

MTHM508_Coursework_II

2023-01-13

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 (modified from MTHM508 Pima Indians lectures)
ethnicity_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) ))
  # make a var denoting if ethnicity has been captured
  df$eth_rep = ethnicity_reported(df$observed_ethnicity)
  df$shooting = as.integer(df$method_of_killing=='Shooting')
  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)
  # binarize the solved status
  if(y_exists){
    df = binarize(df=df, col='solved_status', set_to_one='Solved', other_encoding=0)
  }
  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 classification 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 does not bear any θ parameter, so I will not use it going forward. Therefore, our model is described by a bernoulli distribution with mean $\frac{1}{1+e^{-\eta}}$

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.

$$\eta_j = b_0 + b_{sex}x_{1j} + b_{abuse}x_{2j} + b_{shooting}x_{3j} + b_{eth}x_{4j} + \beta_0 + \beta_{sex}x_{1j} + \beta_{abuse}x_{2j} + \beta_{shooting}x_{3j}$$

b_0 and β_0 signify population and group intercepts.

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

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

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. = $\text{logit}(0.8944) = 2.137$. 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 $\sim 96\%$ 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 $\sim 75\%$ 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 $\sim 2.19 + 3.16$ 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')

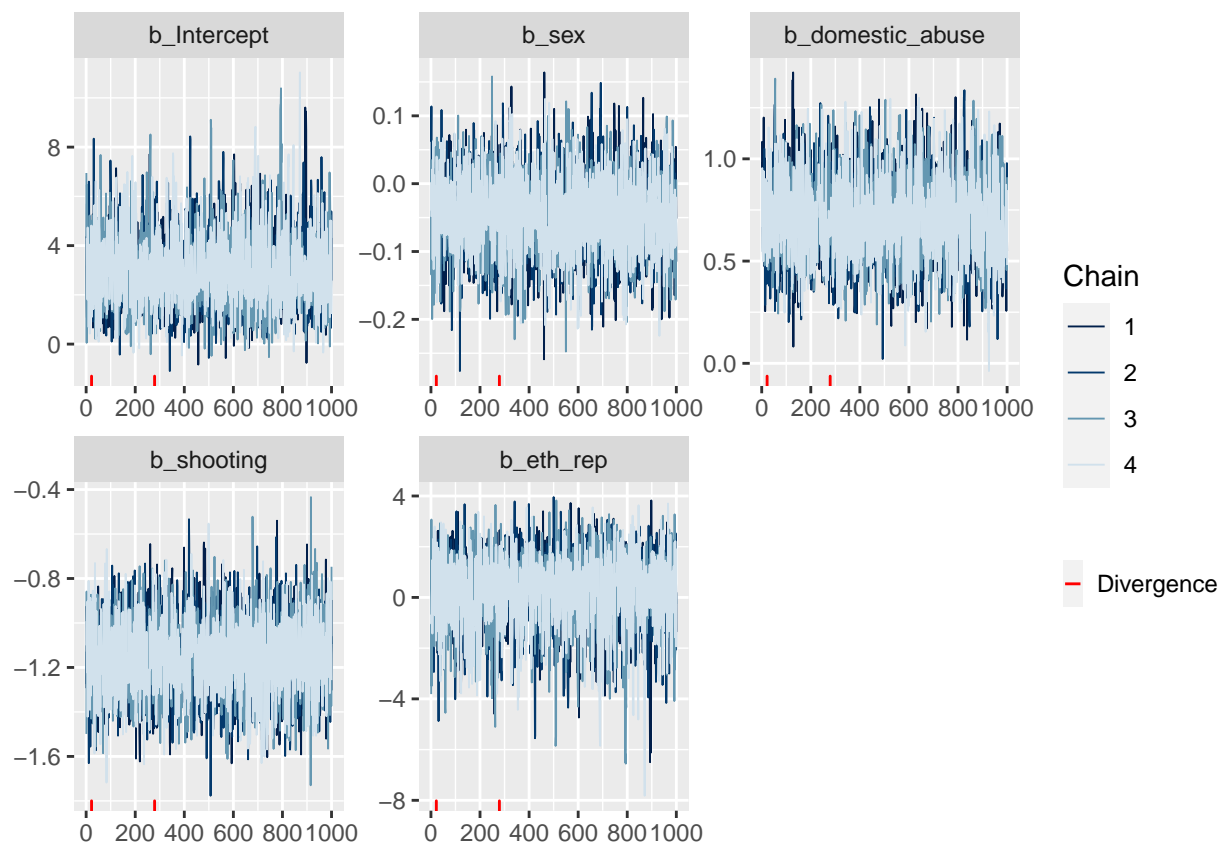
sd_priors_I <- set_prior("normal(0,0.5)",class="sd", group="season", coef='Intercept')
sd_priors <- set_prior("normal(0,0.1)",class="sd", group="season", coef = c('sex','domestic_abuse','shooting'))

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

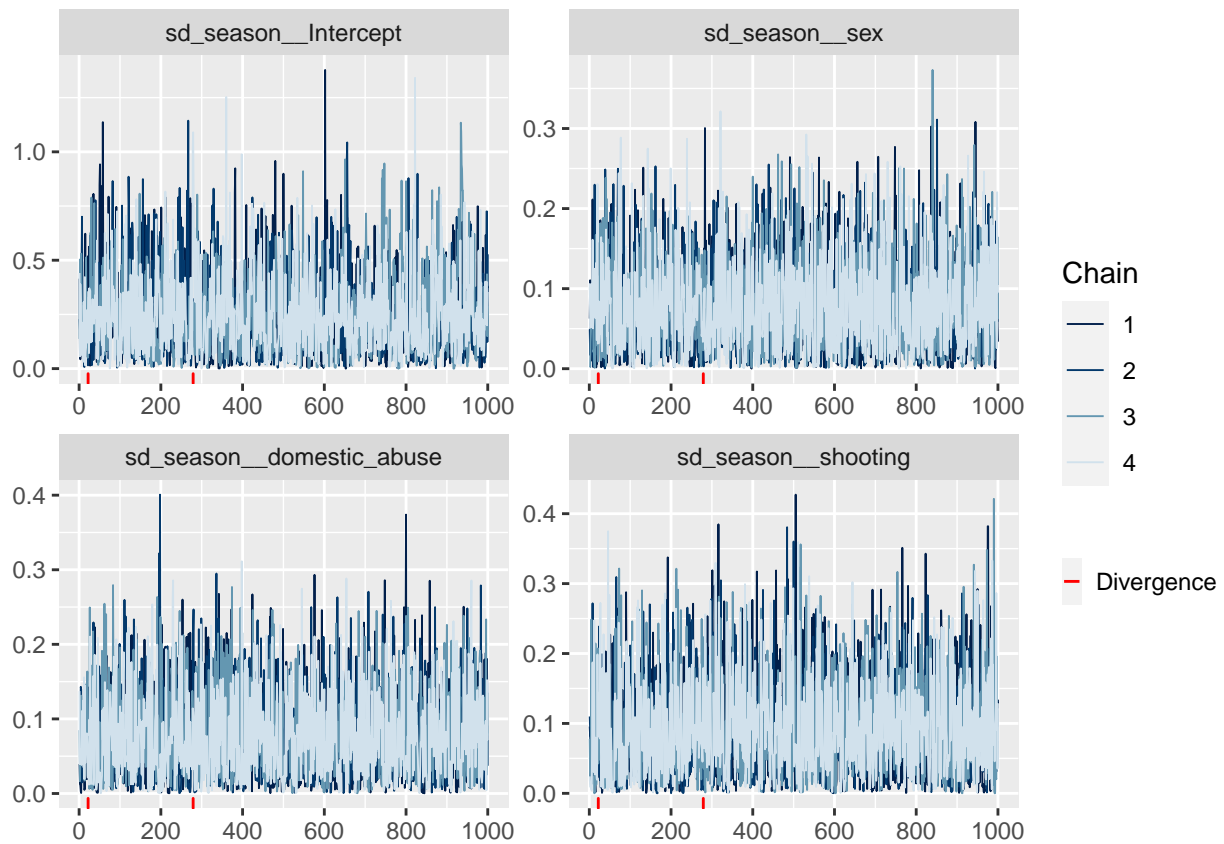
fit_1 = brm(solved_status~sex+domestic_abuse+shooting+eth_rep
            +(sex+domestic_abuse+shooting|season),
            data=homicides_1,
            family=bernoulli,
            prior=my_prior,
            refresh=0)

## Compiling Stan program...
## Start sampling
Trace plots are examined.
mcmc_plot(fit_1 , type='trace', variable = '~b_', regex=TRUE)

```

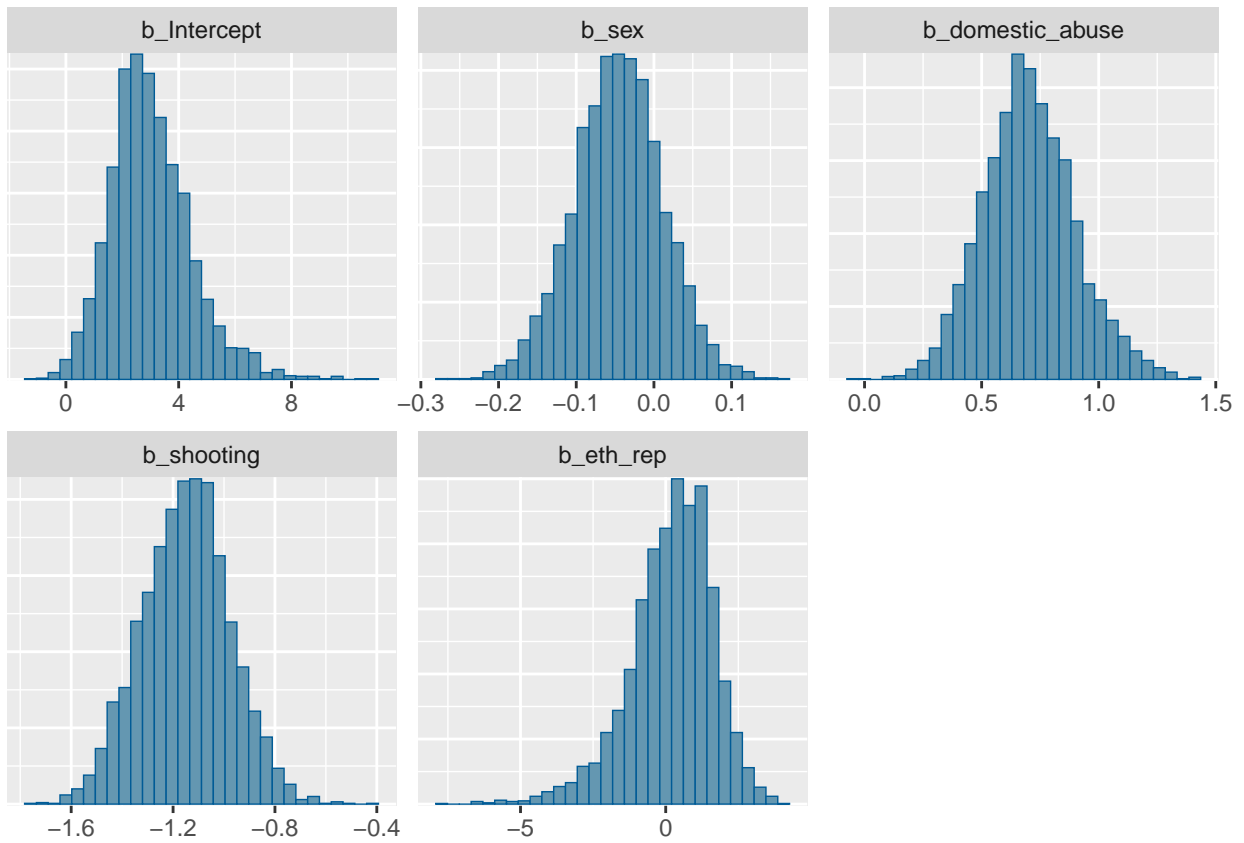


```
mcmc_plot(fit_1 , type='trace', variable = '^sd_', regex=TRUE)
```



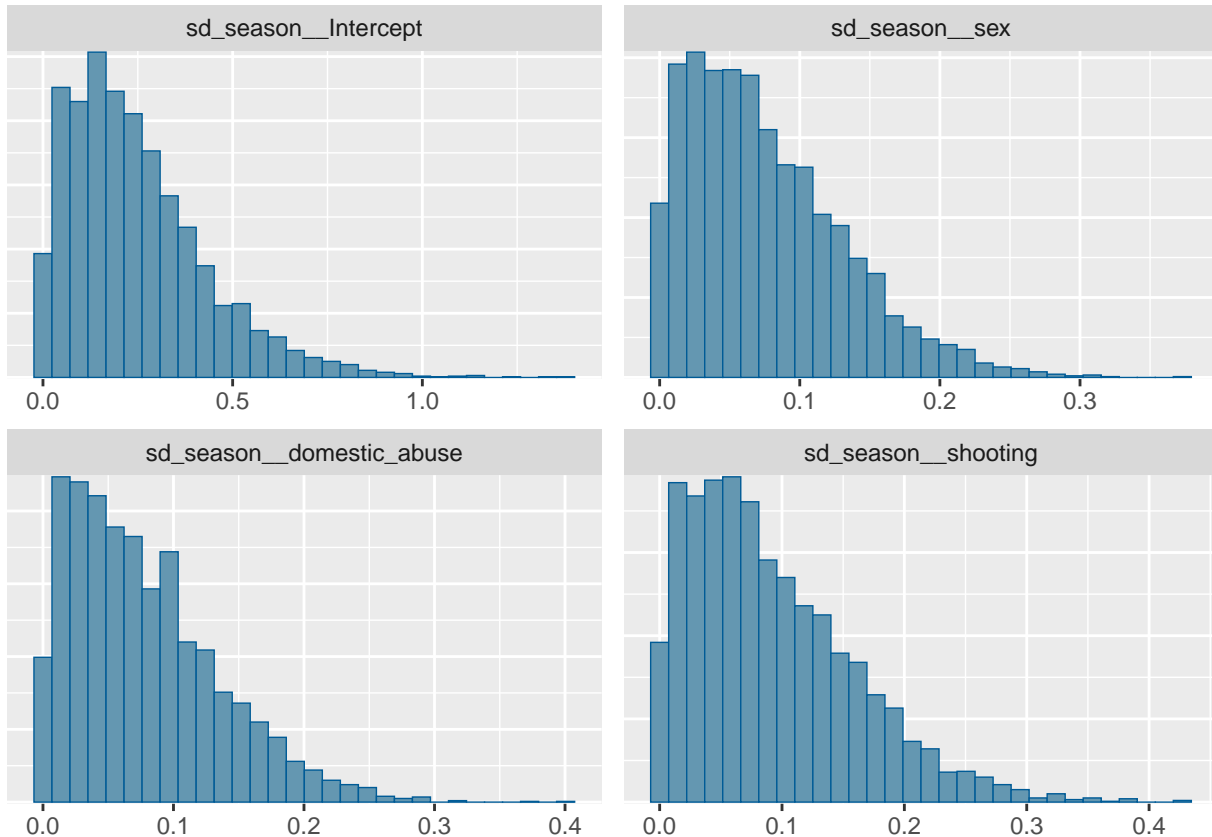
```
mcmc_plot(fit_1 , type='hist', variable = '^b_', regex=TRUE)
```

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



```
mcmc_plot(fit_1 , type='hist', variable = '^sd_', regex=TRUE)
```

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



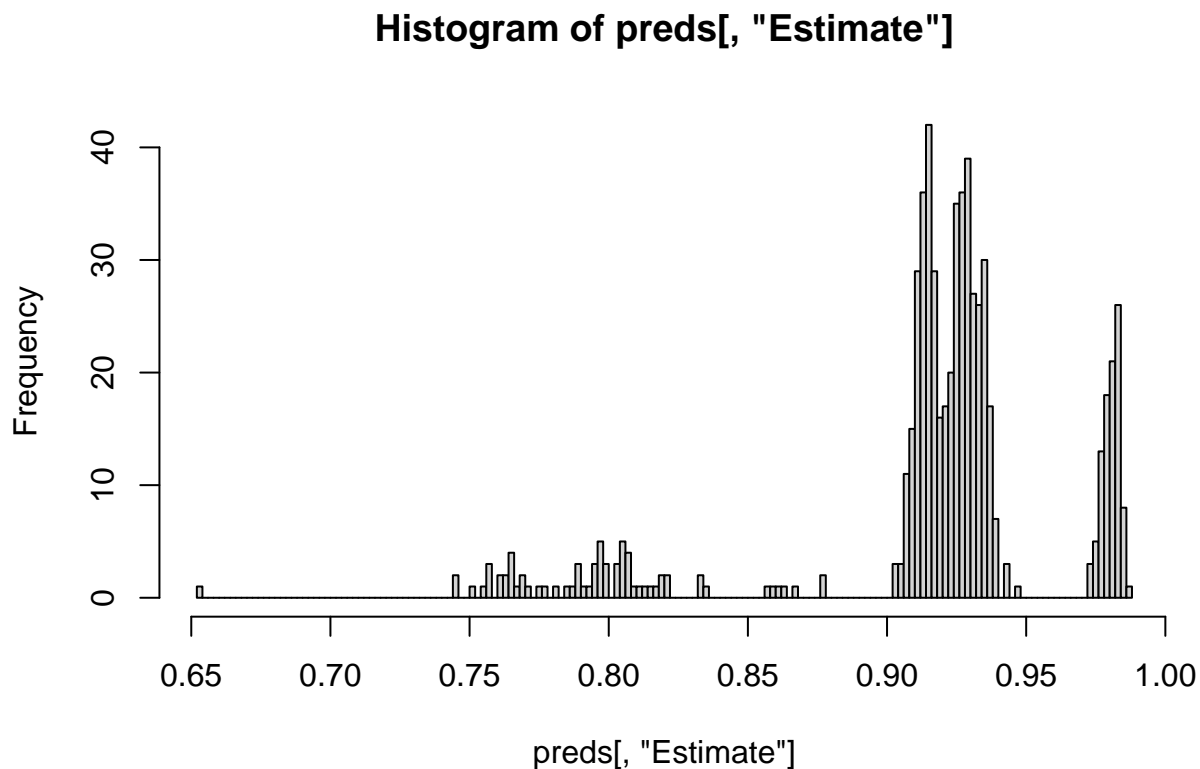
The traceplots suggest that convergence has been reached. To judge whether model has adequately fit the data, I will use it to predict over unseen data (the 5 years of data stowed away at the beginning).

```
# 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 1-95% CI u-95% CI Rhat
## sd(Intercept)          0.25      0.18    0.01    0.70 1.00
## sd(sex)                0.08      0.06    0.00    0.21 1.00
## sd(domestic_abuse)      0.08      0.06    0.00    0.21 1.00
## sd(shooting)            0.09      0.07    0.00    0.26 1.00
## cor(Intercept,sex)     -0.07      0.44   -0.84    0.78 1.00
## cor(Intercept,domestic_abuse) 0.03      0.45   -0.79    0.81 1.00
## cor(sex,domestic_abuse) 0.00      0.45   -0.82    0.81 1.00
## cor(Intercept,shooting) -0.03      0.45   -0.81    0.82 1.00
## cor(sex,shooting)       0.04      0.45   -0.79    0.84 1.00
## cor(domestic_abuse,shooting) 0.00      0.45   -0.81    0.82 1.00
##
##           Bulk_ESS Tail_ESS
## sd(Intercept)        1424    2052
## sd(sex)              2867    1763
## sd(domestic_abuse)   2372    1780
## sd(shooting)         2443    1741
## cor(Intercept,sex)   3539    2683
## cor(Intercept,domestic_abuse) 3452    2427
## cor(sex,domestic_abuse) 3718    2593
## cor(Intercept,shooting) 3600    2665
## cor(sex,shooting)    3832    2748
## cor(domestic_abuse,shooting) 2893    2701
##
## Population-Level Effects:
##           Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept          3.02      1.45    0.52    6.45 1.00    2977    2044
## sex                -0.05      0.06   -0.16    0.07 1.00    4095    3027
## domestic_abuse      0.70      0.20    0.33    1.13 1.00    3104    2329
## shooting            -1.15      0.17   -1.49   -0.81 1.00    3078    2450
## eth_rep             0.19      1.44   -3.23    2.63 1.00    3244    1921
##
## 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).

```

Domestic abuse and shooting play heavily into whether the crime will be solved. The sex of the victim as well as whether ethnicity was reported or not contribute to the model, albeit to a lesser extent (coefficients closer to 0). Recall that sex is encoded as 1 for 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. The ethnicity being reported increases chances of the case being solved - if I think about it, this makes sense as a

missing (or unreported) ethnicity may point towards there being problems in the paperwork, indicating the case may take longer to solve.

Investigate the group level effects.

```
ranef(fit_1)

## $season
## , , Intercept
##
##      Estimate Est.Error      Q2.5      Q97.5
## autumn -0.07427481 0.2038833 -0.5318750 0.3368919
## spring  0.12866347 0.2272994 -0.2471778 0.6944977
## summer  0.17307021 0.2388304 -0.1578140 0.7890939
## winter -0.01644212 0.1987596 -0.4125010 0.4387734
##
## , , sex
##
##      Estimate Est.Error      Q2.5      Q97.5
## autumn -0.01793746 0.08536351 -0.2164975 0.1442509
## spring -0.02662755 0.08843029 -0.2465796 0.1265119
## summer -0.02607010 0.08822201 -0.2567478 0.1300073
## winter -0.02620061 0.08964948 -0.2440299 0.1257073
##
## , , domestic_abuse
##
##      Estimate Est.Error      Q2.5      Q97.5
## autumn  0.003826291 0.08696126 -0.1843256 0.1976417
## spring  0.009555773 0.08914565 -0.1703374 0.2144152
## summer -0.015515306 0.09375141 -0.2429321 0.1717057
## winter  0.034633129 0.09644824 -0.1225886 0.2841744
##
## , , shooting
##
##      Estimate Est.Error      Q2.5      Q97.5
## autumn -0.05673733 0.1227265 -0.3796307 0.1189581
## spring -0.04668248 0.1154198 -0.3521522 0.1298257
## summer -0.03111938 0.1128356 -0.3116758 0.1549773
## winter -0.03037107 0.1111447 -0.3142104 0.1649550
```

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] 0.3985
```

This says that there is ~40% chance that the mean effect of reported ethnicity is of the wrong sign - this is not reliable. **I will remove it from our model.** Before that, let's check a few other posterior samples. From this, it is clear that shooting, sex and domestic abuse have a significant effect on the response variable.

```

# 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] 0.00025

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

## [1] 0.20625

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 0.34075 , proba of <=0 0.65925"
## [1] "r_season[spring,Intercept] proba of >0 0.716 , proba of <=0 0.284"
## [1] "r_season[summer,Intercept] proba of >0 0.78375 , proba of <=0 0.21625"
## [1] "r_season[winter,Intercept] proba of >0 0.4365 , proba of <=0 0.5635"
## [1] "r_season[autumn,sex] proba of >0 0.4235 , proba of <=0 0.5765"
## [1] "r_season[spring,sex] proba of >0 0.4 , proba of <=0 0.6"
## [1] "r_season[summer,sex] proba of >0 0.38775 , proba of <=0 0.61225"
## [1] "r_season[winter,sex] proba of >0 0.40075 , proba of <=0 0.59925"
## [1] "r_season[autumn,domestic_abuse] proba of >0 0.51125 , proba of <=0 0.48875"
## [1] "r_season[spring,domestic_abuse] proba of >0 0.5275 , proba of <=0 0.4725"
## [1] "r_season[summer,domestic_abuse] proba of >0 0.44225 , proba of <=0 0.55775"
## [1] "r_season[winter,domestic_abuse] proba of >0 0.63725 , proba of <=0 0.36275"
## [1] "r_season[autumn,shooting] proba of >0 0.32925 , proba of <=0 0.67075"
## [1] "r_season[spring,shooting] proba of >0 0.36425 , proba of <=0 0.63575"
## [1] "r_season[summer,shooting] proba of >0 0.41425 , proba of <=0 0.58575"
## [1] "r_season[winter,shooting] proba of >0 0.4125 , proba of <=0 0.5875"

```

Looking at the random effects, none of the covariates are that affected across different seasons (apart from Intercept which increases in summer, i.e. there is ~23% chance of there not actually being an increase, which is far less than the other group effects).

Model Checking - Refit 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.

$$\eta_j = b_0 + b_{sex}x_{1j} + b_{abuse}x_{2j} + b_{shooting}x_{3j} + \beta_0 + \beta_{sex}x_{1j} + \beta_{abuse}x_{2j} + \beta_{shooting}x_{3j}$$

b_0 and β_0 signify population and group intercepts (grouped on `observed_ethnicity`).

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

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

```

sd_priors_I <- set_prior("normal(0,0.5)",class="sd", group="observed_ethnicity", coef='Intercept')
sd_priors <- set_prior("normal(0,0.1)",class="sd", group="observed_ethnicity", coef = c('sex','domestic
# Join priors into a vector
my_prior = c(intercept_prior,sex_prior,abuse_prior,shooting_prior,sd_priors_I,sd_priors)

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

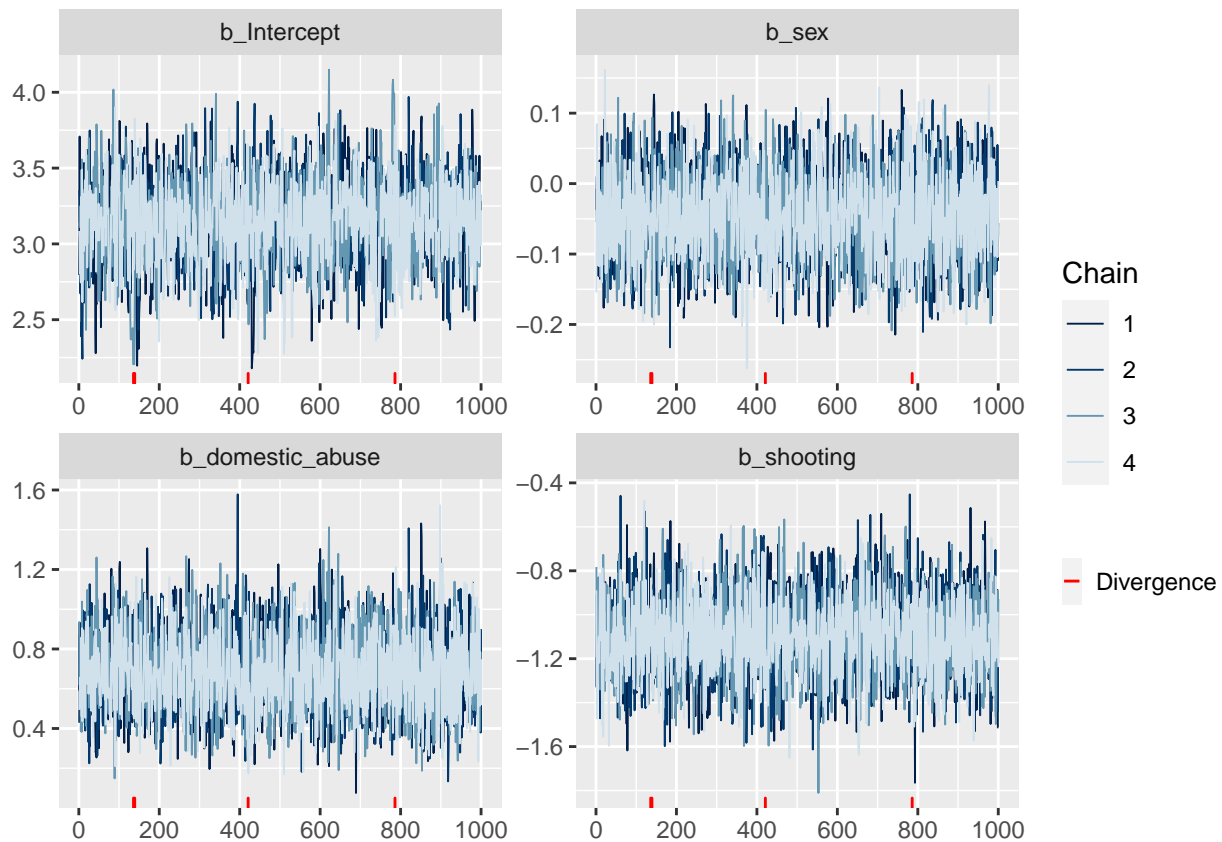
```

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

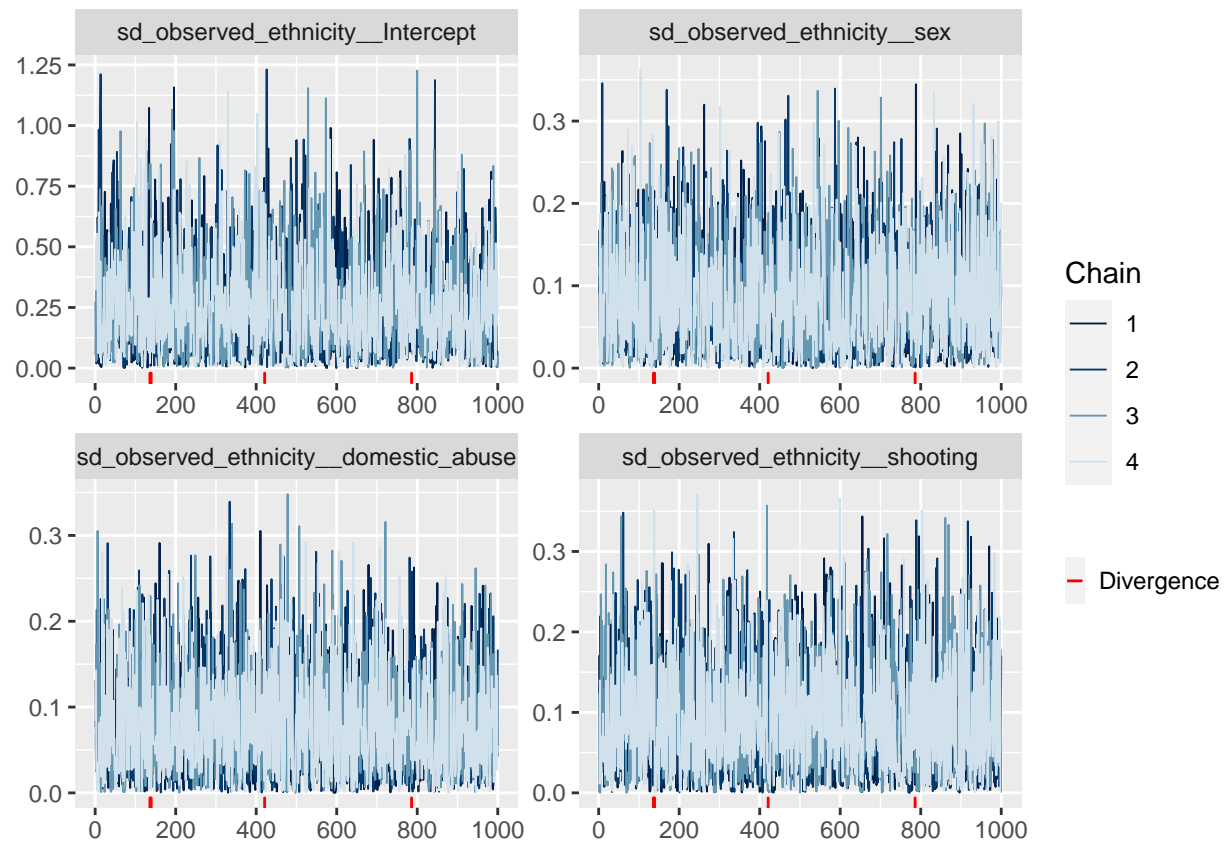
```
## Start sampling
```

```
Assess convergence
```

```
mcmc_plot(fit_2 , type='trace', variable = '^b_', regex=TRUE)
```

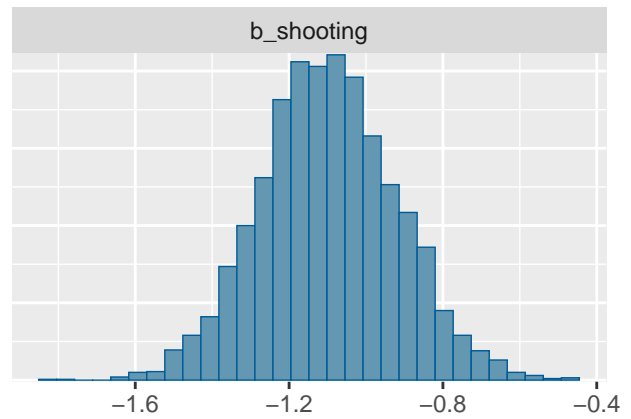
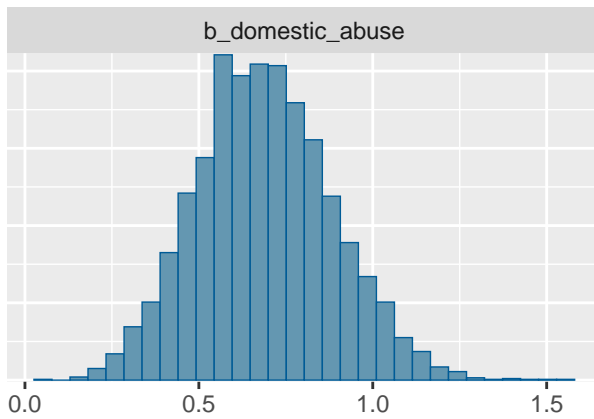
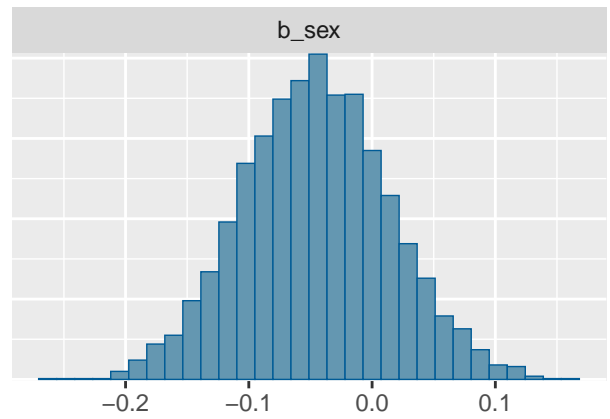
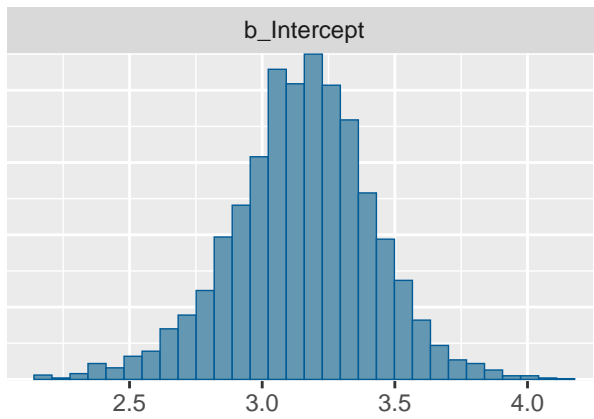


```
mcmc_plot(fit_2 , type='trace', variable = '^sd_', regex=TRUE)
```



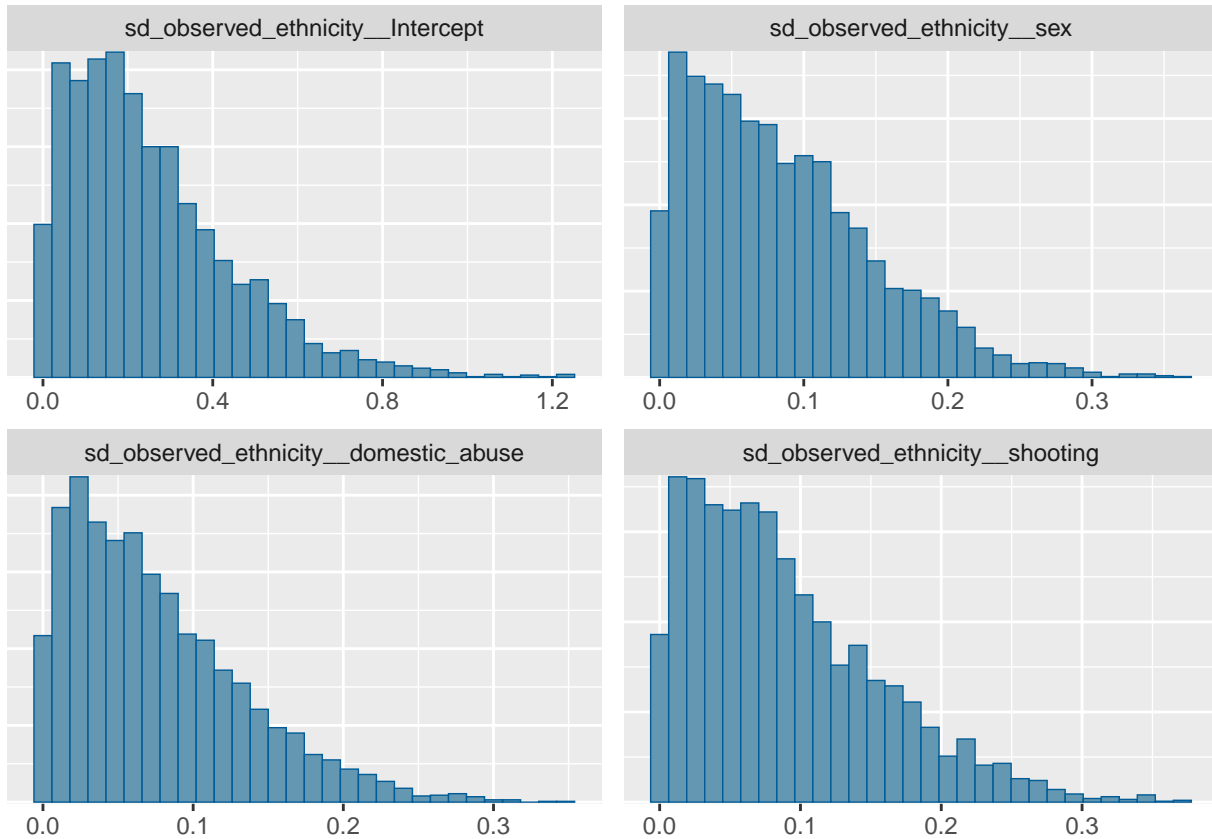
```
mcmc_plot(fit_2 , type='hist', variable = '^b_', regex=TRUE)
```

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



```
mcmc_plot(fit_2 , type='hist', variable = '^sd_', regex=TRUE)
```

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



```
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 + (sex + domestic_abuse + shooting | 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	1-95% CI	u-95% CI	Rhat
sd(Intercept)	0.25	0.20	0.01	0.74	1.00
sd(sex)	0.09	0.06	0.00	0.23	1.00
sd(domestic_abuse)	0.08	0.06	0.00	0.22	1.00
sd(shooting)	0.09	0.07	0.00	0.25	1.00
cor(Intercept,sex)	-0.02	0.44	-0.81	0.80	1.00
cor(Intercept,domestic_abuse)	0.04	0.43	-0.78	0.79	1.00
cor(sex,domestic_abuse)	-0.00	0.44	-0.80	0.82	1.00
cor(Intercept,shooting)	-0.04	0.44	-0.81	0.80	1.00
cor(sex,shooting)	0.04	0.44	-0.78	0.82	1.00
cor(domestic_abuse,shooting)	0.00	0.45	-0.82	0.81	1.00

```
## Bulk_ESS Tail_ESS
```

```
## sd(Intercept)          1397      1824
## sd(sex)                2387      1455
## sd(domestic_abuse)     2982      1655
## sd(shooting)           2230      1496
## cor(Intercept,sex)     4079      2837
## cor(Intercept,domestic_abuse) 4454      3042
## cor(sex,domestic_abuse) 3540      2550
## cor(Intercept,shooting) 3664      2880
## cor(sex,shooting)      3365      2883
## cor(domestic_abuse,shooting) 3181      3395
##
## Population-Level Effects:
##           Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept          3.15      0.27    2.57    3.66 1.00     1800     1085
## sex                -0.05      0.06   -0.16    0.07 1.00     4136     3145
## domestic_abuse      0.69      0.20    0.32    1.09 1.00     4012     2938
## shooting           -1.10      0.18   -1.45   -0.74 1.00     3543     2240
##
## 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          0.036097765 0.2390065 -0.4057534 0.5812631
## Black          -0.002842002 0.2066067 -0.3934064 0.4778897
## Not Reported/Not known -0.024355427 0.2960045 -0.6955977 0.6137760
## Other          -0.030554931 0.2597047 -0.6173484 0.4912525
## White          0.194102137 0.2580085 -0.1678558 0.8250980
##
## , , sex
##
##           Estimate Est.Error      Q2.5      Q97.5
## Asian        -0.053156246 0.10983566 -0.3532423 0.10366171
## Black        -0.052520759 0.10094193 -0.3136879 0.09282259
## Not Reported/Not known -0.007362178 0.10783921 -0.2627512 0.21149097
## Other        -0.006906924 0.10874677 -0.2438865 0.21408143
## White        0.007916529 0.08790058 -0.1740755 0.21129023
##
## , , domestic_abuse
##
##           Estimate Est.Error      Q2.5      Q97.5
## Asian        0.022795790 0.09400101 -0.1385088 0.2624918
## Black        0.005003558 0.08708506 -0.1768467 0.2027444
## Not Reported/Not known 0.003828336 0.09555596 -0.1985246 0.2213839
## Other        0.007855902 0.09347181 -0.1864252 0.2307923
## White       -0.004049490 0.08809920 -0.1912088 0.1836865
##
## , , shooting
##
##           Estimate Est.Error      Q2.5      Q97.5
```



```
## Asian          -0.045218043 0.1232605 -0.3703933 0.1389556
## Black          -0.046154311 0.1104366 -0.3330005 0.1227237
## Not Reported/Not known -0.005597407 0.1141142 -0.2562293 0.2258908
## Other          -0.012824195 0.1124581 -0.2775103 0.2000041
## White          -0.031264925 0.1124256 -0.3065325 0.1604701
```

It appears that the victim being white increases the intercept, i.e. the base probability of the murder being solved. The other covariates do not seem to be affected much by this grouping. I observe that the Est. Error for the intercept has decreased, denoting that some of the variance has been explained away by the differences in ethnicity. This also suggests that this is a better grouping than the previous one.

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] 0.2185
```

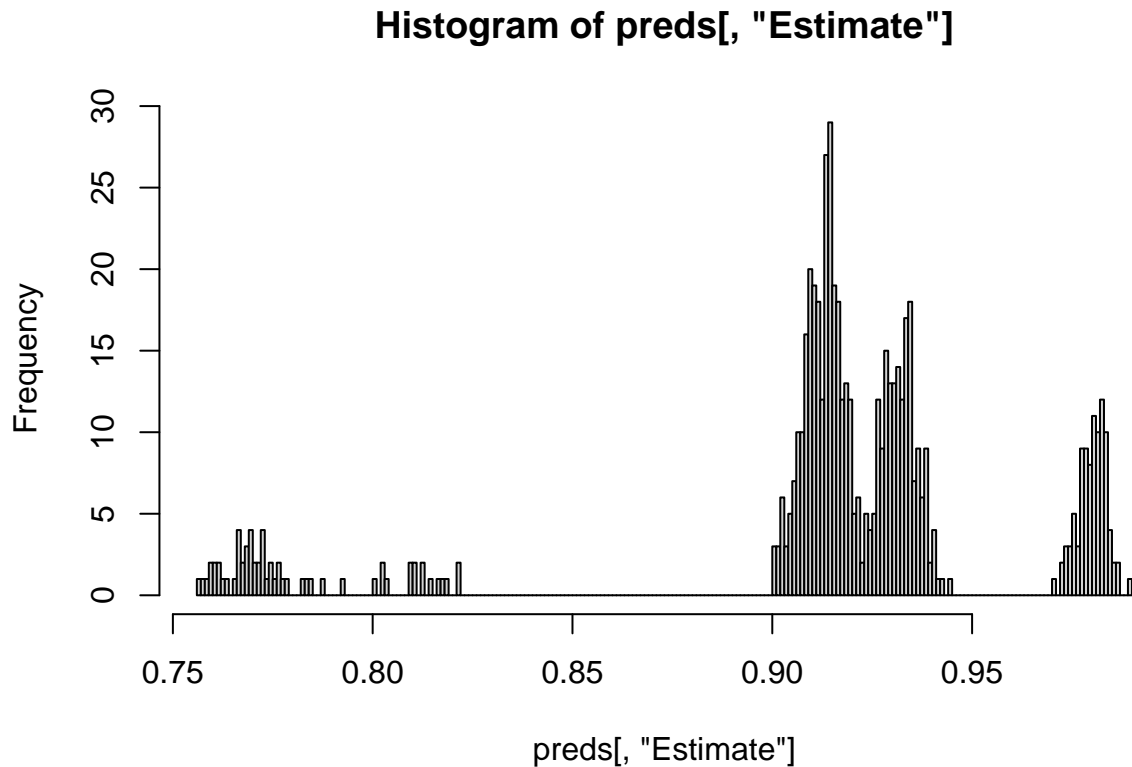
Check if group effects are significant or not. The probability of the intercept being greater than 0 when the victim is white is over 80% - this suggests that it is likely a real effect. The case is more probable to be solved if the victim is white. This confirms a previous argument.

```
for(c in colnames(samples[,15:34])){# raneef 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 0.53775 , proba of <=0 0.46225"
## [1] "r_observed_ethnicity[Black,Intercept] proba of >0 0.45175 , proba of <=0 0.54825"
## [1] "r_observed_ethnicity[Not.Reported/Not.known,Intercept] proba of >0 0.4765 , proba of <=0 0.5235"
## [1] "r_observed_ethnicity[Other,Intercept] proba of >0 0.461 , proba of <=0 0.539"
## [1] "r_observed_ethnicity[White,Intercept] proba of >0 0.7915 , proba of <=0 0.2085"
## [1] "r_observed_ethnicity[Asian,sex] proba of >0 0.327 , proba of <=0 0.673"
## [1] "r_observed_ethnicity[Black,sex] proba of >0 0.303 , proba of <=0 0.697"
## [1] "r_observed_ethnicity[Not.Reported/Not.known,sex] proba of >0 0.4765 , proba of <=0 0.5235"
## [1] "r_observed_ethnicity[Other,sex] proba of >0 0.473 , proba of <=0 0.527"
## [1] "r_observed_ethnicity[White,sex] proba of >0 0.53325 , proba of <=0 0.46675"
## [1] "r_observed_ethnicity[Asian,domestic_abuse] proba of >0 0.57825 , proba of <=0 0.42175"
## [1] "r_observed_ethnicity[Black,domestic_abuse] proba of >0 0.524 , proba of <=0 0.476"
## [1] "r_observed_ethnicity[Not.Reported/Not.known,domestic_abuse] proba of >0 0.51325 , proba of <=0 0.48675"
## [1] "r_observed_ethnicity[Other,domestic_abuse] proba of >0 0.53825 , proba of <=0 0.46175"
## [1] "r_observed_ethnicity[White,domestic_abuse] proba of >0 0.4815 , proba of <=0 0.5185"
## [1] "r_observed_ethnicity[Asian,shooting] proba of >0 0.367 , proba of <=0 0.633"
## [1] "r_observed_ethnicity[Black,shooting] proba of >0 0.353 , proba of <=0 0.647"
```

```
## [1] "r_observed_ethnicity[Not.Reported/Not.known,shooting] proba of >0 0.48825 , proba of <=0 0.51175"
## [1] "r_observed_ethnicity[Other,shooting] proba of >0 0.464 , proba of <=0 0.536"
## [1] "r_observed_ethnicity[White,shooting] proba of >0 0.4155 , proba of <=0 0.5845"

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)
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 2022-08-01  55 to ~ Fema~ Other   Domest~ Merton  Blunt ~    20 AUG     summer
## 2 2022-07-01  55 to ~ Fema~ Not Re~ Domest~ Harrow  Knife ~    20 JUL     summer
## # ... with abbreviated variable names 1: recorded_date, 2: age_group,
## #   3: observed_ethnicity, 4: domestic_abuse, 5: method_of_killing,
## #   6: month.name

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 0.022 and the MC error is 8.1e-05"
print(paste('(',round(phat-1.96*MCError , 6), ',',round(phat+1.96*MCError , 6),') is a 95% C.I'))

## [1] "( 0.02184 , 0.02216 ) is a 95% C.I"

```