is a very high probability that b_7 is greater than 0. A possible explanation of this effect could be that if it is known that the victim was also subject to domestic abuse, then there would be a relatively obvious suspect as to who committed the homicide, making it easier for law enforcement to solve the homicide.

Additionally, every level of the method of killing factor seem to have a significant effect on the chances of solving a homicide. Indeed, the log-odds of solving a homicide are estimated to go up by 0.47 when the victim was murdered using a knife or a sharp object, by 0.27 when the victim died from physical assault, and decrease by 0.84 when the victim was killed by a gun. This corresponds to estimated probabilities of 90.551%, 88.695% and 72.112%, respectively, for solving a homicide. The plots of the respective posterior distributions indicate that the model is confident that these effects are present as 0 is only present at the tail of the distributions, indicating that there is a high probability that these effects are present. Although it does seem strange that using a gun leads to a much lower probability of solving a homicide than using a knife or another sharp object, this could be explained by the fact that one is more likely to get rid of the gun in the aftermath whereas the criminal might not think to dispose of the knife/sharp object, making it easier for police to find the object used to murder and therefore retrieve DNA prints. On the other hand, it makes sense that someone who killed a person with physical assault has higher chances of being found because there would be DNA evidence present on the victim.

Furthermore, our model shows that whether or not the victim is male also has an effect on the probability of solving a homicide. Although we are not too certain of this effect, as there is 30.435% probability that the real value of $b_1 3$ is over 0, and 69.565% probability that it is under it, it seems to be that being a male decreases the chances of solving a homicide. Indeed, our model estimates these chances to be around 82.92, which is 2.773% less than the chances of solving the average homicide. This relationship may be due to the fact that because homicides involving female victims are much rarer than those involving male victims, there might be more incentive to solve them. This effect could also be due to the nature of the deaths themselves. For example, male victims may be more likely to have been involved in gang related activities and therefore harder to solve. Whereas female deaths may be more likely to be due to domestic violence, and as previously established, when domestic violence is involved in a homicide, it is more likely to be solved. The interaction between domestic abuse and sex of the victim on the probabilities of a homicide being solved was fit in this model and will be discussed later.

Looking at the posterior distributions for the coefficients relevant to the poverty level of the borough in which the crime was committed, we can see that we are quite uncertain about whether or not there is an effect of this on the odds of solving a crime as the means are quite close to 0. The means for both a "rich" and "middle" poverty level borough are inferior to 0, indicating that the probability of solving a homicide is greater if committed in a poorer borough. However, for both of the posteriors there is only around 60% probability that the real values of the coefficients are lower than 0, meaning that we are not sure about the direction of the effect, or if there is even an effect to start with.

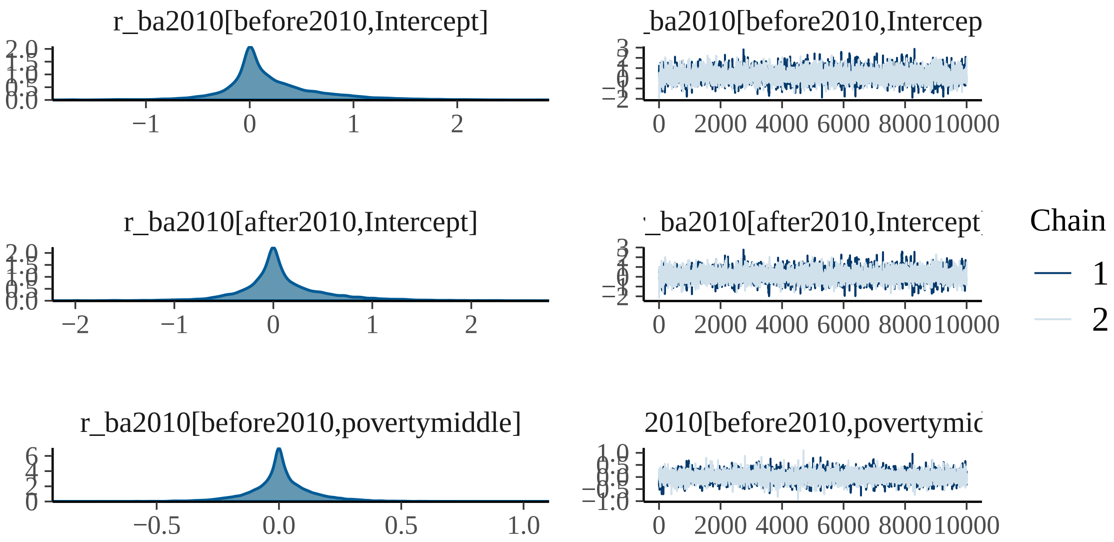
Similarly, the posterior distribution for the coefficient which is activated when the homicide was committed after 2010 shows that the estimated mean is very close to 0, which is reinforced by the 95% credible interval as well as the estimate for the coefficient. The fact that whether the homicide was committed before or after 2010 is contrary to the beliefs I had that homicides would be less likely to be solved after 2010 due to the conservative party defunding the police.

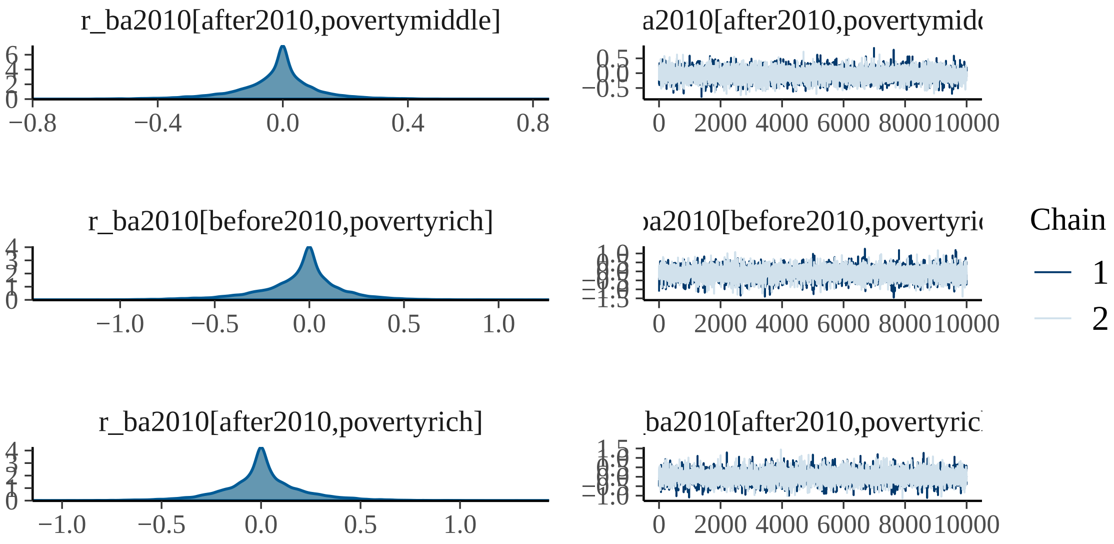
We can now look at the grouping effects:

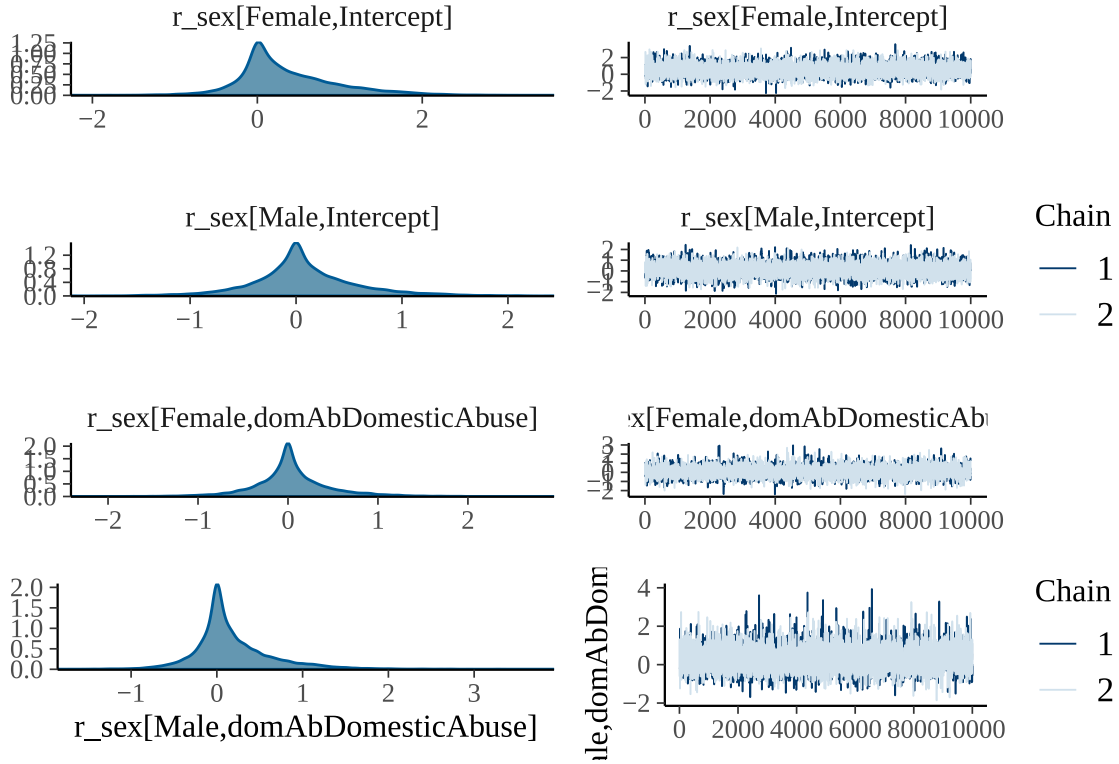
```
ranef(model)
```

```
## $ba2010
## , , Intercept
##
##              Estimate Est.Error       Q2.5    Q97.5
## before2010 0.19524475 0.4679874 -0.5971656 1.370519
## after2010  0.08886787 0.4515326 -0.7369813 1.206241
##
## , , povertymiddle
##
##                Estimate Est.Error       Q2.5     Q97.5
## before2010  0.008497312 0.1357952 -0.2735876 0.3121932
## after2010  -0.019867129 0.1315443 -0.3256111 0.2456982
```

```
##
## , , povertyrich
##
##                 Estimate Est.Error       Q2.5     Q97.5
## before2010 -0.047105318 0.2268805 -0.6073065 0.3855337
## after2010   0.008572733 0.2150533 -0.4455848 0.4812312
##
##
## $sex
## , , Intercept
##
##           Estimate Est.Error       Q2.5    Q97.5
## Female 0.35803225 0.5955742 -0.5867461 1.819132
## Male   0.07420427 0.4871115 -0.8908947 1.221927
##
## , , domAbDomesticAbuse
##
##           Estimate Est.Error       Q2.5    Q97.5
## Female 0.04193239 0.4314269 -0.8208381 1.051372
## Male   0.17476543 0.4600890 -0.5844918 1.319297
```







```
# Corresponding probability of solving a homicide for the coefficient of the before 2010 group intercept to which the "poor"
# level of the poverty level of the borough factor was aliased
poorBefore2010_prob = round(inv_logit_scaled(1.79+0.195)*100, 3)
# Corresponding probability of solving a homicide for the coefficient of the female group intercept to which the "Not Domestic
# Abuse" level of the was domestic abuse involved factor was aliased
notDomAbFemale_prob = round(inv_logit_scaled(1.79+0.385)*100, 3)
# Corresponding probability of solving a homicide for the coefficient of a male victim where domestic abuse was involved
DomAbMale_prob = round(inv_logit_scaled(1.79+0.5-0.21+0.175)*100, 3)
```

```
# Calculate the probability that coefficient for the before 2010 group intercept is greater than 0
ranef_draws = spread_draws(model,r_ba2010[before2010,Intercept])[spread_draws(model,r_ba2010[before2010,Intercept])$before2010 == "before2010",]
ranef_draws = ranef_draws[ranef_draws$Intercept == "Intercept",]
poorBefore2010_over0 = round(mean(ranef_draws$r_ba2010>0)*100, 3)
# Calculate the probability that coefficient for the female group intercept is greater than 0
ranef_draws = spread_draws(model,r_sex[Female,Intercept])[spread_draws(model,r_sex[Female,Intercept])$Female == "Female",]
ranef_draws = ranef_draws[ranef_draws$Intercept == "Intercept",]
notDomAbFemale_over0 = round(mean(ranef_draws$r_sex>0)*100, 3)
# Calculate the probability that coefficient for a male victim where domestic absue was involved is greater than 0
ranef_draws = spread_draws(model,r_sex[Female,Intercept])[spread_draws(model,r_sex[Female,Intercept])$Female == "Male",]
ranef_draws = ranef_draws[ranef_draws$Intercept == "domAbDomesticAbuse",]
DomAbMale_over0 = round(mean(ranef_draws$r_sex>0)*100, 3)
```

The summary table for the effects of the grouping variables and the density plots of the posterior distributions seem to indicate that there are only 3 grouping coefficients that have a relatively high probability of having an effect on the probability of solving a homicide. The first is for the intercept of the before 2010 group, to which the "poor" level of the poverty level factor of the borough in which the homicide was committed has been aliased. The coefficient for this is 0.195, which means that a homicide that was committed before 2010 and in a poor borough has a 87.921% chance of being solved. From the density of the posterior distribution, there is 65.96% probability that this effect is present, therefore it is not really sure that this effect is indeed present. Despite this, an interpretation can still be offered. As the conservation party came into power in 2010 and I believe that they do not care for the poor as much as the labor party, I believe it makes sense that there was a greater chance of solving a homicide in a poor borough before the conservative party came into power.

Furthermore, it also seems that if the victim is female and domestic abuse was not involved, then there is a higher chance of solving the homicide than if the victim was male. The estimate of the coefficient corresponding to a male victim and no domestic abuse is quite close to 0, on the other hand, the estimated coefficient for a female victim and no domestic abuse is 0.358. This means that there is a 89.798% of solving the homicide in this scenario. The posterior distribution's density plot indicates that we can be quite confident that there is a positive effect as 72.81% of the distribution is greater than 0, i.e. there is a 72.81% chance that this coefficient is positive. As discussed previously, the homicide of females is much less common than that of males, meaning that there may be more "incentive" to solve their murders, which would explain this positive relationship.

On the other hand, it seems that if a male is the victim, and domestic abuse is involved, his homicide is more likely to be solved than if the victim was a female. The estimated coefficient corresponding to this scenario is 0.175, which translates to there being a 90.508% of a homicide being solved if the victim is male and domestic abuse is involved. Once again, there is only 63.37% probability that this coefficient is greater than 0 and therefore I am not too confident about the positive effect.

The other grouping effects do not seem to have an effect on the probability of a homicide being solved or not. Indeed, their mean estimates are close to 0 and from looking at the plots of the posterior distributions, it looks like these are symmetrical around 0, i.e. there is a 50% chance that the effect is positive or negative.

Monte Carlo Estimation

Code to generate two new homicides:

```
curated_cols <- c("recorded_date","age_group","sex","observed_ethnicity","domestic_abuse",
                  "borough","method_of_killing")
new_dates <- as.Date(c("2022-04-01", "2022-05-01", "2022-06-01",
                       "2022-07-01", "2022-08-01", "2022-09-01",
                       "2022-10-01", "2022-11-01", "2022-12-01"))
curated_homs <- dplyr::select(homicides, all_of(curated_cols))
hypothetical_homicides <- tibble(recorded_date = sample(new_dates, 2, TRUE))
month_tibble <- read_csv("month_tibble.csv", show_col_types = FALSE)
for(i in 2:length(names(curated_homs))){
  hypothetical_homicides <- cbind(hypothetical_homicides,
                                  sample(as.vector(unlist(unique(curated_homs[,i]))),2,TRUE))
}
names(hypothetical_homicides) <- names(curated_homs)
hypothetical_homicides <- as_tibble(hypothetical_homicides) %>%
  left_join(month_tibble, by = "recorded_date")
```

Using the clean_df() function to perform the same operations on this hypothetical homicides dataframe as on the original homicides dataset. The two new homicides are:

```
hypothetical_homicides = clean_df(hypothetical_homicides)
hypothetical_homicides$borough_class_income = borough_poverty_group(hypothetical_homicides)
hypothetical_homicides = hypothetical_homicides %>% rename(
  poverty = borough_class_income
)
hypothetical_homicides[,c(1,2,3,4,5,6,7,11,12, 8,9,10)]
```

```
## # A tibble: 2 x 12
##   recorded_date age_group sex   ethn  domAb borough methOfK ba2010 poverty  year
##   <date>        <fct>     <fct> <fct> <fct> <fct>   <fct>   <fct>  <fct>   <dbl>
## 1 2022-08-01    20 to 24  Fema~ Not ~ Not ~ Richmo~ Knife ~ after~ rich       20
## 2 2022-04-01    45 to 54  Male  Not ~ Not ~ Isling~ Shooti~ after~ rich       20
## # ... with 2 more variables: month.name <fct>, season <fct>
```

We are asked to estimate $P(A \wedge \tilde{B}|data)$, the probability of homicide A being solved and homicide B not being solved given the data. These lines of code retrieve a 20000 x 2 dataframe of posterior draws from $logit(p)$ as defined in the model definition section to which have been applied the inverse link function $\frac{1}{1+e^{-x}}$, where each row corresponds to the probability of homicide A and B being solved individually for that simulation.

```
preds = as_tibble(posterior_linpred(model, transform = TRUE, newdata = hypothetical_homicides))
preds = preds %>% rename("A" = "V1", "B" = "V2")
```

As described previously, a homicide will be predicted as being solved if the probability returned by the logistic regression is greater than 0.88. The following lines of code add two columns to the predictions dataframe containing the classification of each homicide for every simulation.

```
preds = preds %>% mutate(classA = ifelse(A > bestThreshold, "Solved", "Unsolved")) %>%
  mutate(classB= ifelse(B > bestThreshold, "Solved", "Unsolved"))
head(preds)
```

```
## # A tibble: 6 x 4
##       A     B classA classB
##   <dbl> <dbl> <chr>  <chr>
## 1 0.937 0.744 Solved Unsolved
## 2 0.919 0.578 Solved Unsolved
## 3 0.905 0.670 Solved Unsolved
## 4 0.929 0.658 Solved Unsolved
## 5 0.931 0.582 Solved Unsolved
## 6 0.952 0.745 Solved Unsolved
```

Now to estimate $P(A \wedge \tilde{B}|data)$:

```
prob_A_andNot_B = round(mean(preds$classA=="Solved" & preds$classB=="Unsolved"),3)
MCMCerror = sd(preds$classA=="Solved" & preds$classB=="Unsolved")/sqrt(4289)
```

And so we have that an estimate for $P(A \wedge \tilde{B}|data)$ is around 0.852, i.e. there is around 85.2% probability of homicide A being solved and not homicide B given the data. The Monte Carlo error is given by the $\frac{s}{\sqrt{n_{eff}}}$ where s is the standard deviation of the estimate and n_{eff} the smallest effective sample size in the summary table, which is 4289 here. Therefore the Monte Carlo error is 0.0054299.