Investigating the influence of state on US serial killer's age at last kill

MATH3092: Mixed Models Report

Faye Williams
201308646

Department of Mathematics
University of Leeds
United Kingdom
May 2023

1 Introduction

The average US serial killer commits their last murder at 44 years of age, and this as the age of the killer's first kill increases, their age at last kill increases at an equal rate. A serial killer's race and sexual preference also has a profound impact on the killer's age at last kill, where those that are white and heterosexual have a higher age at last kill, compared to those that are non-white and non-heterosexual. Evidence to support these facts shall be demonstrate in this report, through the analysis of the 'CWsample' dataset. The influence of the state that the murder was committed in has on the killer's age at last kill shall also be examined, as well as the combination of other covariates on this variable, such as sex, race and sexual preference.

'CWsample' is a sample of 366 US serial killers from the 'killers' dataset, which consists of personal information concerning serial killers from 12 US states, as well as information regarding their crimes. This includes the serial killer's ID numbers, their sex and the number of victims of each killer. The 'State' column gives the two-letter abbreviation of the US state the murder was committed in. Other relevant columns we shall utilise in our analyses include the 'AgeLastKill' column, which provides the age of the killer when they committed their last murder and similarly 'AgeFirstKill' denotes their age at the time of their first murder. Some columns that have been created within the dataset include the Boolean variable 'White', which is 'TRUE' for white killers and 'FALSE' for non-white killers, as determined using the 'Race' column of the dataset. Furthermore, the Boolean variable 'Hetero' equals 'TRUE' for killers that are heterosexual and 'FALSE' for those that are either homosexual, bisexual or their sexuality is unknown, as provided by the 'Sexual Preference' column on the original dataset. The packages used to conduct the analysis include the 'Stats' package, used to create generalised linear models via the 'lm' function. The 'lme4' package was also utilised to create random intercept and random slope models, using the 'lmer' function. Note that any R code used in the remainder of the report can be seen in Appendix A.

2 Results

Linear Regression Model. We find that the serial killer's age at time of last kill varies quite greatly between killers, as shown from the histogram plot on the left in figure 1. As we see a roughly bell shaped curve of the age at last kill variable, we can reasonably assume that this variable is normally distributed. In addition, the normal quantile-quantile (Q-Q) plot on the right shows little curvature, further suggesting 'AgeLastKill' is a normally distributed variable.

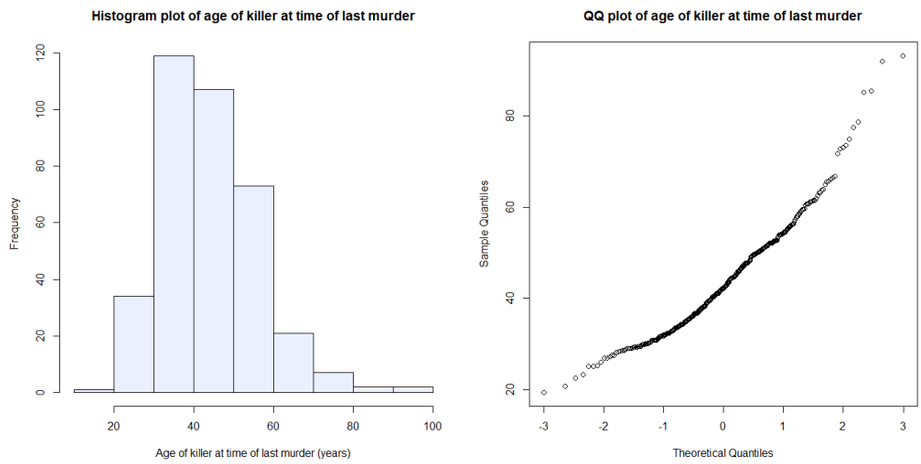

Figure 1: Histogram of killer's age at last kill (right) and a Q-Q plot of the age of killers at last murder (left) based on model (1).

When assuming the ages of killers at their last kill are independent and identically distributed, we propose the model:

$$y_i = \beta_0 + e_i \quad \text{with} \quad e_i \sim N(0, \sigma^2). \tag{1}$$

where y_i denotes the age of last kill of the i-th killer in our sample, such that $i = 1, ..., 366$ and e_i denotes the random error for each individual. Using the method of ordinary least squares via the 'lm' function in R, we find an estimate for the population mean, $\hat{\beta}_0 \approx 43.8$ years (3 s.f.), meaning the average killer's age at last kill is around 44 years. Note that all estimates will be rounded to three significant figures unless stated otherwise. In addition, the model has an estimated variance of $\hat{\sigma_0}^2 \approx 11.9^2$, which is quite a large spread of around 12 years, suggesting that the distribution of age at last kill varies greatly between individuals. In an attempt to decrease this, we propose the following linear regression model, model A, which investigates the relationship between age of killer at last kill and age at first kill,

$$\text{Model A:} \quad y_i = \beta_0 + \beta_1 x_i + e_i \quad \text{with} \quad e_i \sim N(0, \sigma^2). \tag{2}$$

Here x_i denotes the age of the murderer at the time of their first kill, in years. Defining this model in R using the 'lm' function, the estimates of the parameters and variance become $\hat{\beta}_0 \approx 12.4$ years, $\hat{\beta}_1 \approx 0.989$ years and $\hat{\sigma_0}^2 \approx 7.13^2$ respectively. As the variance parameter has decreased from the previous model and the coefficient of β_1 is approximately one, there is evidence to suggest the individuals age at first kill increases at the same rate as the killer's age at last kill.

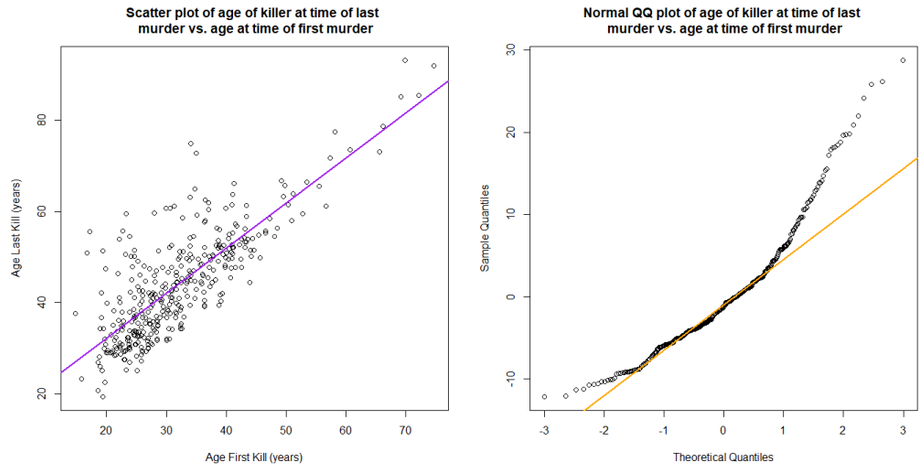

Figure 2: Plot of killers age at last kill against the killers age at first kill, with a purple estimated regression line based on our linear regression model (2) (left) and the corresponding normal Q-Q plot of the estimated error residuals for this model (right), with an orange Q-Q added to the plot.

The Q-Q plot on the right of figure 2 shows a slight curvature, as demonstrated by many of the points lying above the straight orange line. This suggests that the distribution of age of last kill may not support the normality assumption in this linear regression model. The scatter plot on the left however shows the plots are equally spread above and below the purple regression line, suggesting that the linear regression model is a good fit to the data. Also, the positive gradient of this line suggests a positive correlation between a killer's age at first kill and at last kill, as demonstrated by the positive coefficient β_1. To improve this model we shall consider whether the state the murder was committed in has an influence on the relationship between a killer's age at last kill and first kill. The number of serial killers from each of the 12 US states is summarised in a table in figure 3 below.

AL	AR	AZ	CA	CT	DC	GA	KS	MD	MO	MS	NC
10	5	20	155	14	20	33	11	20	32	6	40

Figure 3: Table of the number of killers from each US state in the sample.

The influence of clustering on the killers age at last kill can be more clearly seen in figure 4. We see that the variance around the estimated mean from the original model greatly varies between killers in different states, where the age of last kill of killers in Washington DC all fall below the estimated mean. This suggests that a multilevel model that will takes into account the killer's state would give a better fit to the data than the previous models.

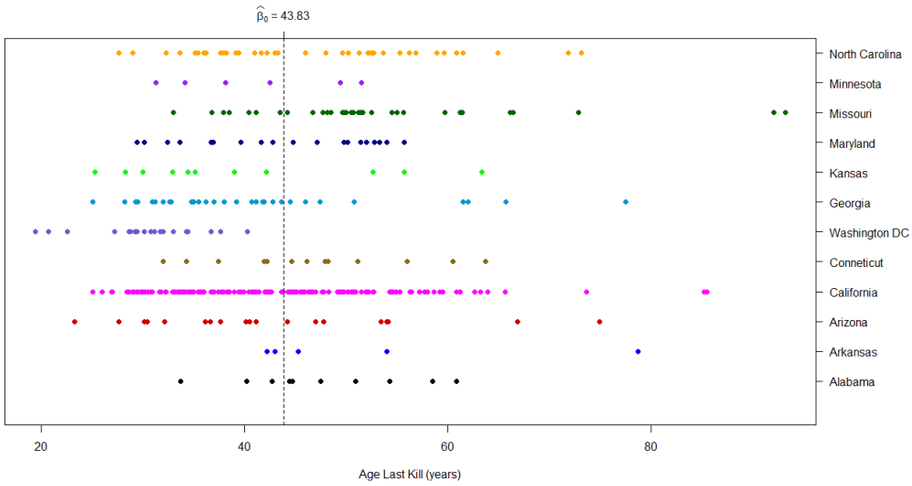

Figure 4: Plot of serial killers age at last kill, grouped by the state they killed in, with a dotted line at the estimated mean of the original model.

Random Intercept Model. Here we are examining the response variable y_{ij}, which is the age of the i-th killer in the j-th state at the time of their last kill. Assuming that the killers age at last kill is independent and identically distributed, we propose the random intercept model,

$$\text{Model B:} \quad y_{ij} = \beta_0 + \beta_1 x_{1ij} + u_{0j} + e_{0ij} \quad \text{with} \quad u_{0j} \sim N(0, \sigma^2_{u0}) \quad \& \quad e_{0ij} \sim N(0, \sigma^2_{e0}). \tag{3}$$

where x_{1ij} denotes the age of first kill of the i-th killer in the j-th state. Defining model B in R using the 'lmer' function, we find the estimated parameters are $\hat{\beta}_0 \approx 14.3$ and $\hat{\beta}_1 \approx 0.947$, and estimates for variance

are given by $\hat{\sigma}^2_{u0} \approx 3.77^2$ at the state level and $\hat{\sigma}^2_{e0} \approx 6.43^2$ at the killer level. The plot of the killer's age at first kill against age at last kill fitted with the random intercept model (3) is given in figure 5. It can be seen that the age of first kill intercepts of the dashed lines vary greatly, suggesting that the plotted points are indeed clustered by state.

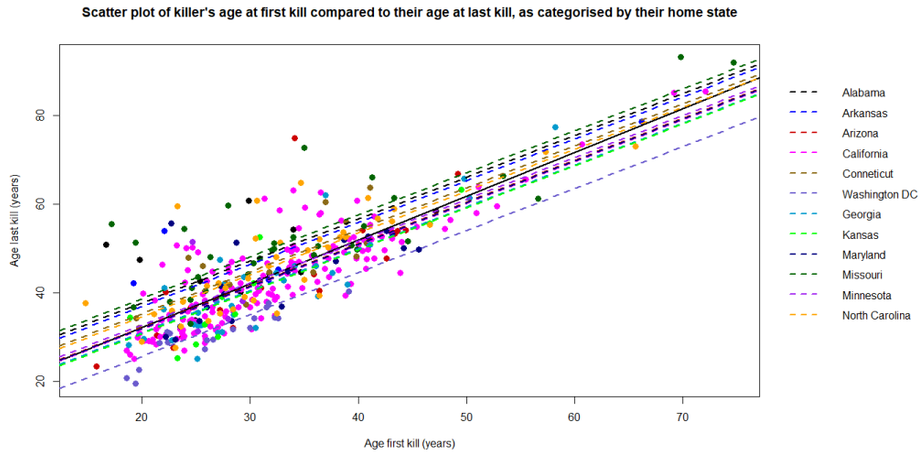

Figure 5: Scatter plot of killer's age at last kill against age at first kill, with a black solid regression line as found in our previous model which ignored clustering. The dashed coloured lines show the regression lines corresponding to the state each killer committed murder in, as determined by the random intercept model.

Likelihood Ratio Test. We conduct a likelihood ratio test to decide whether there is sufficient evidence to select this multi-level model B over the single-level linear regression model A. As the only additional parameter in model B is the level two variance parameter, σ^2_{u0}, we investigate the following hypotheses:

$$H_0 : \sigma^2_{u0} = 0 \quad vs. \quad H_1 : \sigma^2_{u0} > 0. \tag{4}$$

From conducting this test we find the p-value at a 5% significance level is $p = 3.4e^{-12}$ (2 s.f.), meaning we have strong evidence to suggest model B is an improvement to model A at the 5% significance level.

Potenial Covariates. Despite this result suggesting model B is a better fit than alternative models, it can be seen the plotted points in figure 5 are not necessarily close to their corresponding state-clustered regression line. For example, many of the plots for serial killers in Missouri lie above the corresponding dark green dashed regression line. This suggests that the killer's state can only explain some of the variability we see in the killers age at last kill. Therefore, we may investigate the influence of other covariates on the killers age at last kill to find a better fitting model for the data. Narrowing down to the three covariates - sex, race and sexual preference, we will choose to use one or more of these in our final model. To decide which of these to include, let us first compare the box plots of killers age at last kill for these covariates, as seen in figure 6.

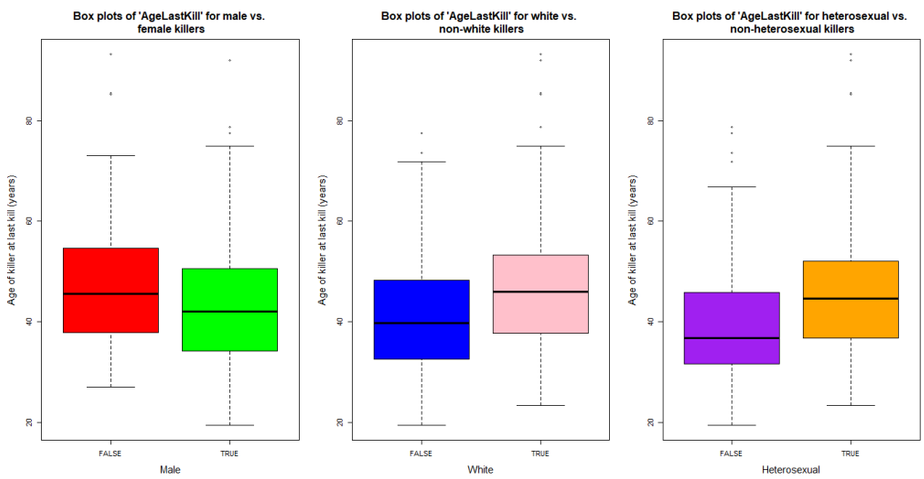

Figure 6: Box plots for three covariates, sex, race and sexual preference, respectively.

We see in all three plots that the mean age of last kill varies between those that are male and female, white and non-white, heterosexual and non-heterosexual. Albeit, we see slightly more difference in the spread and location of the age of last kill between the blue and pink box plots and between the orange and purple box plots, corresponding to the sample when partitioned by race and sexual preference respectively, in comparison to the box plots partitioned by sex. As this is only by eye and is subjective, we shall investigate which of these two covariates has the greatest influence on killer's age at last kill in an objective way. This is done by creating three linear regression models comparing the killers age at last kill with each of these covariates and seeing which has the smallest p-value using the 'summary' function on each model. We find that when comparing 'AgeLastKill' with the killer's sex, $p = 0.0053$ (2 s.f.), when comparing sexual

preference, $p = 5.93e^{-5}$ and when comparing race, $p = 7.47e^{-6}$. Therefore, as the p-value is smallest for race, we shall start with adding this covariate to our random intercept model.

Random Intercept Model with an additional covariate. We define another random intercept model, model C, that includes the influence of race on the killers age at last kill. This model is given as follows,

$$\text{Model C:} \quad y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_{0j} + e_{0ij} \tag{5}$$

where x_{2ij} denotes whether the i-th killer from the j-th state is white or not, 1 if they are white and 0 if they are non-white.

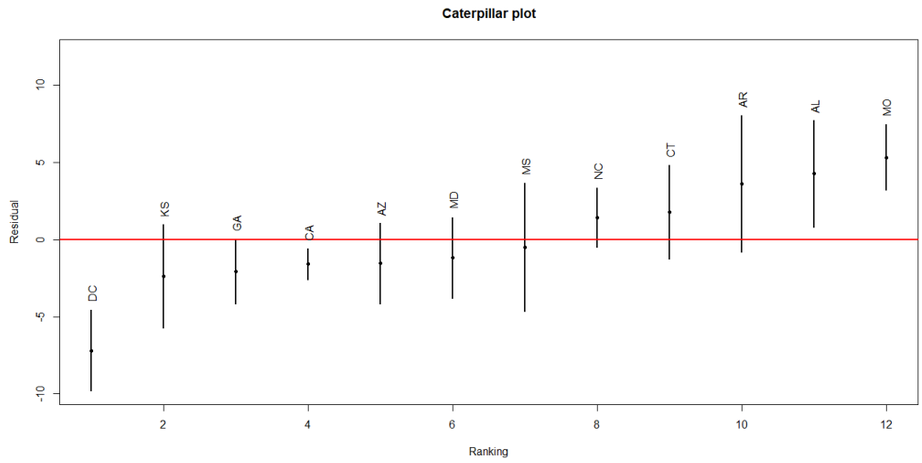

Figure 7: Caterpillar plot showing the level 2 residuals for model C, specific to each state, with 95% confidence intervals. The horizontal red line indicates the average state which would have a level 2 residual equal to zero.

We can compare the effects of the different clusters on killers age at last kill by creating a caterpillar plot based on this model, as seen in figure 7. The plot shows two clusters in particular that have residuals that stand out from the other clusters, these being the DC and MO states, for which the intervals of these residuals do not contain zero. Therefore, these states appear to be significantly different to the average state when considering the killer's age at last kill. Despite this, the intervals for most states are quite wide, especially for AR and MS, suggesting there is an element of uncertainty in our estimates of the state effects, therefore we may want to develop a more complex model to investigate these effects further.

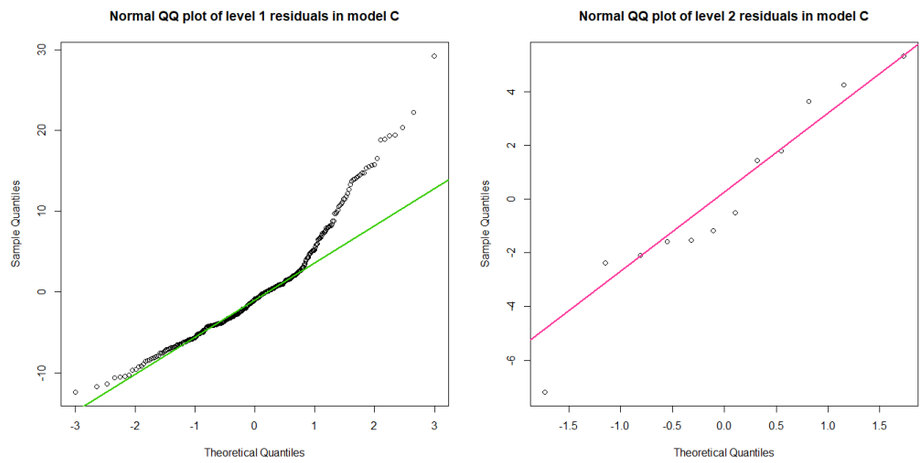

Figure 8: Normal Q-Q plots for level 1 residuals (left), fitted with a green Q-Q line and level 2 residuals (right), fitted with a pink Q-Q line for model C.

Furthermore, we can also assess the normal Q-Q plots to evaluate the assumption that the residuals are from a normal distribution, as shown in figure 8. As we can see, the level 1 residuals appear to follow the pink line, suggesting these are normally distributed, however the level 2 residuals show a slight curvature away from the green line. This suggests we cannot be certain that these residuals are based on a normally distributed variable.

Confidence Intervals on Random Intercept Models. Let us suppose model D, where we investigate the effects of race and sexual preference on the killers age at last kill, clustered by state, i.e.

$$\text{Model D:} \quad y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_{0j} + e_{0ij} \tag{6}$$

where x_{1ij} and x_{2ij} denote the race and the sexual preference of the i-th killer in the j-th state, respectively. We have estimates for the random effects, $\hat{\sigma}^2_{u0} = 4.22^2$ and $\hat{\sigma}^2_{e0} = 10.8^2$ as well as the estimated parameters, $\hat{\beta}_0 = 39.5$, $\hat{\beta}_1 = 4.32$ and $\hat{\beta}_2 = 3.80$. We first investigate the hypothesis that neither race nor sexual preference have any impact on the killer's age at last kill, i.e. $H_0 : \beta_1 = \beta_2 = 0$. Creating an appropriate

matrix C (a 3x1 matrix) that corresponds to this hypothesis, $C = [0\,1\,1]$, and using the 'qchisq' function within R to give us the correct quantile, we calculate the Mahalanobis distance, $D(C\hat{\beta}, k)$ and compare this to $d = \sqrt{\chi^2_q(\alpha)}$ at the 5% significance level. We find that $D = 5.02$ and $d = 2.45$, which means that k lies outside of the confidence region. Therefore we have sufficient evidence to reject the null hypothesis at the 5% significance level. Next we investigate the hypothesis that race and sexual preference combined will have an influence on 'AgeLastKill', i.e. whether the age at last kill of a killer that is both white and heterosexual varies from the rest of the sample. This is done by defining the matrix $C = [0\,1\,1]$ and using the 'qnorm' function within R to test at the correct significance level. Using our calculated values for $\hat{\sigma}_{u0}$ and $\hat{\sigma}_{u0}$, we test the null hypothesis $H_0 : \beta_1 = 0$ at the 10% significance level and find that the confidence interval equals $(5.44, 10.8)$. As zero does not fall within this interval, we have sufficient evidence to reject the null hypothesis at the 10% significance level in this model, i.e. there is reasonable evidence to suggest that race and sexual preference combined have an influence on killer's age at last kill. Furthermore, we find this continues to hold at the 5% significance level, where the confidence interval is $(4.93, 11.3)$. Similarly, we can investigate the hypothesis that the two covariates have an equal influence on 'AgeLastKill', i.e. $H_0 : \beta_1 - \beta_2 = 0$ by defining $C = [0\,1\,-1]$ and carrying out the same test. We find the confidence interval to be $(-3.10, 4.14)$ at the 5% significance level and still $(-2.52, 3.56)$ at the 10% significance level. As zero falls within both of these categories, we cannot reject the null hypothesis at either significance level, suggesting that these two covariates have similar influence on the killer's age at last kill.

Random Slope Model. Let us suppose the killer's age at first kill varies between states also, then we can extend model B to make a random slope model, defined as model E,

$$\text{Model E:} \quad y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_{0j} + u_{1j} x_{1ij} + e_{0ij} \tag{7}$$

with the added random slope of age of first kill of killer i in state j, u_{1j} and the covariate of killer's race, x_{2ij}. We find this model to be an improvement to model B at the 5% significance level, with a p-value of 0.00560, found using the adjusted p-value calculation for 1 additional covariate and 1 additional variance parameter function. The regression lines for this model over the age last kill vs. age first kill scatter plot are shown in figure 9. As this is a random slope model we see that the regression lines vary in intercepts and slopes, depending on the killer's state of murder. We also see the lines vary between those in the 'White' and 'Non-white' categories, where the age of last kill is slightly higher for the killer's that are 'Non-white' compared to those that are 'White'. The killer's from Washington DC and Kansas in particular are shown to have a lower 'AgeLastKill' age and those from Missouri and Maryland whom have a higher age of last kill, which is the same pattern as seen in figure 4 based on model B, further supporting this claim.

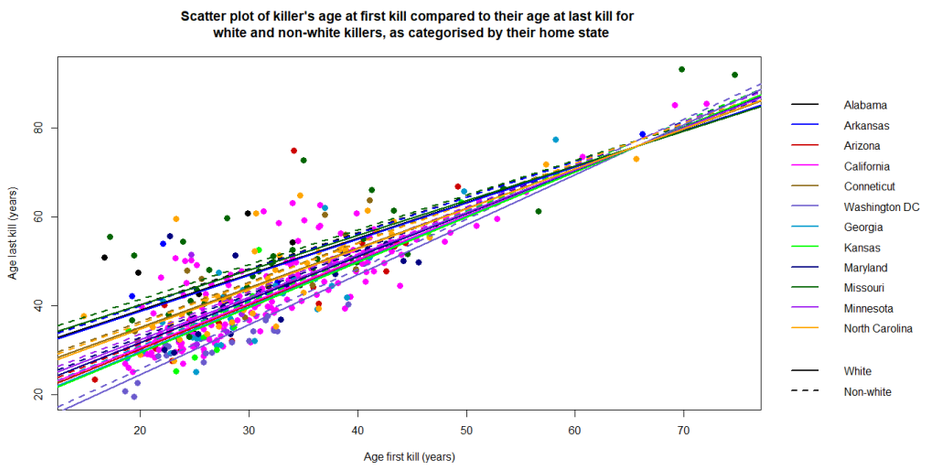

Figure 9: Plot of killer's age at last kill against age at first kill, where the coloured lines show the regression lines corresponding to the state each the killer committed murders in, with dashed lines for non-white killers and solid lines for white killers.

We compare the effects of clustering on killer's age at last kill by creating a caterpillar plot for the level one residuals and level two residuals of this model, as seen in figure 10. The left plot shows two clusters in particular that lie very far from the purple line, these being the DC and MO states, where the intervals of these residuals do not contain zero. This means that these states are significantly different to the average state when considering the killer's age at last kill and race. Similarly in the left figure, we see these states also differ from the purple line, however in the opposite directions to the right plot. In addition, in both figures the intervals for most states are quite wide, especially for AR and MS, suggesting there is an element of uncertainty in our estimates of the state effects on the killer's age at last kill.

Albeit, despite figures 9 and 10 showing a seemingly good fit to the data for this model, we run into an error message when putting this function into R - the boundary fit is singular. This means that the choice of 'AgeFirstKill' as a random effect is too complex and the variance component of this is close to zero, as can be seen from the normal Q-Q plot of these residuals in figure 11 where the sample quantities on the y-axis show a very small range between $(-0.15, 0.2)$.

Furthermore, defining model F by adding a third random effect based on the 'race' variable to model E, $u_{2j} x_{2ij}$, we get a new error - that the model 'failed to converge'. This is possibly due to the two random effects being on very different scales, as age of first kill can vary from roughly 15 to 75, while race is either 1 or 0. Therefore changing the 'AgeFirstKill' variable to be closer to the 'race' covariate may remove this

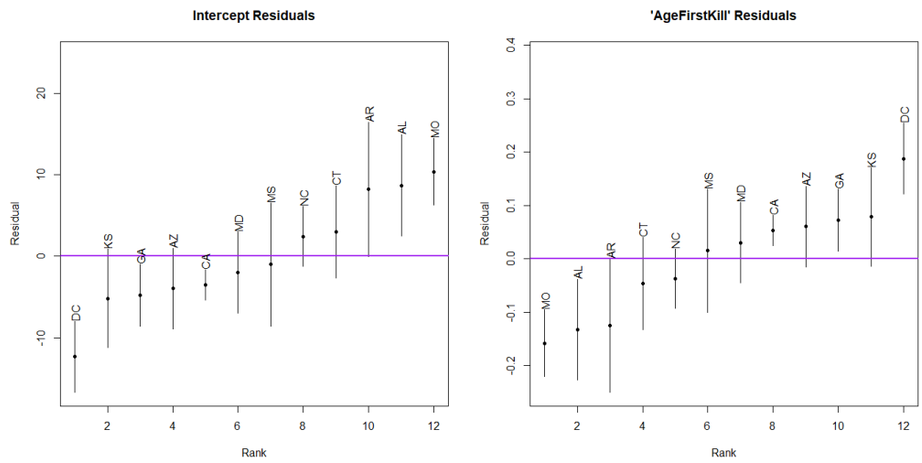

Figure 10: Caterpillar plots for the level one and level two residuals for model E, specified by killer's state with a 95% confidence interval. The horizontal purple lines signify the average state, which would have a residual of zero.

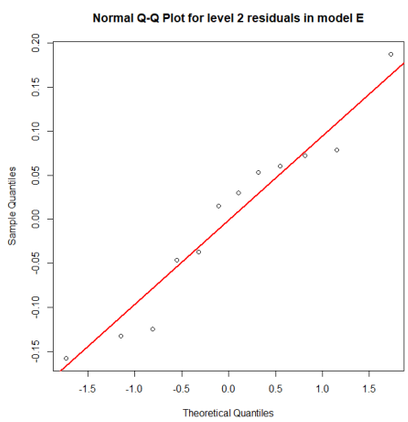

Figure 11: Normal Q-Q plot of level 2 residuals for model E, with a red Q-Q line.

error and we do this by defining a new variable, 'AgeFirstKillDecades'. This divides the original age at first kill variable by 10 so that the answer is given in decades instead of years and we use this variable within the clustered part of the model F instead of 'AgeFirstKill' to produce a new model, model G. Although this works to remove the 'failure to converge' error, this model still runs into an issue of having a singular boundary fit.

3 Discussion

In conclusion, we find from this sample that the average killer's age at last kill is 43.8 years and that there is a positive correlation between a killer's age at first kill and at last kill. Furthermore, we have concluded, by the use of likelihood ratio tests that the killer's state of murder has an influence of the relationship between the killer's age at first kill and age at last kill, with a 95% likelihood of this being true. In particular, killer's in states like Washington DC and Kansas have a higher average age at last kill, compared to those from Missouri or Maryland, perhaps as police force receives greater funding in some states, meaning they have the more resources to help capture killer's sooner. Furthermore, we concluded that race and sexual preference have a strong influence on the killers age at last kill, as well as sex having a slightly weaker but still significant influence. We find from the scatter plots fitted with regression lines for each state that clustering by state has a strong influence on the killer's age at the time of their last kill, as determined by models B and C. Furthermore, by the use of confidence regions, we concluded that race and sexual preference have a similar influence on the killer's age at the time of their last kill.

Despite many strong conclusions being made from this investigation, a weakness of this analysis is the use of the 'unknown' answer in the sexual preference covariate. We assumed that these individuals did not fit into the 'Heterosexual' category which may be an incorrect assumption to make, as there is still a very real possibility that these individuals were heterosexual, it is just unknown. Removing these individuals from the sample may give greater accuracy, however as we already were dealing with a relatively small sample size, these individuals were kept within the analysis. Another weakness of this investigation is that we experienced models that did not converge, like model F. However, we were able to overcome this by rescaling the 'AgeFirstKill' variable. Nevertheless, we had a singular boundary fit error occur in models E and G, which may be due to the small sample size we are studying, as a larger sample should allow for more complex models to be made without experiencing the issue of a singular fit.

A R Code

```
setwd("C:/Users/fayew/Documents/_Year 4 Semester 2/5092M Mixed Models with Medical Applications/
Practicals")
load(file = "killers2.Rdata")
makesample(8646)
save(CWsample, file = "CWsample.RData")
table(CWsample$State)
attach(CWsample)

#Defining new variables
CWsample$Male <- CWsample$Sex == "Male"
CWsample$White <- CWsample$Race == "White"
CWsample$Hetero <- CWsample$SexualPreference == "Heterosexual"

CWsample$States <- as.numeric(CWsample$State)
colours <- c("black", "blue", "red3", "magenta",  "goldenrod4", "slateblue", "deepskyblue3",
"green", "navy", "darkgreen", "purple", "orange")
States_list <- unique(paste(CWsample$State))
States_full <- c("Alabama", "Arkansas", "Arizona", "California", "Conneticut", "Washington DC",
"Georgia", "Kansas", "Maryland", "Missouri", "Minnesota", "North Carolina")

#FIGURE 1
par(mfrow = c(1,2))
hist(CWsample$AgeLastKill,
     xlab = "Age of killer at time of last murder (years)",
     main = "Histogram plot of age of killer at time of last murder",
     col = "#EBF0FF")
qqnorm(CWsample$AgeLastKill, main="QQ plot of age of killer at time of last murder")

#LINEAR MODEL
model0 <- lm(AgeLastKill ~ 1, data = CWsample)
summary(model0)
betahat <- model0$coefficients

#LINEAR REGRESSION MODEL
modelA <- lm(AgeLastKill ~ AgeFirstKill, data = CWsample)
summary(modelA)

#FIGURE 2
par(mfrow = c(1, 2))
plot(CWsample$AgeLastKill ~ CWsample$AgeFirstKill,
     main = "Scatter plot of age of killer at time of last \n murder vs. age at time of first
     murder",
     ylab = "Age Last Kill (years)",
     xlab = "Age First Kill (years)")
abline(modelA, lwd = 2, col = "purple")
qqnorm(modelA$resid, main="Normal QQ plot of age of killer at time of last \n murder vs. age at
time of first murder")
qqline(modelA$resid, col="orange", lwd=2)

#FIGURE 3
library(knitr)
install.packages("kableExtra")
library(kableExtra)
State_tab <- table(t(CWsample$State))
kbl(t(State_tab), align="l",  caption = "Frequency table of Killer's home state") %>%
  kable_styling( full_width = F)

#FIGURE 4
par(mfrow=c(1,1))
par(mar = c(5, 4, 4, 9))
plot(CWsample$States ~ CWsample$AgeLastKill,
     xlab = "Age Last Kill (years)",
     ylab = " ", main = " ", pch = 16,
     yaxt = "n", ylim = c(0, 12),
     col = colours[CWsample$States])

axis(4, at = 1:12, labels = States_full, las = 2)
abline(v = betahat, lty = 2)
axis(3, at = betahat, labels = expression(paste(widehat(beta)[0], " = 43.83")))
```

```
#RANDOM INTERCEPT MODEL
library("lme4")
modelB <- lmer(AgeLastKill ~ AgeFirstKill + (1|State), data = CWsample, REML=FALSE)
summary(modelB)
randomeffects <- ranef(modelB)
u <- randomeffects$State$"(Intercept)"
beta0 <- modelB@beta[1]
beta1 <- modelB@beta[2]

#FIGURE 5
par(mfrow=c(1,1))
par(mar = c(5,5,5,13))
plot(CWsample$AgeLastKill ~ CWsample$AgeFirstKill,
     ylab = "Age last kill (years)",
     xlab = "Age first kill (years)",
     main = "Scatter plot of killer's age at first kill compared to their age at last kill,
     as categorised by state",
     pch = 16,
     cex = 1.2,
     col = colours[CWsample$States])
legend(80,95, cex = 1, bty = "n", ncol=1, legend = States_full, x.intersp=0.5, seg.len=1.5,
col = colours, lty=2, lwd=2, xpd = TRUE)
abline(modelA, lwd = 2)
for(j in 1:12){
  abline(a = beta0 + u[j], b = beta1, col = colours[j], lty=2, lwd=2)
}

#LIKELIHOOD RATIO TEST
A_log = logLik(modelA)
B_log = logLik(modelB)
chi = 2*(B_log - A_log)
p = 1-pchisq(chi, df=1)

#Equivalently
anova(modelB,modelA)

#FIGURE 6
par(mfrow = c(1,3))
boxplot(CWsample$AgeLastKill ~ CWsample$Male, cex.main=1.5, cex.lab=1.4, col=c("red", "green"),
main = "Box plots of 'AgeLastKill' for male vs. \n female killers", xlab="Male" , ylab="Age of
killer at last kill (years)")
boxplot(CWsample$AgeLastKill ~ CWsample$White, cex.main=1.5, cex.lab=1.4, col=c("blue", "pink"),
main = "Box plots of 'AgeLastKill' for white vs. \n non-white killers", xlab="White", ylab="Age of
killer at last kill (years)")
boxplot(CWsample$AgeLastKill ~ CWsample$Hetero, cex.main=1.5, cex.lab=1.4, col=c("purple", "orange"),
main = "Box plots of 'AgeLastKill' for heterosexual vs. \n non-heterosexual killers", xlab=
"Heterosexual", ylab="Age of killer at last kill (years)")

#POTENTIAL COVARIATES
modelAm <- lm(AgeLastKill ~ Male, data = CWsample)
summary(modelAm)
modelAw <- lm(AgeLastKill ~ White, data = CWsample)
summary(modelAw)
modelAh <- lm(AgeLastKill ~ Hetero, data = CWsample)
summary(modelAh)

#RANDOM INTERCEPT MODEL WITH AN ADDITIONAL COVARIATE
modelC <- lmer(AgeLastKill ~ AgeFirstKill + White + (1 | State), data = CWsample, REML = FALSE)
summary(modelC)

#FIGURE 7
beta0 <- modelC@beta[1]
beta1 <- modelC@beta[2]
beta2 <- modelC@beta[3]
randomeffects <- ranef(modelC)
u <- randomeffects$State$"(Intercept)"

#Posterior variances:
str(attr(randomeffects$State, "postVar"))
v <- attr(randomeffects$State, "postVar")[1, 1, ]

group <- rownames(randomeffects$State)
```

```
# Create data frame:
level2 <- data.frame(group, u, v)
# Name the columns:
colnames(level2) <- c("Group", "Residual", "PostVar")
# Order by size of the residuals:
level2 <- level2[order(level2$Residual), ]
# Include a column containing the ranks:
level2 <- cbind(level2, 1:nrow(level2))
# Name the new column:
colnames(level2)[4] <- "Rank"
level2$Lower <- level2$Residual - qnorm(0.975)*sqrt(level2$PostVar)
level2$Upper <- level2$Residual + qnorm(0.975)*sqrt(level2$PostVar)

plot(level2$Rank, level2$Residual,
     xlab = "Ranking", ylab = "Residual",
     pch = 20, main = "Caterpillar plot",
     ylim = c(min(level2$Lower), 1.5*max(level2$Upper)))
# Plot the credible intervals:
segments(level2$Rank, level2$Lower, level2$Rank, level2$Upper, lwd = 2)
abline(h = 0, lty=1, col="red", lwd=2)
# Add group names to the plot:
groupname = paste(level2$Group)
text(x = 1:12 , y = level2$Upper, labels = groupname, srt = 90,  adj = -0.5, cex = 1)

#FIGURE 8
par(mfrow = c(1, 1))
par(mfrow = c(1, 2))
qqnorm(resid(modelC), main="Normal QQ plot of level 1 residuals in model C")
qqline(resid(modelC), col=c("#33CC00"), lwd=2)
qqnorm(level2$Residual, main="Normal QQ plot of level 2 residuals in model C")
qqline(level2$Residual, col="#FF3399", lwd=2)

#CONFIDENCE INTERVALS ON RANDOM INTERCEPT MODELS
modelD <- lmer(AgeLastKill ~ White + Hetero + (1 | State), data = CWsample, REML = FALSE)
summary(modelD)

sigmahat_u0 <- 4.224
sigmahat_e0 <- 10.775
betahat <- matrix(modelD@beta, nrow = 3, ncol = 1)

#WHAT ABOUT beta_1=beta_2=0? 5%
n <- nrow(CWsample)
intercept <- rep(1, n)
covariates <- CWsample[, c("White", "Hetero")]
X <- cbind(intercept, covariates)
X <- as.matrix(X)
I <- diag(n)
G <- matrix(NA, nrow = n, ncol = n)
for(h in 1:n){
  for(k in 1:n){

    G[h, k] <-  as.numeric(CWsample$State[h] == CWsample$State[k])
  }
}
Sigmahat <- (sigmahat_e0^2)*I + (sigmahat_u0^2)*G

q <- 2
d <- sqrt(qchisq(1 - 0.05, df = q))
C <- matrix(data = 0, nrow = 2, ncol = 3)
C[1, 2] <- 1
C[2, 3] <- 1
C
k = matrix(c(0,0), ncol = 1, nrow = 2)
k
V <- C %*% solve(t(X) %*% solve(Sigmahat) %*% X) %*% t(C)
D =  sqrt( t(C%*%betahat - k) %*% solve(V) %*% (C%*%betahat - k) )
D > d
D
d

#H_0 : beta_1+beta_2=0, 10% significance
C <- matrix(c(0,1,1), nrow = 1, ncol = 3)
```

```
r <- qnorm(1 - 0.1/2)
lower <- C %*% betahat - r * sqrt(C %*% solve(t(X) %*% solve(Sigmahat) %*% X) %*% t(C))
upper <- C %*% betahat + r * sqrt(C %*% solve(t(X) %*% solve(Sigmahat) %*% X) %*% t(C))
c(lower, upper)

#5% significance
C <- matrix(c(0,1,1), nrow = 1, ncol = 3)
r <- qnorm(1 - 0.05/2)
lower <- C %*% betahat - r * sqrt(C %*% solve(t(X) %*% solve(Sigmahat) %*% X) %*% t(C))
upper <- C %*% betahat + r * sqrt(C %*% solve(t(X) %*% solve(Sigmahat) %*% X) %*% t(C))
c(lower, upper)

#H_0 : b_1-b_2=0, 10% significance level
r <- qnorm(1 - 0.1/2)
C <- matrix(c(0,1,-1), nrow = 1, ncol = 3)
l <- C %*% betahat - r * sqrt(C %*% solve(t(X) %*% solve(Sigmahat) %*% X) %*% t(C))
u <- C %*% betahat + r * sqrt(C %*% solve(t(X) %*% solve(Sigmahat) %*% X) %*% t(C))
c(l, u)

#5% significance level
r <- qnorm(1 - 0.05/2)
C <- matrix(c(0,1,-1), nrow = 1, ncol = 3)
l <- C %*% betahat - r * sqrt(C %*% solve(t(X) %*% solve(Sigmahat) %*% X) %*% t(C))
u <- C %*% betahat + r * sqrt(C %*% solve(t(X) %*% solve(Sigmahat) %*% X) %*% t(C))
c(l, u)


#RANDOM SLOPE MODEL
modelE <- lmer(AgeLastKill ~ AgeFirstKill + White + (1 + AgeFirstKill| State), data = CWsample,
REML = FALSE)
#WARNING/ERROR MESSAGE
summary(modelE)

E_log = logLik(modelE)
B_log = logLik(modelB)
chi = 2*(E_log - B_log)
p1 = 1 - pchisq(chi, 1)
p2 = 1 - pchisq(chi, 2)
p = (p1 + p2)/2

#FIGURE 9
beta0 <- modelE@beta[1]
beta1 <- modelE@beta[2]
beta2 <- modelE@beta[3]
randomeffects <- ranef(modelE)
u0 <- randomeffects$State$"(Intercept)"
u1 <- randomeffects$State$"AgeFirstKill"

plot(CWsample$AgeLastKill ~ CWsample$AgeFirstKill,
     ylab = "Age last kill (years)",
     xlab = "Age first kill (years)",
     main = "Scatter plot of killer's age at first kill compared to their age at last kill, as
     categorised by their home state",
     pch = 16,
     cex = 1.2,
     col = colours[CWsample$States])

legend(75,100, cex = 1, bty = "n", ncol=1, legend = States_full, x.intersp=0.2, seg.len=1.5,
col = colours, lty=1, xpd = TRUE)
legend(75,20, bty="n", cex=1, legend = c("White", "Non-white"), x.intersp=0.2, seg.len=1.5,
lty=c(1:2), xpd=TRUE)
for(j in 1:12){
  abline(a = beta0 + beta2 + u0[j], b = beta1 + u1[j], col = colours[j], lty = 2)
}
for(j in 1:12){
  abline(a = beta0 + u0[j], b = beta1 + u1[j], col = colours[j])
}

#FIGURE 10
randomeffects$State
str(attr(randomeffects$State, "postVar"))
attr(randomeffects$State, "postVar")
```

```
varu0 <- attr(randomeffects$State, "postVar")[1,1 , ]
varu1 <- attr(randomeffects$State, "postVar")[2,2 , ]

group <- rownames(randomeffects$State)
level2 <- data.frame(group, u0, u1, varu0, varu1)
colnames(level2) <- c("Group", "Intercept Residuals", "'AgeFirstKill' Residuals", "IntPostVar",
"AFKPostVar")
level2


par(mfrow = c(1, 1))
par(mfrow = c(1, 2))
for(i in 1:2){

  # Order by the rankings of the effects in column i+1:

  level2 <- level2[order(level2[ , 1 + i]), ]

  # Calculate the credible intervals for the
  # effects in column i+1:

  Lower <- level2[ , i + 1] - qnorm(0.975)*sqrt(level2[ , i + 3])
  Upper <- level2[ , i + 1] + qnorm(0.975)*sqrt(level2[ , i + 3])

  # Plot the effects against the rankings:

  plot(1:12, level2[ , i + 1], pch = 20,
       xlab = "Rank", ylab = "Residual",
       main = names(level2)[i + 1],
       ylim = c(min(Lower), 1.5*max(Upper)))

  # Plot the credible intervals:

  segments(1:12, Lower, 1:12, Upper)

  abline(h = 0, col = "purple", lwd=2)

  groupname = paste(level2$Group)

  text(x = 1:12 , y = Upper, labels = groupname, srt = 90,  adj = 0, cex = 1)

}

#FIGURE 11
par(mfrow = c(1, 1))
qqnorm(u1, main="Normal Q-Q Plot for level 2 residuals in model E")
qqline(u1, col=c("red"), lwd=2)


#ERRORS
modelF <- lmer(AgeLastKill ~ AgeFirstKill + White + (1 + AgeFirstKill + White| State),
data = CWsample, REML = FALSE)
summary(modelF)

CWsample$AgeFirstKilldecades <- CWsample$AgeFirstKill/10
modelG <- lmer(AgeLastKill ~ AgeFirstKill + White + (1 + AgeFirstKilldecades + White | State),
data = CWsample, REML = FALSE)
summary(modelG)
```

Public Health Intervention - Reducing murder by delaying killer's age at first kill

MATH5092M: Mixed Models with Medical Applications Report

Faye Williams
201308646

Department of Mathematics
University of Leeds
United Kingdom
May 2023

1 Trial design

This section will detail the trials primary endpoint as well as it's proposed design to reach this endpoint. The underlying outcome model shall be specified, in addition to primary analysis of this model. Any assumptions that have been made within this model will be specified, and the implications of their violations shall be discussed. Lastly, the reasoning for the chosen trial design shall be specified, which will explain why this design was chosen over other potential designs.

1.1 Endpoint

The primary endpoint of this investigation is to determine whether implementing the new public health intervention leads to a higher age at which serial killers commit their first kill, in comparison to no intervention being made.

1.2 Trial Design

To investigate the effectiveness of this intervention in delaying serial killer's age at first kill, a clustered randomised trial design shall be used, which is clustered by the 12 US states included in the 'CWSample' sample, taken from the 'killers2.Rdata' dataset. This dataset consists of personal information concerning 366 serial killers from 12 US, as well as information regarding their crimes. This includes the serial killer's ID numbers, the two letter abbreviation of the US state they committed their murders in, the killers age at the time of their first kill and the year of their first kill, which ranges from 1908-2015. Any R code featured in the remainder of the report can be found in Appendix A.

As the intervention is at state-level, it involves not just focusing on the individuals who are at high risk of committing these crimes but instead the whole population in that state. Therefore data shall be collected on the murders committed by first-time killer's in each of these states as they occur and the killer's age at the time of the first kill. As every participant within a cluster is in the same arm of the trial, we say these clusters are nested within the intervention and can call this a nested design. Further, it will be assumed here that the trial designs are balanced, such that each arm has the same number of clusters, meaning there will be 6 clusters in each arm, and each arm will be assumed to have the same number of participants, n. The time period of this study shall be determined by however long it takes to reach the required amount of participants in each arm of the study, i.e. have n new murderers arise in the six intervention states and n new murderers in the six non-intervention states. Once all the data has been collected, a model shall be fitted on the age of first kill of those in the non-intervention clusters and another on those in the intervention clusters. Creating two linear models from this, similar to the model seen in equation (3) and comparing the estimated population means, β_0 shall indicate to us the average age of first kill in both arms of the trial. From here we can compare these ages to see if the age of first kill is higher in the intervention group than the non-intervention group.

The clustered randomised trial design involves intervention being introduced either for all individuals in one state or for no individuals in that state. This involves the six states being matched together in pairs, where one of the two states will be randomly chosen for intervention and the other will act as a control state. The pairs will be chosen based on the state's average ages at first kill and these values are shown in the table in figure 1 below. Ranking these states according to average age of first kill, the first pairs will be the two states with the highest average age at first kill, then the next two and so on. Therefore, the pairs chosen will be AR and MO, NC and AZ, MD and CA, CT and GA, AL and KS, MS and DC. From here the names of the pairs of states will be put into a random choice generator and whichever state is chosen first will be chosen to have intervention. For example, for the first pair AR and MO, the two state names will be put into an online random choice generator wheel and the first state that is produced will be chosen to have intervention, while the state that isn't selected shall not introduce any intervention.

State	AL	AR	AZ	CA	CT	DC	GA	KS	MD	MO	MS	NC
AFK	29.35	34.36	32.38	31.90	31.41	26.44	30.94	30.30	32.25	34.79	29.25	33.08

Figure 1: Table of the killers average age at first kill in each state, rounded to two decimal places.

One key reason for choosing a clustered randomised trial over an individually randomised trial design is that is it more convenient to apply the intervention either entirely to a state or not at all, which would not be the case if we chose to split each state into two intervention arms. Further, having both intervention and non-intervention individuals from the same state may cause issues of individuals that haven't received intervention influencing their peers who have in the same state, or vice versa. This could greatly impact the behaviour of individuals in that state, as those receiving intervention may encourage those that are not to change their attitudes, therefore decreasing their likelihood to kill. Therefore, if we choose an individually randomised trial design, the data we receive may not be as accurate, due to the influence of other's behaviour. In addition, we want to examine how the implementation of this intervention impacts murder rates within a state overall, which is easier to examine when individuals in one state are all receiving the same treatment.

Another alternative design would involve using the sample of the 'killers2.Rdata' dataset as the control group and applying the intervention to all states and recording the murders that happen. Although this would reduce the amount of data we need to collect, therefore saving time and money, we would end up having data on the age at first kill of serial killers from completely different timelines. It would take many years to collect the new data from the intervention group and as the control group data has already been recorded from 1908-2015, we would see a large gap with no overlap of the recorded years of murders. Many

covariates could influence how the average age of last kill would change over this time frame, such as humans living longer in more recent years. Therefore, it is hard to accurately compare data from these two periods of time and instead it is best to record the data for the control group and intervention group at the same time.

1.3 Outcome Model

Assuming each arm has the same number of clusters, which we have chosen to be $m = 6$ in each and assuming that each arm has the same number of participants n, the outcome model implied by this trial design is given in the equation (1). The response variable y_{ijk} denotes the age of last kill of the i-th killer in the j-th cluster in the k-th arm of the trial, where $i = 1, ..., p$, $j = 1, ..., m$ and $k = 0, 1$. Therefore, we have $n = 6p$ participants in each arm of the trial, taken from the p participants in each of the 6 states in each arm.

$$y_{ijk} \sim N(\mu_k + \mu_j, \sigma_e^2) \quad \text{with} \quad u_j \sim N(0, \sigma_u^2). \tag{1}$$

Writing this model in the style of a random intercept model with no additional covariates, we have

$$y_{ijk} = \beta_0 + \beta_1 x_k + u_j + e_{ijk} \quad \text{with} \quad u_j \sim N(0, \sigma_u^2) \quad \& \quad e_{ijk} \sim N(0, \sigma_e^2), \tag{2}$$

where x_k is a binary indicator of intervention arm, such that $x_k = 1$ for those receiving intervention and 0 for those not receiving intervention.

2 Sample size requirements

This section will first determine the parameter estimations that will be required to perform sample size calculation. This sample size calculation will be demonstrated using the necessary formula(e) as well as using R. Lastly, we shall evaluate the influence changing the type I and type II error rates has on the required sample size, as well as changing the treatment effect and variance assumptions.

2.1 Parameters

We shall use the data within the 'CWsample' to obtain estimates for the variance of the outcome model implied by the chosen trial design, as well as the estimated variance of the killer's state of murder. Assuming that killer's age at first kill is a normally distributed variable, we find the variance, σ^2, for this sample using the basic linear model,

$$y_i = \beta_0 + e_i \quad \text{with} \quad e_i \sim N(0, \sigma^2), \tag{3}$$

Putting this model into R using the 'lm' function, we find the estimate for the population mean is $\beta_0 \approx 31.8$ years (3 s.f.), meaning the average killer's age at first kill is around 32 years of age. Note that estimates will be rounded to three significant figures, unless stated otherwise. In addition, the function gives an estimated variance of $\sigma^2 \approx 9.62$. Furthermore, as we have assumed u_j is normally distributed with a variance of σ_u^2 in equations (1) and (2), we find this variance by creating the linear model of the killer's state of murder, given by

$$u_j = \beta_0 + e_j \quad \text{with} \quad e_j \sim N(0, \sigma_u^2), \tag{4}$$

for some estimated population mean β_0. Putting this model into R gives us an estimated variance of $\sigma_u^2 \approx 3.09^2$, which will be used within the sensitivity analysis to calculate the intracluster correlation coefficient.

2.2 Sample size

To begin, we calculate the sample size required for type I and II error rates of 0.025 (one-sided) and 0.2, respectively, and will then explore other choices of nominal error rates and how this affects the required sample size. The required sample size is calculated using the following formula, as provided in the answer to question 7.3 of the MATH5092 exercises [1],

$$n = \left(\frac{\sqrt{2\sigma^2}(z_{1-\alpha} - z_\beta)}{\delta_1 - \delta_0} \right)^2 \tag{5}$$

As we are aiming to increase the average age of last kill by 2.5 years, we have the null hypothesis $H_0 : \delta = \delta_0$ and alternative hypothesis $H_1 : \delta = \delta_1$. Here δ denotes the treatment effect and is given by $\delta = \mu_0 - \mu_1$, which denote the average of the first kill for those before intervention and after. As we are aiming to increase the age of first kill in murderers by 2.5 years, the difference between these will be 2.5, meaning we define the original treatment effect before intervention as $\delta_0 = 0$ and the proposed treatment effect to be $\delta_1 = 2.5$. Substituting in the type I and type II errors, $\alpha = 0.025$ (one-sided) and $\beta = 0.2$ respectively, our estimated value for variance, σ^2, and the values for δ_0 and δ_1, we find the calculated sample size to be

$$n = \left(\frac{\sqrt{2(9.62)^2}(z_{1-0.025} - z_{0.2})}{2.5 - 0} \right)^2 = 232.51 \, (5 \text{ s.f.}) \tag{6}$$

where $z_x = \Phi^{-1}(x)$ is given by the inverse standard normal distribution and here $z_{0.975} \approx -1.96$ and $z_{0.2} \approx 0.842$. As this value is the minimum number of participants required, we must round to give a required $n = 233$ participants for the specified type I and type II errors. This can also be calculated within R using

the 'power.t.test' function, where we specify $\delta = 2.5$, standard deviation $\sigma = 9.62$ and power $1 - \beta = 0.8$. We find the required sample size to be $n = 233.4$ (4 s.f.) therefore using this function we need $n = 234$ participants. The slight discrepancy between number of participants required in each arm can be explained by taking the estimations of inverse standard normal distribution, rather than the exact values of these that would be used within the R function.

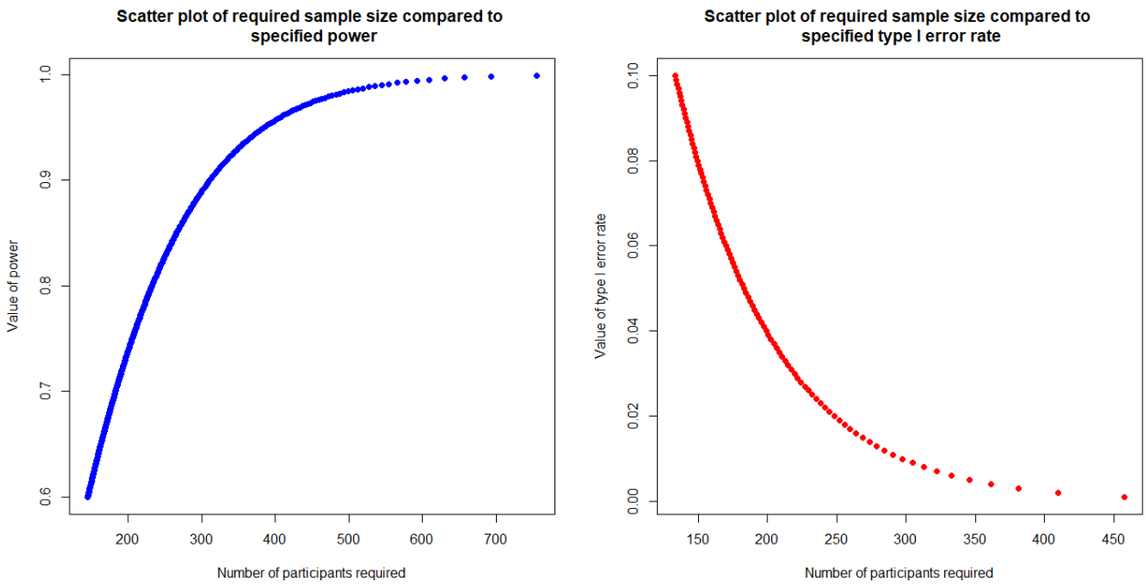

Figure 2: Scatter plot of required sample size compared to the specified power value, $1 - \beta$, (left) and compared to the specified type I error rate (right).

We see from figure 2 that a higher type I error rate leads to a smaller amount of required participants, and as a higher value of power leads to a greater amount of required participants, we also see a lower type II error leading to a higher number of participants required. Further, as we have assumed the age of first kill is a normally distributed variable based on the 'CWsample' dataset, with variance $\sigma \approx 9.62^2$, we can also see how changing this variance will influence the required sample size, as shown in figure 3 below.

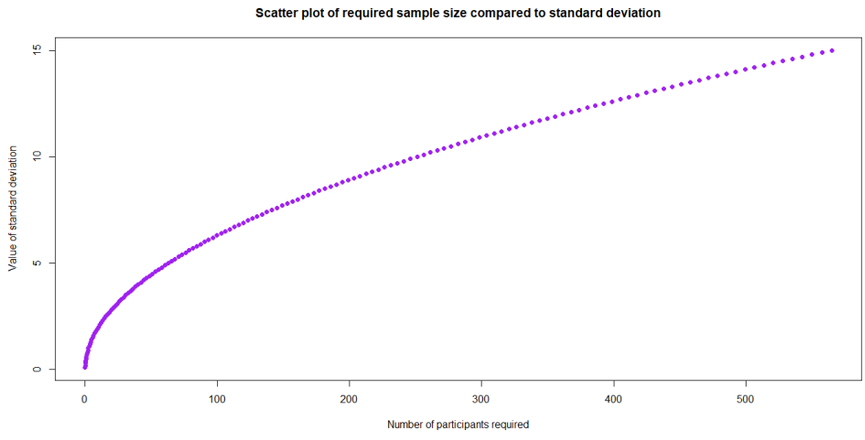

Figure 3: Scatter plot of required sample size compared to the assumed standard deviation.

2.3 Sensitivity analysis

We found the variance of age first kill $\sigma^2 \approx 9.62^2$ and variance of state to be $\sigma_u^2 \approx 3.09^2$ and since the total variance is $\sigma^2 = \sigma_u^2 + \sigma_e^2$, we can combine the two estimated variances to give $\sigma_e^2 \approx 9.62^2 - 3.09^2 \approx 9.11^2$. It is from this model that we can estimate the intracluster correlation coefficient to be

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} \approx \frac{3.09^2}{9.62^2} \approx 0.103. \tag{7}$$

Therefore, when accounting for clusters with this intracluster correlation coefficient of $\rho \approx 0.103$, we have a design effect of $1 - \rho \approx 0.897$ and multiplying this by our proposed sample size of $n = 232.51$, we find the required sample size reduces to $n = 209$ per arm. However, as this is based on the assumption that age at first kill and state are normally distributed variables, the variances are only estimates and so it can be considered risky to reduce the same size this much based only on assumptions. Therefore, we shall stick with the original required sample size of $n = 233$ in each arm, to increase the accuracy of our results.

3 References

MATH5092 Mixed Models with Medical Applications. *Exercises.* Available from: Here

A R Code

```
setwd("C:/Users/fayew/Documents/_Year 4 Semester 2/5092M Mixed Models with Medical Applications
/Practicals")
load(file = "killers2.Rdata")
makesample(8646)
save(CWsample, file = "CWsample.RData")
attach(CWsample)

model0 <- lm(AgeFirstKill ~ 1, data = CWsample)
summary(model0)

CWsample$States <- as.numeric(CWsample$State)
model1 <- lm(States ~ 1, data = CWsample)
summary(model1)

sigmasq <- 9.62**2
sigmasq_u <- 3.09**2
sigmasq_e <- sigmasq - sigmasq_u
sqrt(sigmasq_e)

n <- power.t.test(n = NULL, delta = 2.5, sd = 9.621, power = 0.8)$n

ICC <- sigmasq_u/(sigmasq)
DE <- 1 - ICC

#FIGURE 1

library(dplyr)
library(knitr)
install.packages("kableExtra")
library(kableExtra)
state_tbl <- CWsample %>% group_by(State) %>%
  summarise(AFK = (round(mean(AgeFirstKill), digits=2)),
            .groups = 'drop')
kable(t(state_tbl), caption = "Table of average age of first kill of killer's in each state") %>%
  kable_styling(full_width = F) %>%
  row_spec(1, bold = TRUE)

#FIGURE 2

calc_p <- function(delta,sd,power)
{
  (sqrt(2*sd**2)*(-qnorm(1 - 0.05/2, mean = 0, sd = 1) - qnorm(power/1000, mean = 0, sd = 1))/
  delta)**2
}

X <- calc_p(delta=2.5, sd=9.62, power=600:999)
power=600:999/1000
par(mfrow = c(1, 2))
plot(X, power, xlab="Number of participants required", ylab="Value of power", main="Scatter plot
of required sample size compared to \n specified power", pch=16, col="blue")

calc_a <- function(delta,sd,alpha)
{
  (sqrt(2*sd**2)*(-qnorm(1 - alpha/1000, mean = 0, sd = 1) - qnorm(0.8, mean = 0, sd = 1))/
  delta)**2
}

Y <- calc_a(delta=2.5, sd=9.62, alpha=1:100)
alpha=1:100/1000
plot(Y, alpha, xlab="Number of participants required", ylab="Value of type I error rate",
main="Scatter plot of required sample size compared to \n specified type I error rate", pch=16,
col="red")

#FIGURE 3

par(mfrow = c(1, 1))
calc_sd <- function(delta,sd,alpha)
{
  (sqrt(2*sd**2)*(-qnorm(1 - alpha, mean = 0, sd = 1) - qnorm(0.8, mean = 0, sd = 1))/
  delta)**2
```

```
}

W <- calc_sd(delta=2.5, sd=1:150/10, alpha=0.025)
sd=1:150/10
plot(W, sd, xlab="Number of participants required", ylab="Value of standard deviation", main=
"Scatter plot of required sample size compared to standard deviation", pch=16, col="purple")
```

UNIVERSITY OF LEEDS

School of Mathematics

Declaration of Academic Integrity
for Individual Pieces of Work

I am aware that the University defines plagiarism as presenting someone else's work as your own. Work means any intellectual output, and typically includes text, data, images, sound or performance.

I promise that in the attached submission I have not presented anyone else's work as my own and I have not colluded with others in the preparation of this work. Where I have taken advantage of the work of others, I have given full acknowledgement. I have read and understood the University's published rules on plagiarism and also any more detailed rules specified at School or module level. I know that if I commit plagiarism I can be expelled from the University and that it is my responsibility to be aware of the University's regulations on plagiarism and their importance.

I re-confirm my consent to the University copying and distributing any or all of my work in any form and using third parties (who may be based outside the EU/EEA) to monitor breaches of regulations, to verify whether my work contains plagiarised material, and for quality assurance purposes.

I confirm that I have declared all mitigating circumstances that may be relevant to the assessment of this piece of work and that I wish to have taken into account. I am aware of the School's policy on mitigation and procedures for the submission of statements and evidence of mitigation. I am aware of the penalties imposed for the late submission of coursework.

Student Signature  Date 23/04/2023

Student Name Faye Williams Student Number 201308646

Please note.
When you become a registered student of the University at first and any subsequent registration you sign the following authorisation and declaration:
"I confirm that the information I have given on this form is correct. I agree to observe the provisions of the University's Charter, Statutes, Ordinances, Regulations and Codes of Practice for the time being in force. I know that it is my responsibility to be aware of their contents and that I can read them on the University web site. I acknowledge my obligation under the Payment of Fees Section in the Handbook to pay all charges to the University on demand.
I agree to the University processing my personal data (including sensitive data) in accordance with its Code of Practice on Data Protection http://www.leeds.ac.uk/dpa . I consent to the University making available to third parties (who may be based outside the European Economic Area) any of my work in any form for standards and monitoring purposes including verifying the absence of plagiarised material. I agree that third parties may retain copies of my work for these purposes on the understanding that the third party will not disclose my identity.'"