MTHM508_Coursework_II

2023-01-07

1. Present a single Bayesian analysis consisting of a hierarchical model, fit using brms, for the event that a particular homicide is solved given the values of the relevant covariates in this data set. Load the required libraries and the data.

```
# Load libraries - suppress messages
suppressMessages(library(bayesplot))
suppressMessages(library(brms))
suppressMessages(library(reshape2))
suppressMessages(library(dplyr))
suppressMessages(library(ggplot2))
suppressMessages(library(arm))
suppressMessages(library(boot))
# Read data
homicides = read.csv("homicides.csv")
# train test split
homicides_train = filter(homicides,year<15)
homicides_test = filter(homicides,year>=15)
#look at raw data
head(homicides_train, n=2)
```

```
##   recorded_date age_group  sex observed_ethnicity     domestic_abuse
## 1      1/1/2003  25 to 34 Male              White Not Domestic Abuse
## 2      1/1/2003  45 to 54 Male              White Not Domestic Abuse
##       borough        method_of_killing solved_status year month.name season
## 1 Westminster Knife or Sharp Implement        Solved    1        JAN winter
## 2    Haringey Knife or Sharp Implement        Solved    1        JAN winter
```

Of the 20 years in the dataset, I will consider the first 15 as training and reserve the last 5 for testing.

```
# train test split
homicides_train = filter(homicides,year<15)
homicides_test = filter(homicides,year>=15)
```

Some functions to 'binarize' the data are applied, as most of them are categorical (i.e. factors). If the crime was gun-related, it is grouped as "1" and if not, it is grouped as "0". This new variable is named "shooting".

```
# Functions

# binarize categorical variables
binarize = function(df,col,set_to_one,other_encoding){
  df[col]=ifelse(df[col]==set_to_one,1,other_encoding)
  return(df)
}
# capture if ethnicity has been reported (modififed from  MTHM508 Pima Indians lectures)
ethinicity_reported = function(vec){
  ans <- as.character(vec)
  for(i in 1:length(vec)){
    if(vec[i] == 'Not Reported/Not known'){
      ans[i] = 0
    }
    else{
      ans[i] = 1
    }
  }
  return(as.numeric(ans))
}
# Pipeline Function for cleaning and converting data types
pipeline = function(df,y_exists=TRUE){
  # Don't need recorded date -> year and month.name suffice
  df = tibble(subset(df, select = -c(recorded_date) ))
  df$eth_rep = ethinicity_reported(df$observed_ethnicity)
  df$shooting = as.integer(df$method_of_killing=='Shooting')
  # binarize the solved status
  if(y_exists){
    df = binarize(df=df, col='solved_status', set_to_one='Solved', other_encoding=0)
  }
  df = binarize(df=df, col='sex', set_to_one='Male', other_encoding=-1)
  df = binarize(df=df, col='domestic_abuse', set_to_one='Domestic Abuse', other_encoding=-1)
  return(df)
}

#___________________________________________________________
# transformations
homicides_1 = pipeline(homicides_train)
homicides_2 = pipeline(homicides_test)
# look at data
head(homicides_1,n=2)
```

```
## # A tibble: 2 x 12
##   age_group   sex observe~1 domes~2 borough metho~3 solve~4  year month~5 season
##   <chr>     <dbl> <chr>       <dbl> <chr>   <chr>     <dbl> <int> <chr>   <chr>
## 1 25 to 34      1 White          -1 Westmi~ Knife ~       1     1 JAN     winter
## 2 45 to 54      1 White          -1 Haring~ Knife ~       1     1 JAN     winter
## # ... with 2 more variables: eth_rep <dbl>, shooting <int>, and abbreviated
## #   variable names 1: observed_ethnicity, 2: domestic_abuse,
## #   3: method_of_killing, 4: solved_status, 5: month.name
```

This is a classfication task, so the model will be Logistic Regression. The model structure is

$$y_{ij}|\eta_{ij},\theta \sim \mathcal{D}(g^{-1}(\eta_{ij}),\theta)$$

where g is the 'link function'. In this case, it is the *logit* function, such that $g^{-1}(\eta) = \frac{1}{1+e^{-\eta}}$ The model for logistic regression dows not bear any θ parameter, so I will not use it going forward.

$$\eta_j = b_0 + b_1 x_{1j} + b_2 x_{2j} + b_3 x_{3j} + b_1 x_{4j} + \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j}$$

$$\beta_j \sim \mathcal{N}(0,\Sigma)$$

$$\pi(b,\Sigma)$$

Some factors that seem likely to influence the target variable can be sex, domestic abuse and whether the homicide was gun-related or not. I am also including whether ethnicity was reported or not, but I may take it out later.

Justification for Priors Here, I will sequentially look at setting the priors for the model.

Intercept prior I look at the data to judge the level/ average rate of solving homicides, which is around 0.89. Opting for something weakly informative is advisable as there is no weight of literature behind setting these priors. As a result, choosing an intercept prior that is allowed to reach 0.89 (when put through the inverse logit function) as an extreme observation (~2 s.d.) can be reasonable. Thus, a normal distribution with mean 0 can be allowed to have 2 s.d. $= logit(0.8944) =$ 2.137. The following prior seems appropriate

$$\mathcal{N}(0, 1.2)$$

This yields a 2 s.d of ~2.19.

The following are all originally categorical variables encoded as 1 and 0. Their priors will be similar to each other and different from the intercept prior.

Sex prior Males are encoded as 1 and females are encoded as -1. Considering the intercept, a coefficient that is allowed to vary till 0.5 will roughly yield somewhere around $g^{-1}(2.19 + 0.5 * 1) = 0.93$ and $g^{-1}(2.19 - 0.5 * 1) = 0.84$, which is within bounds I would expect. Hence, following the same calculation as above, I choose a prior of

$$\mathcal{N}(0, 0.0625)$$

Domestic abuse prior When domestic abuse is involved in the homicide, I hypothesize that it is 'easier' to find the perpetrator. Living in the same household must greatly narrow down the search. That being said, I would not be comfortable with placing a very wide prior on the coefficient, as it can potentially take the response variable to above 0.99 chance of being solved - that seems unreasonable. Looking at the data, it appears that domestic abuse cases get solved ~96% of the time. $g(0.96) = 3.18$. Considering the intercept of around 2.19, that is an increment of ~0.987. Let's round that to 1 so that the distribution is allowed to go up to 1 roughly 95% of the time (i.e. ~2 s.d.) This lands us with a prior

$$\mathcal{N}(0, 0.25)$$

Shooting prior It can be reasoned that shootings make it relatively harder to solve the case as compared to physical assaults or stabbings. The latter two are more likely to leave behind DNA evidence or other identifiers at the scene, whereas a shooting can leave behind as little as a bullet shell (or casing), which may or may not be found at the crime scene. This is the reason why the shootings and non-shootings have been grouped separately, and are being considered in the first fit of the model.

As for the prior, I can believe the probability of solving slips down to ~75% if the crime was gun-related. In this scenario, $2.19 - g^{-1}(0.75) = 1.1$, i.e. the prior should be allowed to reach 1.3 in the 95th percentile (of its distribution). Therefore, $s.d. = \frac{1.1}{2} = 0.55$. The prior is thus

$$\mathcal{N}(0, 0.3025)$$

Group priors:

Priors for standard deviation of β_j are next. Left alone, Stan would revert to a half-student_t(3, 0, 2.5). This can seem a little too wide. Suppose, for the intercept, which is already around 2.19, having a s.d. on top of that, reaching upto $\sqrt{2.5} * 2 = 3.16$ would mean passing ~$2.19 + 3.16$ through the inverse logit function, bringing it very near to 1 (albeit this would be a highly improbable scenario). I would like to keep the s.d. priors tight around their original b priors. A zero-mean normal distribution with small variance (~0.1) for the b priors and slightly larger (~0.5) for Intercept prior sounds plausible.

$$\mathcal{N}(0, 0.1) \ ; \ \mathcal{N}(0, 0.5)$$

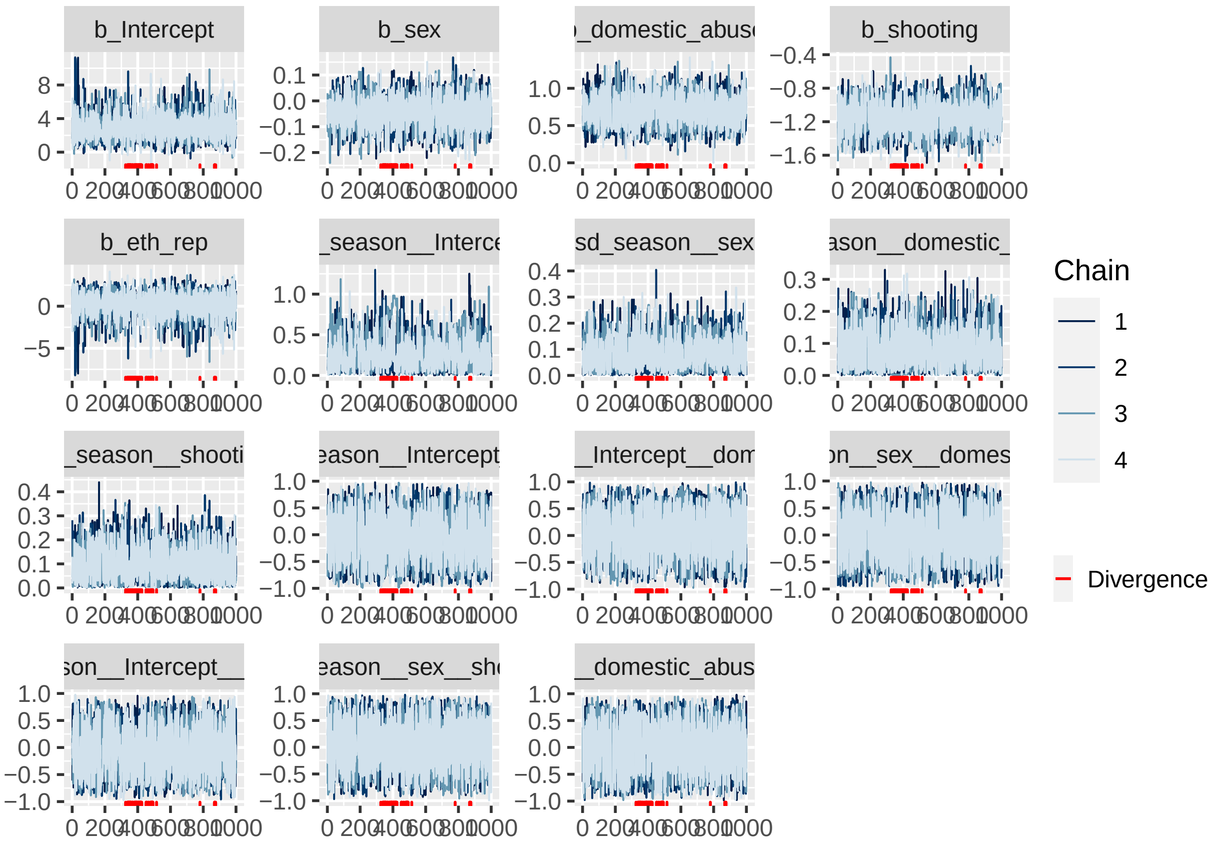
```
# Set priors
intercept_prior = set_prior('normal(0,1.2)',class='Intercept')
sex_prior = set_prior('normal(0,0.0625)',class='b',coef='sex')
abuse_prior = set_prior('normal(0,0.4225)',class='b',coef='domestic_abuse')
shooting_prior = set_prior('normal(0,0.3025)',class='b',coef='shooting')

sd_priors_I <- set_prior("normal(0,0.5)",class="sd", group="season", coef='Intercept')
sd_priors <- set_prior("normal(0,0.1)",class="sd", group="season", coef = c('sex','domestic_abuse','sho

# Join priors into a vector
my_prior = c(intercept_prior,sex_prior,abuse_prior,shooting_prior,sd_priors_I, sd_priors)

fit_1 = brm(solved_status~sex+domestic_abuse+shooting+eth_rep
            +(sex+domestic_abuse+shooting|season),
          data=homicides_1,
          family=bernoulli,
          prior=my_prior,
          refresh=0)
```

```
## Compiling Stan program...
```

```
## Start sampling
```

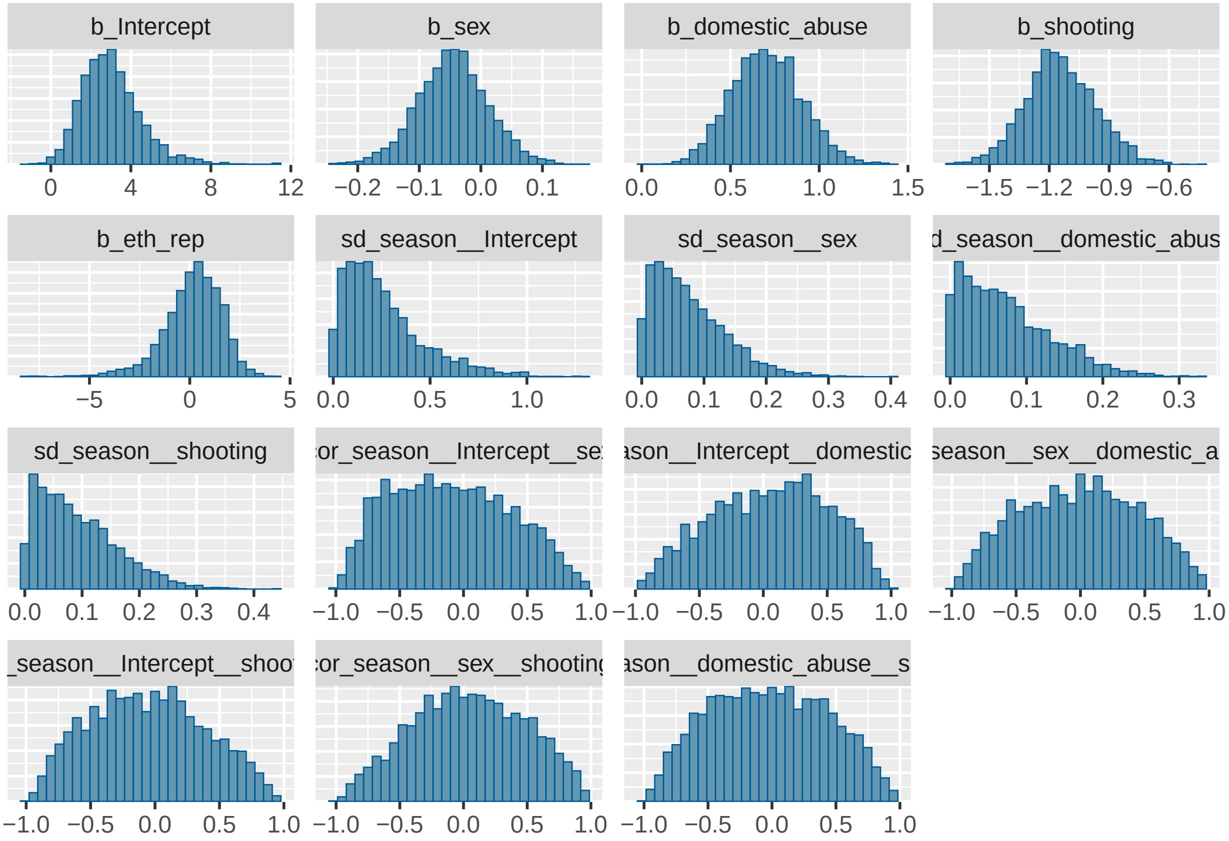
Trace plots are examined.

```
mcmc_plot(fit_1 , type='trace')
```



```
mcmc_plot(fit_1 , type='hist')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

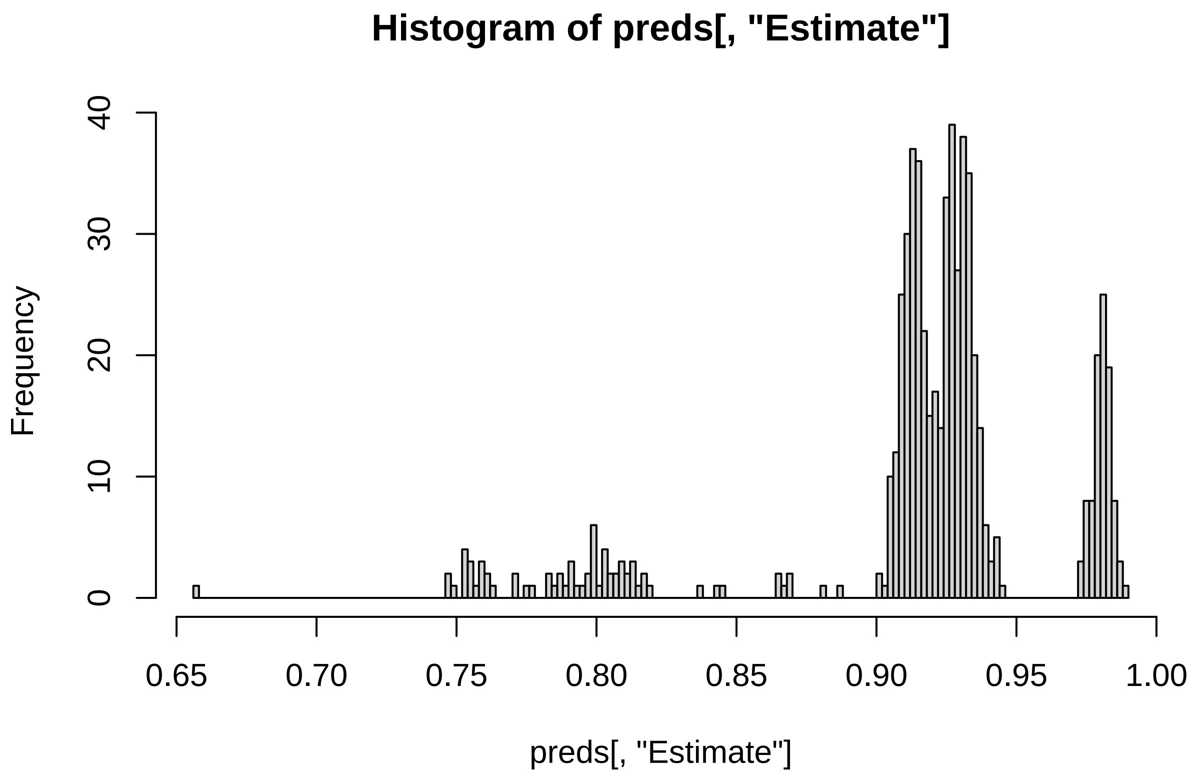


The traceplots suggest that convergence has been reached. To judge whether model has adequately fit the data, I will use it to predict over unseen data (the 5 years of data stowed away at the beginning).

```
# Validation (taken from MTHM508 Pima Indians lectures)
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 Negative", "Actual Positive")
  colnames(ConMatrix) <- c("Pred Negative", "Pred Positive")
  return (ConMatrix)
}

preds <- predict(fit_1, newdata=homicides_2)
```

The threshold can be iteratively changed to see which yields the best confusion matrix (with least Type-1 and Type-2 errors).

```
hist(preds[,"Estimate"],breaks=200)
```



I will keep the threshold at 0.89. This yields an accuracy of ~82%

```
# Use a sensible threshold value
a_classifier <- preds[,"Estimate"]>=0.89
conmat <- ConfusionMatrix(a_classifier, as.integer(homicides_2$solved_status==1))
conmat
```

```
##                   Pred Negative Pred Positive
## Actual Negative              28            65
## Actual Positive              44           472
```

```
sum(diag(conmat))/sum(conmat)
```

```
## [1] 0.8210181
```

Critical Evaluation of model performance Since there are 1904 solved cases amidst the 2130 recorded homicides, having a 'dumb' classifier predict every crime as 'solved' would lead to ~89% accuracy over the dataset. The fitted model does not achieve that accuracy level. However, one thing it is being able to do is correctly classify at least some of the cases which were not actually solved.

2. Use your model to infer how the features of any particular homicide in London affect the probability that the case has been solved (to date).

Looking at the summary of the fit, I can deduce the following:

```
summary(fit_1)
```

```
##  Family: bernoulli
##   Links: mu = logit
## Formula: solved_status ~ sex + domestic_abuse + shooting + eth_rep + (sex + domestic_abuse + shooting
##    Data: homicides_1 (Number of observations: 1521)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Group-Level Effects:
## ~season (Number of levels: 4)
##                                Estimate Est.Error l-95% CI u-95% CI Rhat
## sd(Intercept)                      0.26      0.20     0.01     0.79 1.02
## sd(sex)                            0.08      0.06     0.00     0.22 1.01
## sd(domestic_abuse)                 0.07      0.06     0.00     0.21 1.01
## sd(shooting)                       0.09      0.07     0.00     0.25 1.00
## cor(Intercept,sex)                -0.10      0.46    -0.85     0.76 1.01
## cor(Intercept,domestic_abuse)      0.06      0.45    -0.79     0.82 1.00
## cor(sex,domestic_abuse)           -0.01      0.45    -0.82     0.80 1.01
## cor(Intercept,shooting)           -0.06      0.44    -0.82     0.77 1.00
## cor(sex,shooting)                  0.04      0.44    -0.78     0.84 1.00
## cor(domestic_abuse,shooting)      -0.02      0.45    -0.83     0.80 1.00
##                                Bulk_ESS Tail_ESS
## sd(Intercept)                       302      165
## sd(sex)                            1055     1619
## sd(domestic_abuse)                  413      119
## sd(shooting)                       1761     1928
## cor(Intercept,sex)                 1267     2417
## cor(Intercept,domestic_abuse)      1348     1188
## cor(sex,domestic_abuse)             931     1000
## cor(Intercept,shooting)            3063     2558
## cor(sex,shooting)                  3129     3099
## cor(domestic_abuse,shooting)       1528     1629
##
## Population-Level Effects:
##                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept          3.01      1.50     0.61     6.54 1.00     1790     1663
## sex               -0.05      0.06    -0.16     0.07 1.00     3998     2948
## domestic_abuse     0.70      0.20     0.33     1.10 1.01      991      678
## shooting          -1.14      0.18    -1.48    -0.79 1.00     2400     1755
## eth_rep            0.15      1.46    -3.29     2.49 1.00     3734     1552
##
## 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).
```

The sex of the victim as well as whether the crime was gun-related or not significantly contribute to lowering the chance of the crime being solved. Recall that sex is encoded as 1 or male and -1 for female. Similarly, shooting is encoded as 1 for death by shooting and -1 for death by other methods. A negative coefficient for sex is saying that a crime is more likely to be solved if the victim is a woman. For shooting, the negative coefficient backs up our earlier hypothesis of shooters being relatively more difficult to track down. A decently high positive coefficient on domestic abuse points towards crimes having a higher probability of being solved if it was committed by someone who knew the victim personally or lived in the same household. I think this makes sense as it would cut down on a lot of guesswork for law enforcement and narrow the search to a few possible suspects.

Investigate the group level effects.

```
ranef(fit_1)
```

```
## $season
## , , Intercept
##
##            Estimate Est.Error       Q2.5     Q97.5
## autumn -0.03487684 0.2510678 -0.4722183 0.6674262
## spring  0.16337192 0.2973759 -0.2302115 1.0751432
## summer  0.21438193 0.3266631 -0.1531405 1.2728075
## winter  0.01336951 0.2496383 -0.4246773 0.7114502
##
## , , sex
##
##            Estimate  Est.Error       Q2.5     Q97.5
## autumn -0.01995478 0.08384443 -0.2159153 0.1375168
## spring -0.02664128 0.08809011 -0.2491048 0.1231836
## summer -0.02889447 0.09673184 -0.2688048 0.1293667
## winter -0.03262680 0.09261245 -0.2635908 0.1191194
##
## , , domestic_abuse
##
##            Estimate  Est.Error       Q2.5     Q97.5
## autumn  0.00461985 0.08709022 -0.1813113 0.1947153
## spring  0.00913092 0.08720588 -0.1671973 0.2191715
## summer -0.01544943 0.09202368 -0.2396356 0.1589314
## winter  0.02971771 0.09092672 -0.1282500 0.2682736
##
## , , shooting
##
##            Estimate Est.Error       Q2.5     Q97.5
## autumn -0.06433048 0.1282057 -0.4071453 0.1136594
## spring -0.05020581 0.1167297 -0.3505431 0.1271242
## summer -0.03636694 0.1142630 -0.3208302 0.1508584
## winter -0.02857531 0.1127005 -0.3225194 0.1803714
```

The only noticeable deviations from the population means I can see is for the Intercept, which goes to say that murders happening in summer have a slightly higher chance of being solved. I reserve commenting on crimes in spring, autumn and winter, because even if it looks like they have slightly lower probability of being solved, their distribution is placed pretty evenly on both sides of 0. I am not sure of whether this effect actually exists or not. Checking the posterior samples will help narrow down those covariates that may actually have an effect on the response variable.

```
samples = as_draws_df(fit_1)
eth_rep_samples = samples$b_eth_rep
sum(eth_rep_samples<0)/length(eth_rep_samples)
```

```
## [1] 0.41025
```

This says that there is ~40% chance that the mean effect of reported ethnicity is of the wrong sign - this is not reliable. **We will remove it from our model.** Before that, let's check a few other posterior samples.

```
# Probability that the effect of shooting is actually positive
shooting_samples = samples$b_shooting
```

```r
sum(shooting_samples>0)/length(shooting_samples)
```

```
## [1] 0
```

```r
# Probability that the effect of abuse is actually negative
abuse_samples = samples$b_domestic_abuse
sum(abuse_samples<0)/length(abuse_samples)
```

```
## [1] 0.00025
```

```r
# Probability that the effect of being male is actually positive
sex_samples = samples$b_sex
sum(sex_samples>0)/length(sex_samples)
```

```
## [1] 0.206
```

```r
for(c in colnames(samples[,16:31])){# ranef posteriors begin at 16 and end at 31
  r_posterior = samples[c]
  gt_zero = sum(r_posterior>0)/nrow(r_posterior)
  print(paste(c , "proba of >0", gt_zero , ", proba of <=0", 1 - gt_zero))
}
```

```
## [1] "r_season[autumn,Intercept] proba of >0 0.37675 , proba of <=0 0.62325"
## [1] "r_season[spring,Intercept] proba of >0 0.71725 , proba of <=0 0.28275"
## [1] "r_season[summer,Intercept] proba of >0 0.7855 , proba of <=0 0.2145"
## [1] "r_season[winter,Intercept] proba of >0 0.4715 , proba of <=0 0.5285"
## [1] "r_season[autumn,sex] proba of >0 0.4185 , proba of <=0 0.5815"
## [1] "r_season[spring,sex] proba of >0 0.39625 , proba of <=0 0.60375"
## [1] "r_season[summer,sex] proba of >0 0.3965 , proba of <=0 0.6035"
## [1] "r_season[winter,sex] proba of >0 0.38775 , proba of <=0 0.61225"
## [1] "r_season[autumn,domestic_abuse] proba of >0 0.5425 , proba of <=0 0.4575"
## [1] "r_season[spring,domestic_abuse] proba of >0 0.5295 , proba of <=0 0.4705"
## [1] "r_season[summer,domestic_abuse] proba of >0 0.45 , proba of <=0 0.55"
## [1] "r_season[winter,domestic_abuse] proba of >0 0.6185 , proba of <=0 0.3815"
## [1] "r_season[autumn,shooting] proba of >0 0.32 , proba of <=0 0.68"
## [1] "r_season[spring,shooting] proba of >0 0.34625 , proba of <=0 0.65375"
## [1] "r_season[summer,shooting] proba of >0 0.387 , proba of <=0 0.613"
## [1] "r_season[winter,shooting] proba of >0 0.4055 , proba of <=0 0.5945"
```

From this, it is clear that shooting, sex and domestic abuse have a significant effect on the response variable. Looking at the random effects, none of the covariates are that effected across different seasons (apart from Intercept which increases in summer, i.e. there is ~20% chance of there not acutally being an increase, which is far less than the other group effects). A second model can be fit by removing eth_rep and grouping on something other than season. I pick observed_ethnicity as it seems like there may be group-wise variability in how 'solvable' a homicide will be. I don't believe sex, domestic abuse or shooting will be much affected, hence I will remove them from grouped effects.

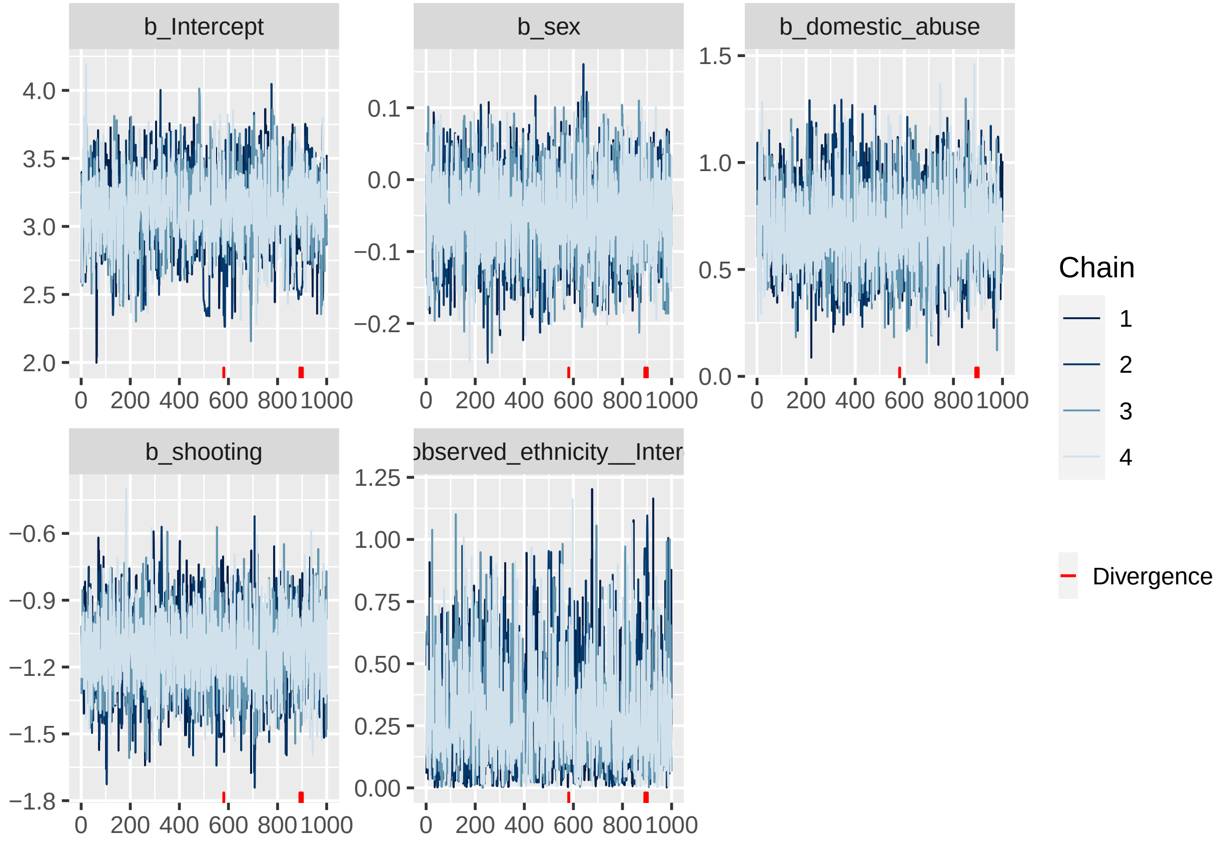
```r
sd_priors_I <- set_prior("normal(0,0.5)",class="sd", group="observed_ethnicity", coef='Intercept')
# Join priors into a vector
my_prior = c(intercept_prior,sex_prior,abuse_prior,shooting_prior,sd_priors_I)

fit_2 = brm(solved_status~sex+domestic_abuse+shooting
              +(1|observed_ethnicity),
            data=homicides_1,
            family=bernoulli,
            prior=my_prior,
            refresh=0,
            iter=2000)
```

```
## Compiling Stan program...
```

```
## Start sampling
```

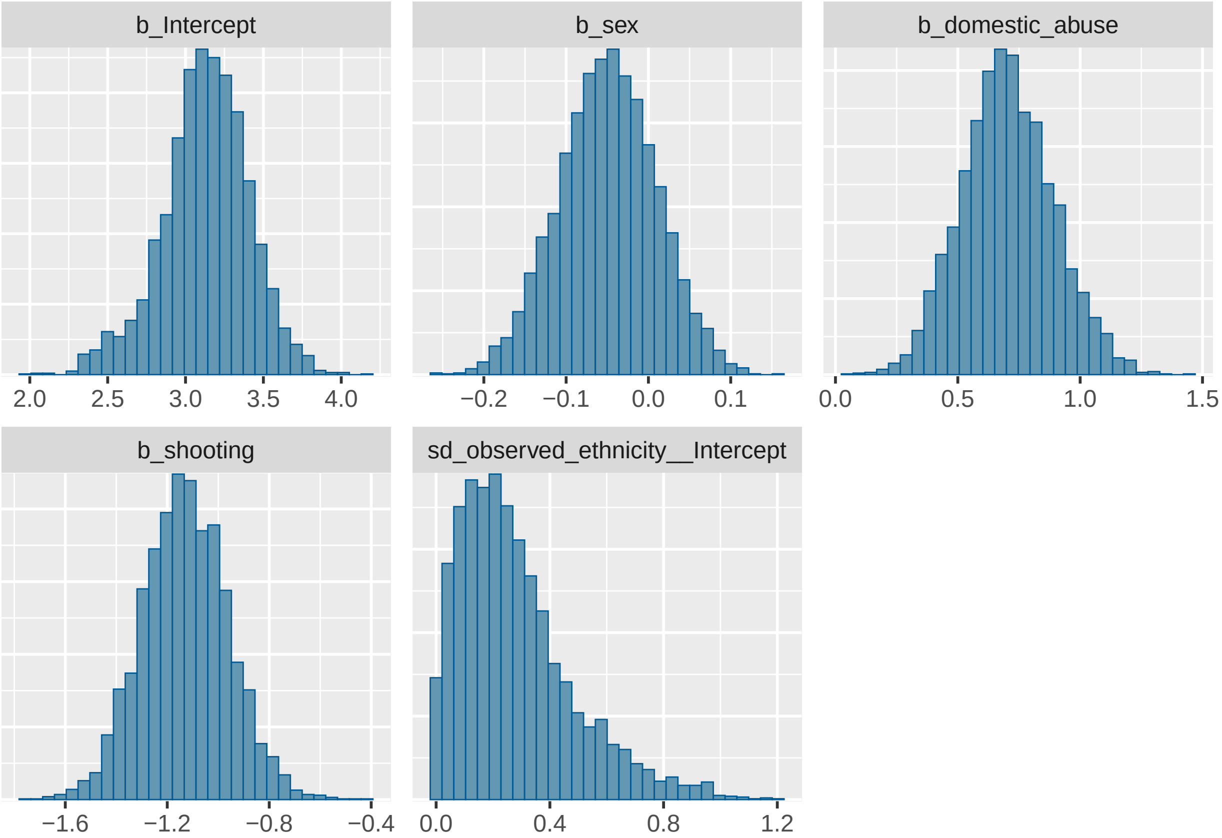
Assess convergence

```
mcmc_plot(fit_2 , type='trace')
```



```
mcmc_plot(fit_2 , type='hist')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```r
summ = summary(fit_2)
# store the min bulk ess - will need it later for monte carlo estimation
n_eff = as.integer(min(summ$fixed$Bulk_ESS))
# Show summary
summ
```

```
##  Family: bernoulli
##   Links: mu = logit
## Formula: solved_status ~ sex + domestic_abuse + shooting + (1 | observed_ethnicity)
##    Data: homicides_1 (Number of observations: 1521)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Group-Level Effects:
## ~observed_ethnicity (Number of levels: 5)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.28      0.20     0.01     0.81 1.01      939     1285
##
## Population-Level Effects:
##                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept          3.13      0.28     2.50     3.64 1.01      947      501
## sex               -0.05      0.06    -0.17     0.07 1.00     3187     1544
## domestic_abuse     0.70      0.19     0.35     1.09 1.00     3306     2985
## shooting          -1.13      0.17    -1.45    -0.79 1.00     3133     2701
##
## 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).
```

```
ranef(fit_2)
```

```
## $observed_ethnicity
## , , Intercept
##
##                           Estimate Est.Error       Q2.5     Q97.5
## Asian                   0.02093840 0.2242264 -0.3948104 0.5761648
## Black                  -0.01613653 0.2108538 -0.4109751 0.5266056
## Not Reported/Not known -0.02346508 0.3125472 -0.7108595 0.6535404
## Other                  -0.03348994 0.2765412 -0.6438964 0.5502347
## White                   0.23401522 0.2593597 -0.1100129 0.9044470
```

Check if the population effects are significant or not.

```
samples = as_draws_df(fit_2)
# Probability that the effect of shooting is actually positive
shooting_samples = samples$b_shooting
sum(shooting_samples>0)/length(shooting_samples)
```

```
## [1] 0
```

```
# Probability that the effect of abuse is actually negative
abuse_samples = samples$b_domestic_abuse
sum(abuse_samples<0)/length(abuse_samples)
```

```
## [1] 0
```

```
# Probability that the effect of being male is actually positive
sex_samples = samples$b_sex
sum(sex_samples>0)/length(sex_samples)
```

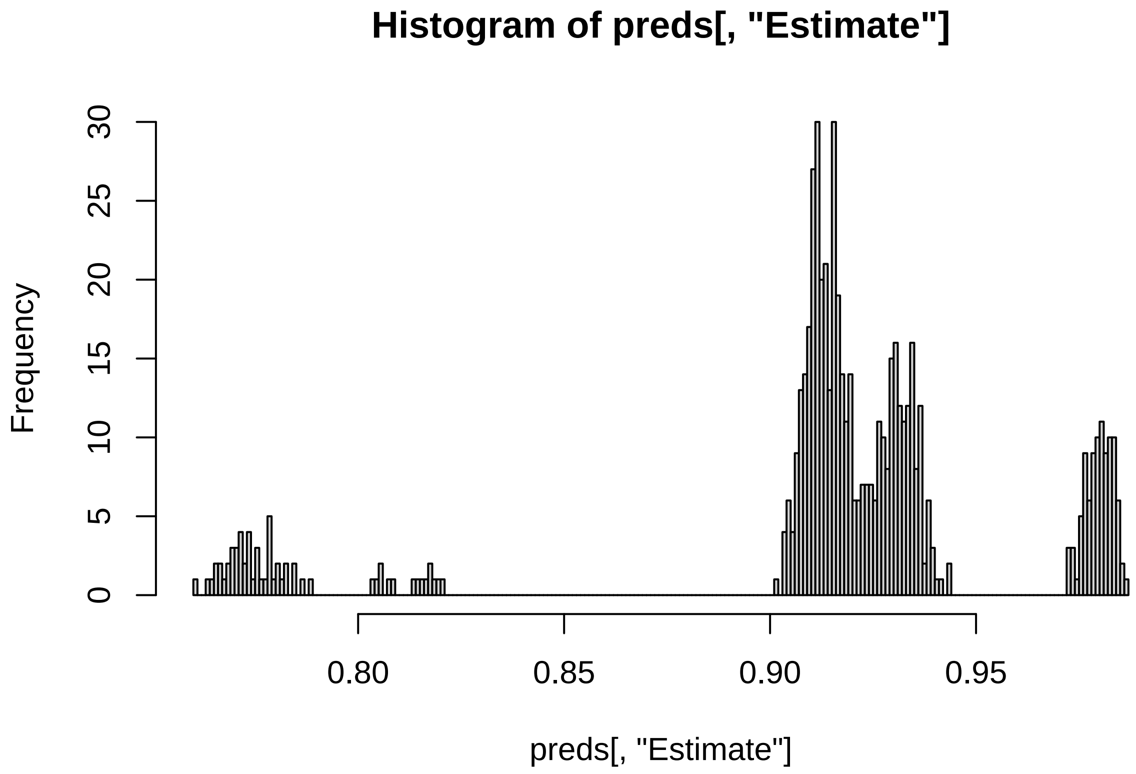
```
## [1] 0.202
```

Check if group effects are significant or not. It appears that the victim being White significantly increases the chance of the homicide being solved (only ~15% chance that the sign of this effect is actuallyy negative). The rest are quite indecisive.

```
for(c in colnames(samples[,6:10])){# ranef posteriors begin at 16 and end at 31
  r_posterior = samples[c]
  gt_zero = sum(r_posterior>0)/nrow(r_posterior)
  print(paste(c , "proba of >0", gt_zero , ", proba of <=0", 1 - gt_zero))
}
```

```
## [1] "r_observed_ethnicity[Asian,Intercept] proba of >0 0.508 , proba of <=0 0.492"
## [1] "r_observed_ethnicity[Black,Intercept] proba of >0 0.41075 , proba of <=0 0.58925"
## [1] "r_observed_ethnicity[Not.Reported/Not.known,Intercept] proba of >0 0.47525 , proba of <=0 0.5247
## [1] "r_observed_ethnicity[Other,Intercept] proba of >0 0.447 , proba of <=0 0.553"
## [1] "r_observed_ethnicity[White,Intercept] proba of >0 0.853 , proba of <=0 0.147"
```

```
preds <- predict(fit_2, newdata=homicides_2)
hist(preds[,"Estimate"],breaks=200)
```



I will keep the threshold at 0.89. This yields an accuracy of ~82%

```
# Use a sensible threshold value
a_classifier <- preds[,"Estimate"]>=0.89
conmat <- ConfusionMatrix(a_classifier, as.integer(homicides_2$solved_status==1))
conmat
```

```
##                  Pred Negative Pred Positive
## Actual Negative            23            70
## Actual Positive            39           477
```

```
sum(diag(conmat))/sum(conmat)
```

```
## [1] 0.8210181
```

This model is the one I will stick with.

3. The following block of code will generate 2 “hypothetical homicides” during the year after March (when the data officially ends).

```
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")
month_tibble$recorded_date <- as.Date(month_tibble$recorded_date)
```

```r
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")
head(hypothetical_homicides, n=2)
```

```
## # A tibble: 2 x 10
##   recorded_~1 age_g~2 sex   obser~3 domes~4 borough metho~5  year month~6 season
##   <date>      <chr>   <chr> <chr>   <chr>   <chr>   <chr>   <int> <chr>   <chr>
## 1 2022-05-01  0 to 12 Fema~ White   Not Do~ Camden  Blunt ~    20 MAY     spring
## 2 2022-09-01  0 to 12 Fema~ Asian   Not Do~ Ealing  Physic~    20 SEP     autumn
## # ... with abbreviated variable names 1: recorded_date, 2: age_group,
## #   3: observed_ethnicity, 4: domestic_abuse, 5: method_of_killing,
## #   6: month.name
```

```r
homicides_3 = pipeline(hypothetical_homicides,y_exists=FALSE)
preds = data.frame(predict(fit_2, newdata=homicides_3, summary=FALSE))
names(preds) = c('hom_A','hom_B')
preds$A_not_B = as.integer((preds$hom_A==1)&(preds$hom_B==0))

# Monte Carlo Estimate
phat = sum(preds$A_not_B)/length(preds$A_not_B)
# MC error
MCerror = sqrt(phat*(1-phat))/n_eff

print(paste('The MC estimate is', round(phat, 6), 'and the MC error is' ,round(MCerror , 6)))
```

```
## [1] "The MC estimate is 0.07625 and the MC error is 0.00028"
```

```r
print(paste('(',round(phat-1.96*MCerror , 6), ',',round(phat+1.96*MCerror , 6),') is a 95% C.I'))
```

```
## [1] "( 0.075701 , 0.076799 ) is a 95% C.I"
```