Multiple Levels Of The Criminal Mind: Modeling, Profiling & Predicting Serial Killers

Name: Marcus Sinclair

Student ID: 201219978

Module Code: MATH5872M

Module Title: Dissertation in Data Science and Analytics

Supervisor: Dr Benjamin Thorpe

Date: ****

Contents

List of Figures

List of Tables

1 Objectives

The primary aim of this project is not just to carry out a data analysis, but rather to demonstrate various multilevel models and how they might be applied in criminology. We will use a database of over 1902 serial killers with 124 variables per killer. The project should start by introducing more familiar statistical models (e.g. linear regression or GLMs) and then should carefully show how they can be extended to multilevel models. There are many directions in which this project could be taken. For example, logistic models could be used to analyse binary data such as whether each killer used a weapon. Poisson models could be used to analyse the number of victims or the number of convictions received. An ambitious student could even explore the use of multiple membership models to analyse killers who were active in multiple states.

2 Generalized Linear Model (GLM)

We start our discussion on multilevel modeling by first considering the **g**eneralized **l**inear **m**odel (**GLM**), a more common and well-known modeling tool in the statistical community. Reasoning for this stems from the GLM being regarded as a specific constrained variant of multilevel modeling [10]. A generalized linear model is any model where the expectation of the output Y is a function of some linear combination of the inputs $\beta X = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$, where $\beta_0, \ldots, \beta_p \in \mathbb{R}$ are input parameter constants [3]. That is:

$$\mathbb{E}(Y = y | X_1 = x_1, \ldots, X_p = x_p) = f(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p), \text{where } f() \text{ is referred to as the link function.} \quad (1)$$

For example, in linear regression, $f()$ is defined as the probability density function of the normal distribution. This results in the model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + e_i, \quad (2)$$

where $e_i \sim N(0, \sigma^2)$ are refereed to as the random error or residuals of the model and are assumed to be **i**ndependent and **i**dentically **d**istributed (**i.i.d**) copies $(e_i)_{i \in 1,\ldots,n}$.

2.1 Application

In this motivating example we focus on illustrating the limitations of generalized linear models applied to serial killers and how even the most simple multilevel model, the variance components model, introduced in section **(3)**, may be more applicable and have a greater potential to yield fruitful results from the data [4].

Consider the following questions:

1. How does the age of a serial killer vary when they commit their first murder?
2. Does the distribution of these lengths depend on the gender of the serial killer? For example, is a killer's age at first kill, on average, higher for females rather than male killers?

The age a serial killer commits kills their first victim is given to the nearest year by **AgeFirstKill** , a numerical variable with known records for 93% of the serial killers recorded in the Radford/FGCU database, 1763 out of the 1902 individuals.

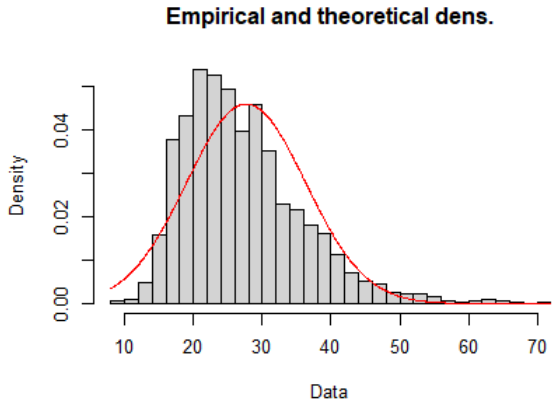
With the focus on our first question regarding variability between **AgeFirstKill** lengths, let us assume the duration periods are independent and identically distributed, with normal distribution. Thus we propose the model:

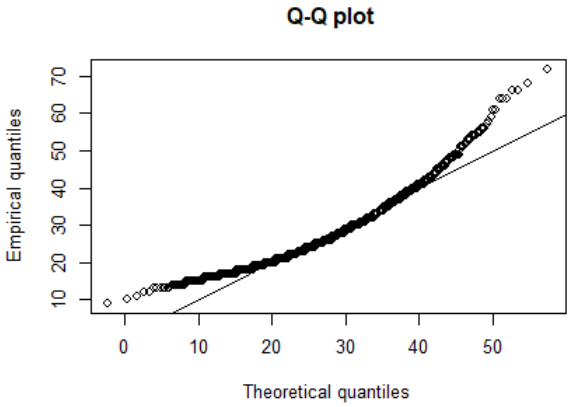
$$y_i \sim N(\beta_0, \sigma^2), \text{with i.i.d copies } (y_i)_{i \in 1,\ldots,n}, \quad (3)$$

where y_i denotes the age at first kill of the i-th serial killer, for individuals $i \in 1, \ldots, n = 1763$. Equivalently, we can write this model as

$$y_i = \beta_0 + e_i, \text{with } e_i \sim N(0, \sigma^2), \text{ i.i.d copies } (e_i)_{i \in 1,\ldots,n}, \quad (4)$$

where each e_i denotes the residual of the i-th serial killer. It is useful to contextualize equation **(4)** as the generalized linear model with normal link function **(2)** such that $x_{1i} = \ldots = x_{pi} = 0$, i.e., no inputs are present. The purpose of writing the model in this format will become apparent following generalized linear model extensions to multilevel models in the following section. Before we answer our questions by applying model **(4)**, we must first check the model assumptions are met that is, normality is present in our residuals e_i. Since the residuals are directly proportional to the output **AgeFirstKill**, looking at this variable seems like the natural approach here.





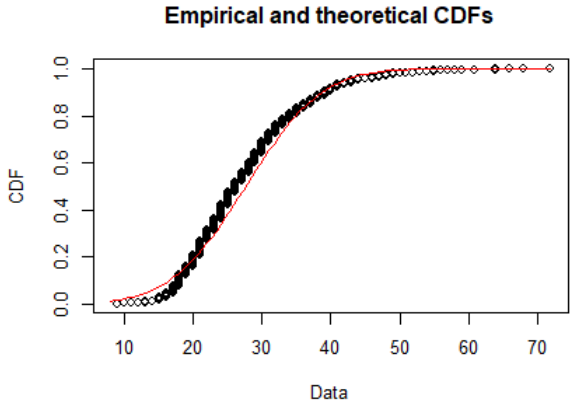


Figure 1: Age Of First Kill Distribution
Histogram, quantile-quantile (Q-Q) plot and cumulative distribution plot for **AgeFirstKill**. The plots indicates the variable's underlying distribution could, to some extent, reasonably be assumed to be normal as evident via the shape of the density and CDF together with the linearity present in the Q-Q plot [1]. Some positive skewness is present- serial killers tend to commit their first murder at early ages, with the most common age being the mid-twenties - we must take this skewness into account when interpreting any results.

	AgeFirstKill		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	27.64	27.24-28.05	<0.001
standard error	8.67		
Observations	1763		

Table 1: GLM Estimates For Model **(4)**, AgeFirstKill ~ 1

The model estimates our unknown parameter β_0 by the method of ordinary least squares (finding the β_0 that minimises $\sum_{i=1}^{n} e_i^2$). Given individuals $n = 1763$, this produces estimates of the population mean, $\hat{\beta}_0 = 27.64$ and residual standard error $\hat{\sigma}^2 = 8.67^2$. From table **(1)**, we see a 95% confidence interval (*CI*) and p-value (*p*) for the intercept parameter β_0. These are defined in the usual sense for instance, the confidence interval is constructed by considering the **estimate** $\pm$ **standard error** $\times t_{2.5\%,n-1}$, where $t_{2.5\%,n-1}$ is the 97.5% quantile of the t-distribution with $n-1$ degrees of freedom [13].

Note, the degrees of freedom in our interval can be written in a more general sense as $n-p-1$, with p referring to the number of inputs present as referenced prior in equation **(2)**. The p-value measures the probability of obtaining a regression coefficient result at least as extreme as the observed result produced by the GLM- it is typically agreed in the statistical community that a p-value < 0.05 is deemed to be useful indicator showing the input as a predictor of the output [11]. The age a serial killers commits their first murder is, on average, seen to occur in their mid-twenties with two-thirds of all killers starting their killings between twenty and forty years of age, with the remaining tailing off at at the younger and older years.

Regarding the second question on the possible dependence of the age at first kill and the gender of a serial killer, we would usually estimate a linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + e_i, \text{with } e_i \sim N(0, \sigma^2), \text{ i.i.d copies } (e_i)_{i \in 1,\ldots,n}, \tag{5}$$

where x_i denotes the gender of the i-th serial killer. This input is given by the variable **Sex**, a categorical variable that takes binary values (*Male*, *Female*) ~ (1745,150) individuals. Only 7 individuals in the Radford/FGCU database have an unknown gender. It may seem strange to apply a categorical variable to a linear regression problem however, given the binary nature of the output of **Sex** and the fact that **Sex** is the only feature input present in the regression model, application of such a model will lead to interpretable and useful results [12]. The variable is encoded with numerical values (*Male*, *Female*) ~ (0,1) respectively, with parameter estimates calculated via minimizing the ordinary least squares of the residuals as usual. Again, standard practice dictates to check the model assumptions of **(5)** prior to analysis that is, normality is present within the residuals e_i.

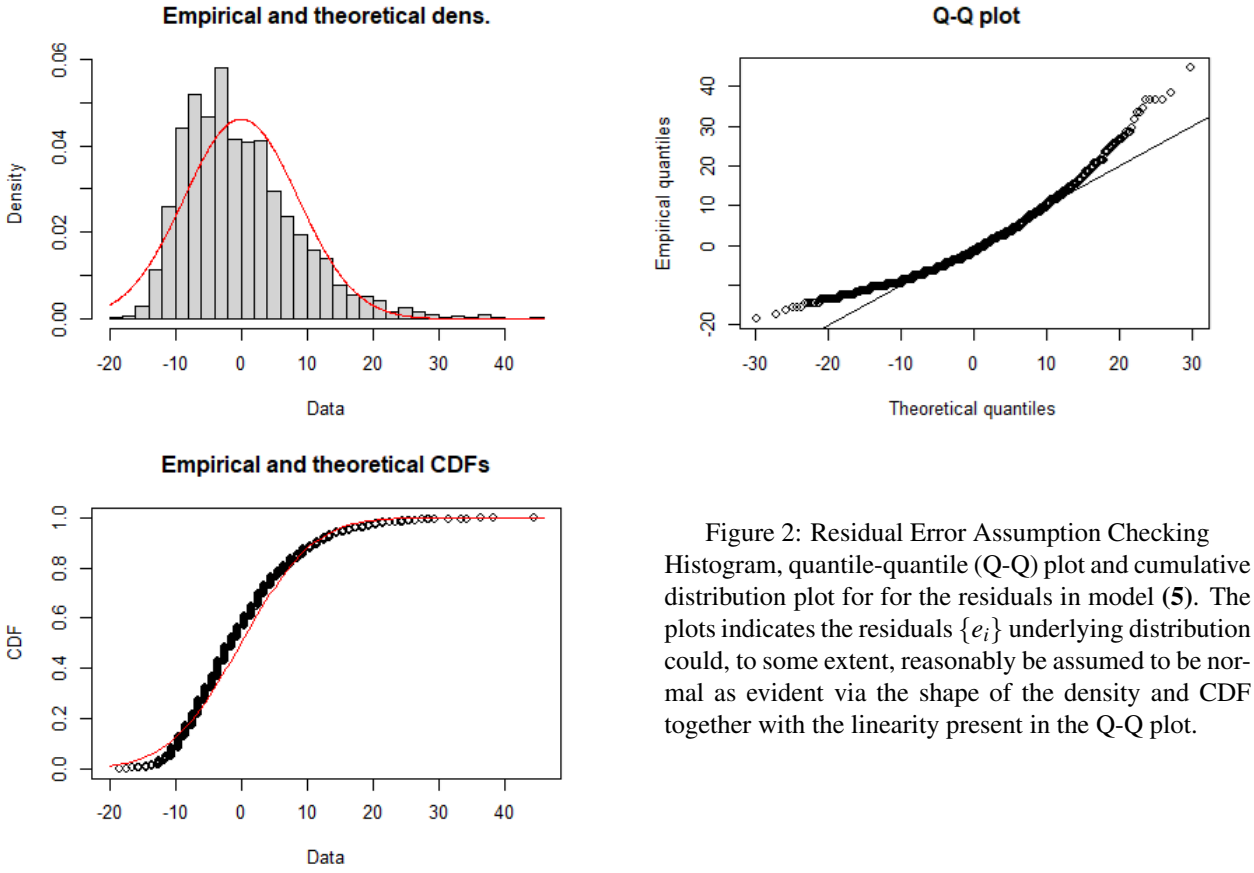


Figure 2: Residual Error Assumption Checking
Histogram, quantile-quantile (Q-Q) plot and cumulative distribution plot for for the residuals in model **(5)**. The plots indicates the residuals $\{e_i\}$ underlying distribution could, to some extent, reasonably be assumed to be normal as evident via the shape of the density and CDF together with the linearity present in the Q-Q plot.

In regards to the distribution of the residuals $\{e_i\}$, it is worth noting that, similar to the distribution of **AgeFirstKill**, a substantial amount of positive skewness is present within the residuals as evident via the positive curvature seen at both extremes of the theoretical quantiles in the Q-Q plot and shape of the histogram in figure **(2)** - our model results should again be interpreted with care.

	AgeFirstKill		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	27.48	27.05-27.90	<0.001
Sex[Female]	2.12	0.62-3.62	0.006
Standard error	8.65		
Observations	1763		
R^2	0.004		
F-statistic	7.71		

Table 2: GLM Estimates For Model **(5)**, AgeFirstKill ~ Sex

Again, estimating via ordinary least squares results in parameter estimates $\hat{\beta}_0 = 27.48$ and $\hat{\beta}_1 = 2.12$, with an estimate of the residual standard error, $\hat{\sigma}^2 = 8.65^2$. Given the presence of a feature input **Sex** in the model, we now have some new statistical objects seen in table **(2)** to consider that is, the R-squared and F-statistic values. An R-squared value helps determine how well the regression models explains the observed data [12]. For instance, here we find only 0.4% of the variability in the age of first kill of a serial killer is explained by the gender of the killer. Female killers, on average, tend to start two years later than their male counterparts.

The large F-statistic of 7.71, coupled with the small p-value $p = 0.006 \ll 0.05$ and confidence interval $CI = [0.62, 3.62]$ situating in the positive domain, suggests that this result is significant throughout serial killers. The distribution of the lengths of age at first kill does indeed depend on the gender of the killer albeit a small amount. A killer's age at first kill is, on average, higher for females rather than male killers.

2.2 Limitations

Application of this generalized linear model approach seems reasonable thus far however, there are at least three issues regarding this methodology:

- We have **assumed independence**. Our sample of $n = 1763$ serial killers may be correlated in relation to some feature variable present in the data. For instance, do you think that the average age at first kill of a serial killer remains constant throughout say, different serial killer motives? Its seems natural to hypothesize motives such as *Financial/personal gain* or *Black widow* (the killing of spouses) to occur at a later age in comparison to motives such as *Organised crime* or *Cult*-related serial killings. If such correlation between individuals is present then a GLM assumption is violated resulting is unreliable findings.
- We probably haven't answered the questions as they were intended. For example, the wording "vary between serial killers" is not very clear. Do we want to know about the killer-to-killer variability in age at first kill? Or do we want to know about the variability in age at first kill with respect to some feature input such as **Race**, **Motive** or **KilledWithAccomplice**? Or do we want to know about both?
- Are our normality assumptions justified? We have applied two generalized linear models, both of which assumed normality within the residuals. As seen by our assumption investigations in figures **(1)** and **(2)**, these normalities assumed could be questioned. Both distributions, that of **AgeFirstKill** and the residuals $\{e_i\}$ in model **(5)** have a substantial amount of positive skewness. Real-world data is never "perfect" and some liberties must be taken in regards to fitting to model assumptions. It is up to the statistician to determine if the amount of such liberties taken is justified given the context of the hypothesis in question [4].

Let's investigate the independence assumption for our first model **(4)** for which we estimated the average age at first kill to be $\hat{\beta}_0 = 27.64$ years.

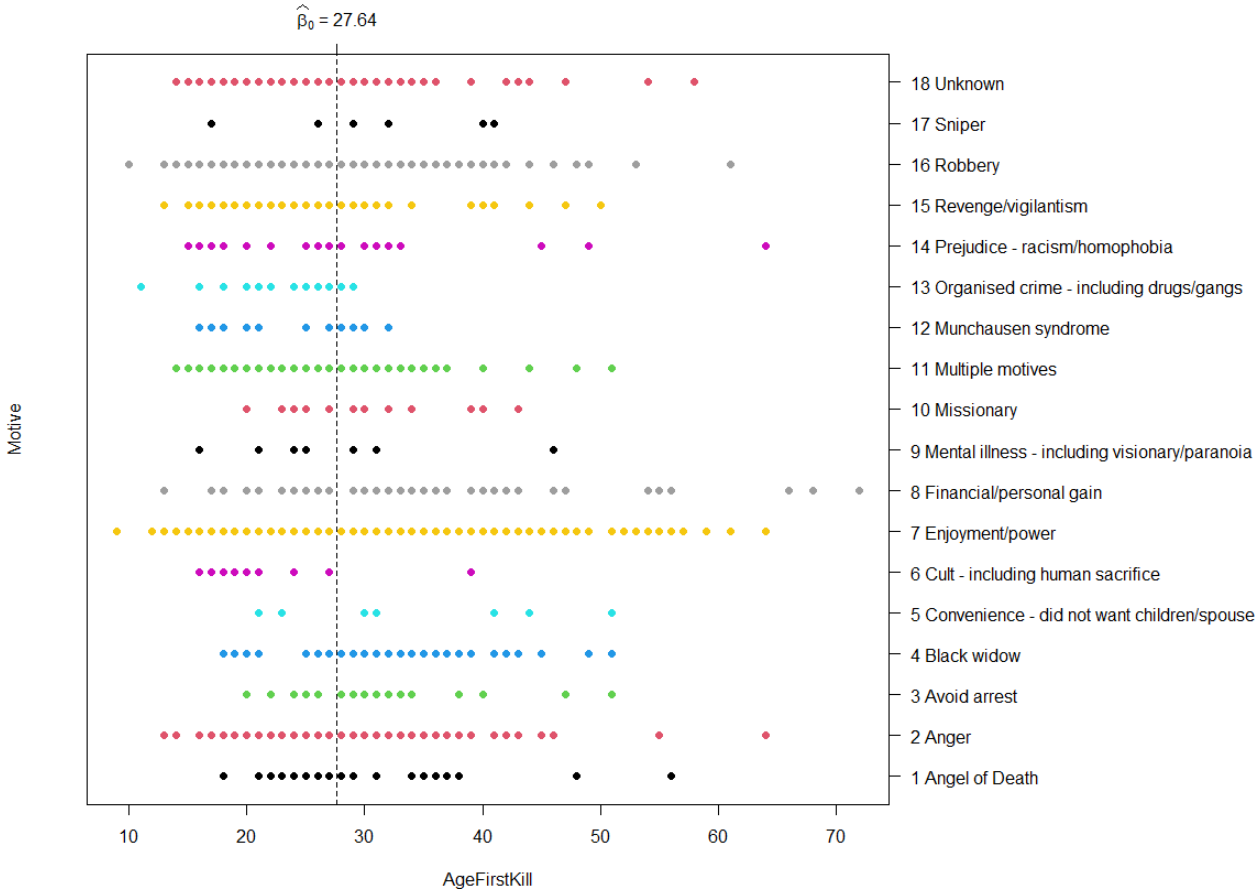


Figure 3: Age At First Kill Clustered By Motive

Plot of **AgeFirstKill** clustered by **Motive**, a categorical variable taking 18 unique values as seen in the plot, from *Angel of Death* to *Unknown* motives. The estimated age at first kill across all killers is given by the dashed vertical line, with a value of $\hat{\beta}_0 = 27.64$ years. It appears that the average duration varies between killers with different motives.

As seen in figure **(3)**, the age at first kill may be correlated in relation to the motive of the killer. As hypothesised prior, serial killers relating to *Organised crime* seem to start killings at an earlier age than average. Vice versa, *Financial/personal gain* incentives and the *Black widow* killings occur at later ages than average. These natural clusterings seen are ignored in a standard generalized linear model framework. As such, our independence assumptions of the response variable $\{y_i\}$ made in the generalized linear model case could be argued to be invalidated [2].

We now introduce the main focus of our discussion that is, multilevel models- a tool that can extend the generalized linear model framework to internalise such groupings present in the data.

3 Variance Components Model (VCM)

Multilevel modeling is a strategy that internalises the clustered nature of our data within the model. The idea is to produce a framework with multiple levels where level 1 refers to individuals or observations as seen in the standard generalized linear model, and level 2 refers to the cluster-level whereby certain individuals behave similarly within respective clusters [2]. This framework could be built upon in perpetuity for instance, a level 3 super-cluster could be internalised and so-forth. For simplicity, let us first consider the level 2 model with the absence of inputs/covariates X:

$$y_{ij} = \beta_0 + u_{0j} + e_{0ij}, \tag{6}$$

with $u_{0j} \sim N(0, \sigma^2_{u0})$ i.i.d copies $(u_{0j})_{j\in 1,\dots,m}$ and $e_{0ij} \sim N(0, \sigma^2_{e0})$ i.i.d copies $(e_{0ij})_{i\in 1,\dots,n, j\in 1,\dots,m}$. Here, u_{0j} is a level 2 (cluster-level) error formally refereed to as the random effect of the j-th group on output/response variable y_{ij}. The value of u_{0j} is assumed to be the same for all individuals pertaining to the j-th cluster, and is assumed to be independent to level 1 random errors e_{0ij}. The zero subscript used in level 1 and 2 errors e_{0ij} and u_{0j} may seem somewhat cumbersome however, its inclusion allows model extension to appear intuitive and complete. A model of the form given in equation **(6)** is refereed to as a **v**ariance **c**omponents **m**odel (**VCM**).

3.1 Application

Recall the first question given in section **(2.1)** :

1. How does the age of a serial killer vary when they commit their first murder?

Given this hypothesis, our goal here is to determine whether VCM may be more applicable and have a greater potential to yield fruitful results from the data compared to the GLM application. Let y_{ij} denote the age at first kill corresponding to the i-th serial killer (the i-th level 1 individual) of the j-th cluster (level 2). Define level 2 clusters as the motives seen and labeled in figure **(3)** i.e., $j = 1$ refers to the motive *Angel of Death*, $j = 2$, *Anger* and so forth up to $j = 18$, *Unknown* motive. Before applying the VCM we must first check our model assumptions that is, normality of not just the level 1 standard error residuals $\{e_{0ij}\}$, but we must find normality present within the level 2 random error $\{u_{0j}\}$ too.

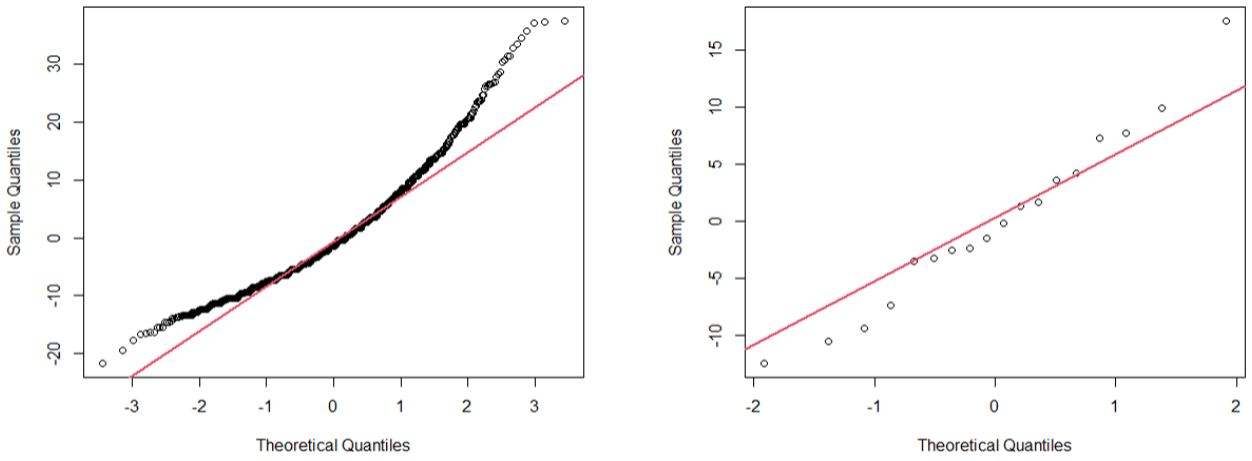


Figure 4: VCM Quantile-Quantile Plots Of Level 1 & 2 Residual Error
Left: Level 1 standard error residuals $\{e_{0ij}\}$ quantile-quantile plot. Similar issues (as expected) arise for assumption checking for model **(6)** to that seen in application **(2.1)**, that of positive skewness present at the level 1 state. Right: Level 2 random error $\{u_{0j}\}$ quantile-quantile plot. 18 points are seen which are reference to the 18 unique values the variable **Motive** can take. Linearity at the level 2 state is seen throughout the points - normality model assumption is justified.

	AgeFirstKill		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	27.87	26.19-29.56	<0.001
Random Effects			
σ^2_{e0}	8.41^2		
$\sigma^2_{u0\text{Motive}}$	3.15^2		
ICC	0.13		
N_{Motive}	18		
Observations	1763		

Table 3: VCM Estimates For Model **(6)**, AgeFirstKill ~ 1 + (1|Motive)

As shown in table **(3)**, applying VCM yields the estimate $\hat{\beta}_0 = 27.87$ years for the population mean $E(y_{ij}) = \beta_0$. Unlike in the generalized linear model case whereby one variance parameter is estimated, for the variance components model we require to estimate two: $\hat{\sigma}^2_{u0} = 3.15^2$ at the motive cluster level, and $\hat{\sigma}^2_{e0} = 8.41^2$ at the killer individual level. These suggest that while most variability in the age at first kill of a serial killer stems from differences between killers, a substantial amount of variability is explained by the motive behind such killings. An interesting and potentially surprising finding is that the population mean for VCM marginally differs from that of the generalized linear model applied in section **(2)**, $\hat{\beta}_0 = 27.87$ compared to that of 27.64 years seen prior.

Whilst the generalized linear model applied in section **(2)** estimates model parameters via maximum likelihood methods i.e, for linear regression, by the ordinary least squares method (OLS); the variance components model uses a more general framework to estimate model parameters- the general least squares (GLS) method. The details regarding GLS will not be discussed here, but can be found in Goldstein's book on multilevel modeling [4].

Parameters estimated via the iterative general least squares method will converge to equivalent parameter values produced via maximum likelihood. The differences in parameter values $\hat{\beta}_{0\text{VCM}} = 27.87$ compared to that of $\hat{\beta}_{0\text{GLM}} = 27.64$ is due to the difference in parameter estimation methods.

Additional statistics will be examined during our multilevel modeling journey with the first being the **i**ntraclass **c**orrelation **c**oefficient (**ICC**) seen in table **(3)** taking a value of 0.13 for the variance components model applied. This statistic is a measure of the correlation between two different individuals in the same cluster, $corr(y_{ij,}y_{kj})$ for $i \neq k$ [2]. It describes how strongly individuals in the same cluster resemble each other, taking a value from 0 to 1 with larger values indicating a high resemblance in the cluster. For a variance components model the intraclass correlation coefficient can be calculated by

$$ICC = \frac{\sigma^2_{u0}}{\sigma^2_{u0} + \sigma^2_{e0}} = \frac{3.15^2}{3.15^2 + 8.41^2} = 0.13, \text{ for the example above.} \tag{7}$$

The formula makes intuitive sense for VCM as the intraclass correlation coefficient is just the proportion of level 2 variation seen in regards to the total variance at levels 1 and 2. Whilst interpreting parameter estimates and statistics is important and useful for deepening our understanding of serial killers, we should be careful about blindly applying models, inferring their results, but focus on the mechanism behind model construction too as the following figure illustrates.

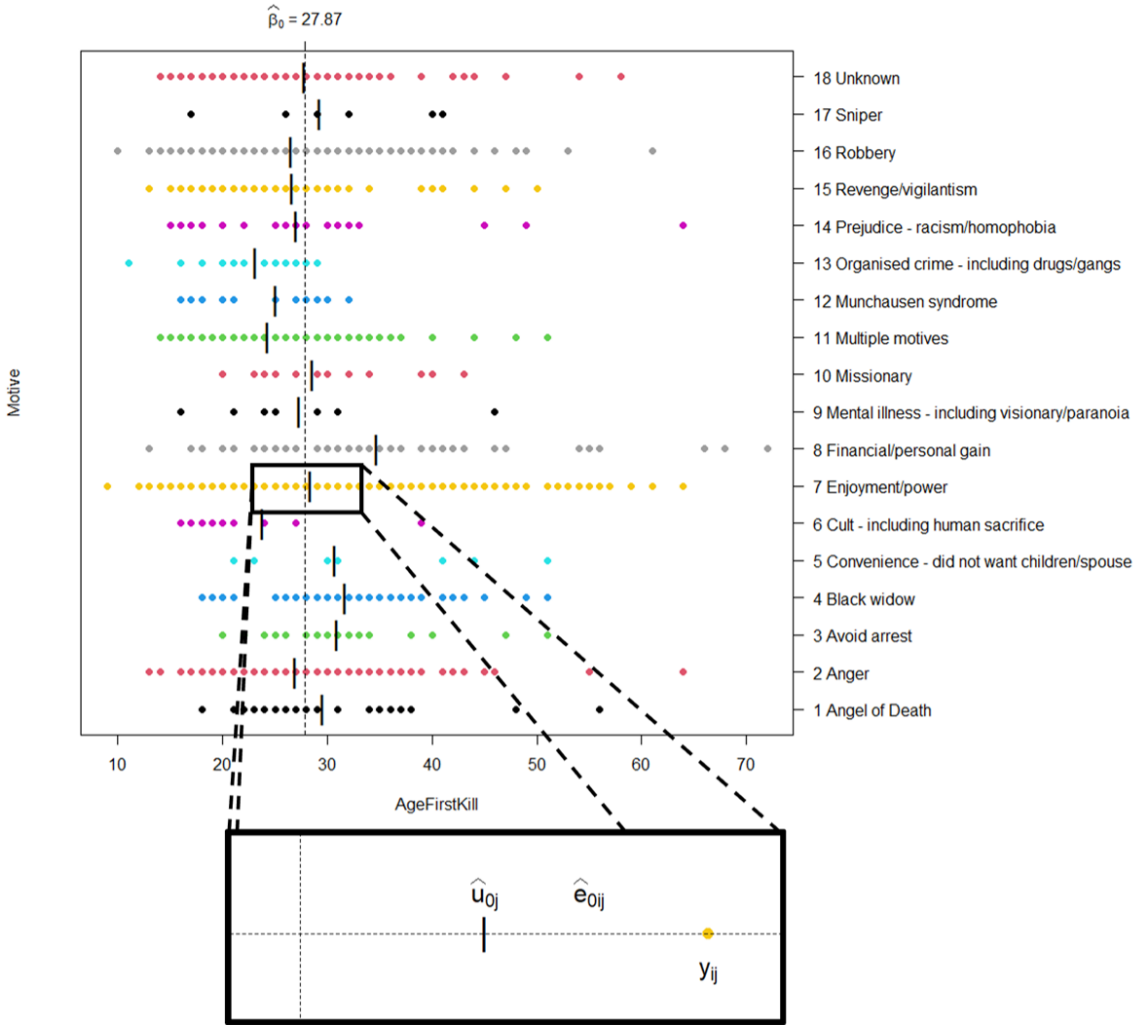


Figure 5: Visualisation Of VCM

Expanded plot to that seen in figure **(3)** with the parameter estimate $\hat{\beta}_0 = 27.87$ produced by VCM seen by the dotted line. Black bars are shown for each of the 18 individual motives that represent the estimated cluster means $\hat{\beta}_0 + \hat{u}_{0j}$. A grid zoom is shown for an individual y_{ij} for the motive *Enjoyment/power* that is, for $j = 7$. The grid zoom shows the mechanism of model **(6)**. Individuals in a specific cluster are normally distributed with noise $e_{0ij} \sim N(0, \sigma^2_{e0})$ around their specific cluster mean - the black bar, with the cluster means being themselves normally distributed with noise $u_{0j} \sim N(0, \sigma^2_{u0})$ around the parameter β_0.

3.2 Limitations

Hopefully this example has given insight into the usefulness of multilevel modeling. Whilst only the most simple multilevel model (the VCM) has been introduced thus far, more complex extensions are still relatively straight forward and even routine to implement using standard statistical software. Therefore, it is of great importance to reflect on whether model assumptions are met and hierarchical structures are prevalent before blindly applying such models. It is common for standard single level models such as the GLM to suffice. The VCM cannot be applied to hypotheses regarding variability between feature inputs and a response variable, considering an alternative multilevel model in such cases will now be explored.

4 Random Intercepts Model (RIM)

Here we introduce an extension to the level 2 VCM seen in section **(3)**. Unlike the variance components model, the **r**andom **i**ntercept **m**odel (RIM) allows the inclusion of inputs/covariates $\beta X = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$ where $\beta_0, \ldots, \beta_p \in \mathbb{R}$ are input parameter constants [4]. A random intercept model assumes multivariate normality of our sample $\{y_{ij}\}$ such that

$$y_{ij} = (\beta_0 + u_{0j}) + \sum_{l=1}^{p} \beta_l x_{lij} + e_{0ij}, \tag{8}$$

where all random terms on the right hand side are mutually independent and normally distributed

$$e_{0ij} \sim N(0, \sigma^2_{e0}) \text{ and } u_{0j} \sim N(0, \sigma^2_{u0}).$$

By multivariate normality, it is sufficient to just think of each random term (and thus the response variable y_{ij} too) in the model i.e., e_{0ij} and u_{0j} as being normally distributed. It is important to understand and distinguish the fixed and random effects in multilevel models. β_l is the constant fixed effect of the l-th covariate $l = 1, ..., p$. u_{0j} is the level 2 random effect of cluster j. e_{0ij} is the level 1 random error for the i-th individual in cluster j. Since the level 2 random effect u_{0j} is constant within a specific cluster j, we can visualise a random intercepts model as m regression models, one for each cluster, with different intercepts $\beta_0 + u_{0j}$. We have $p+3$ unknown, constant parameters to estimate: fixed effects $\beta_0, ..., \beta_p$, and variance parameters $\sigma^2_{e0}$and σ^2_{u0}.

4.1 Application: Imbalanced Covariate

Recall the second question given in section **(2.1)**:

1. Does the distribution of the age at first kill depend on the gender of the serial killer?

Our goal here is to show the usefulness of a random intercepts model and how model results vary in relation to its level 1, generalized linear model counterpart. Applying RIM is a good choice here. Justification for applying the random intercepts model stems from clustering in our data. Normally, we would fit just the single level regression model as applied in section **(2.1)**. This is ill-advised because of the grouping present within the data. Furthermore, the first question is about means, and we can answer it using the fixed part, that is, the slope β_1 of the overall regression line. This is analogous to fitting standard single level regression model where we'd use our estimate of β_1 to answer the question.

It seems reasonable to suggest a variance components model could be used regarding the question given above however, VCM is not applicable here because we need to control for the differences in the feature input **Sex**. RIM does indeed allow control for those differences. We can answer the question using the random part, the level 2 variance, σ^2_{u0}.

We propose the model:

$$y_{ij} = (\beta_0 + u_{0j}) + \beta_1 x_{1ij} + e_{0ij}, \tag{9}$$

where y_{ij} denotes the **AgeFirstKill** of the i-th serial killer in the j-th cluster, **Motive**. This model is equivalent to **(8)** with only one covariate present x_{1ij}, the **Sex** of the i-th serial killer in the j-th cluster.

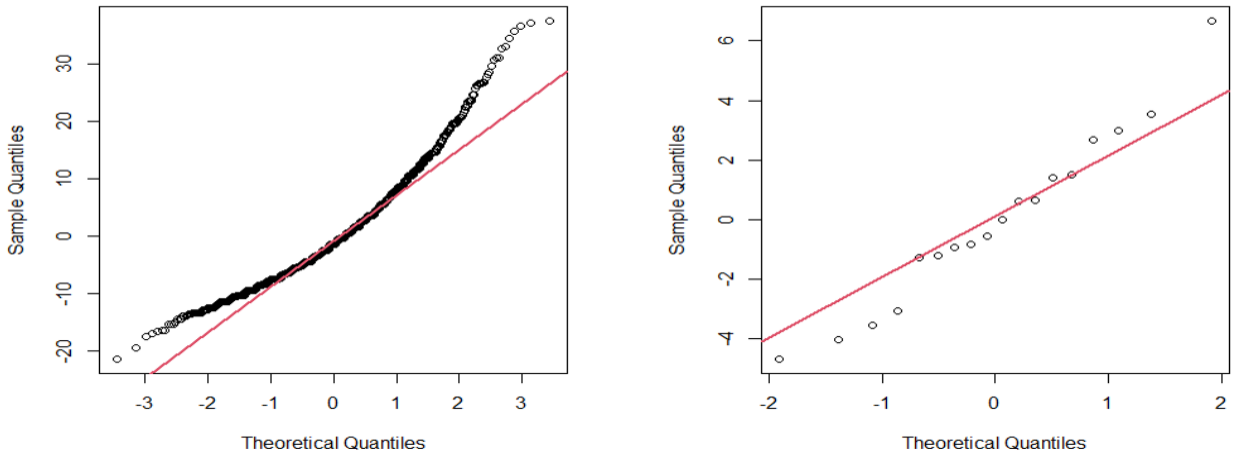


Figure 6: RIM (Sex) Quantile-Quantile Plots Of Level 1 & 2 Residual Error

Left: Level 1 standard error residuals $\{e_{0ij}\}$ quantile-quantile plot. Right: Level 2 random error $\{u_{0j}\}$ quantile-quantile plot. 18 points are seen which are reference to the 18 unique values the variable **Motive** can take. Only a marginal difference is seen in the distribution of RIM residuals in comparison to VCM residuals - I refer you to figure **(4)** for implications.

	AgeFirstKill		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	27.76	26.05–29.46	<0.001
Sex[Female]	0.52	-1.21–2.24	0.558
Random Effects			
σ^2_{e0}	8.41^2		
$\sigma^2_{u0\text{Motive}}$	3.19^2		
ICC	0.13		
N_{Motive}	18		
Observations	1763		

Table 4: RIM Estimates For Model **(9)**, AgeFirstKill ~ Sex + (1|Motive)

From figure **(6)**, it seems justifiable to assume the underlying distribution of this variable to be normal, thus the assumption of multivariate normality of our sample $\{y_{ij}\}$ required by RIM is met. Application of the random intercepts model yields parameter estimates $\hat{\beta}_0 = 27.76$ and $\hat{\beta}_1 = 0.52$ however, unlike in our GLM example in section **(2.1)** whereby we observed a meaningful impact of the gender of a killer on the age at first kill, the analogous RIM version does not. Our confidence interval $CI = [-1.21, 2.24]$ spans a negative and positive domain. Our p-value $p = 0.558$ is large.

Given the similarities in both models, it may seem strange to see such a large disparity in terms of the significance of our findings. What limitation of multilevel modeling have we just stumbled across in this example?

4.2 Limitations

Plot showing model **(9)** for the most prevalent motives *Enjoyment/power* and *Robbery*. Violin plots are seen, two for each motive with the raw data points shown by the black dots. For a given age of first kill, the greater the width of the violin plot the greater the density of serial killers - (750,9) ~ (Male, Female) for *Enjoyment/power* and (367,18) ~ (Male, Female) for *Robbery*.

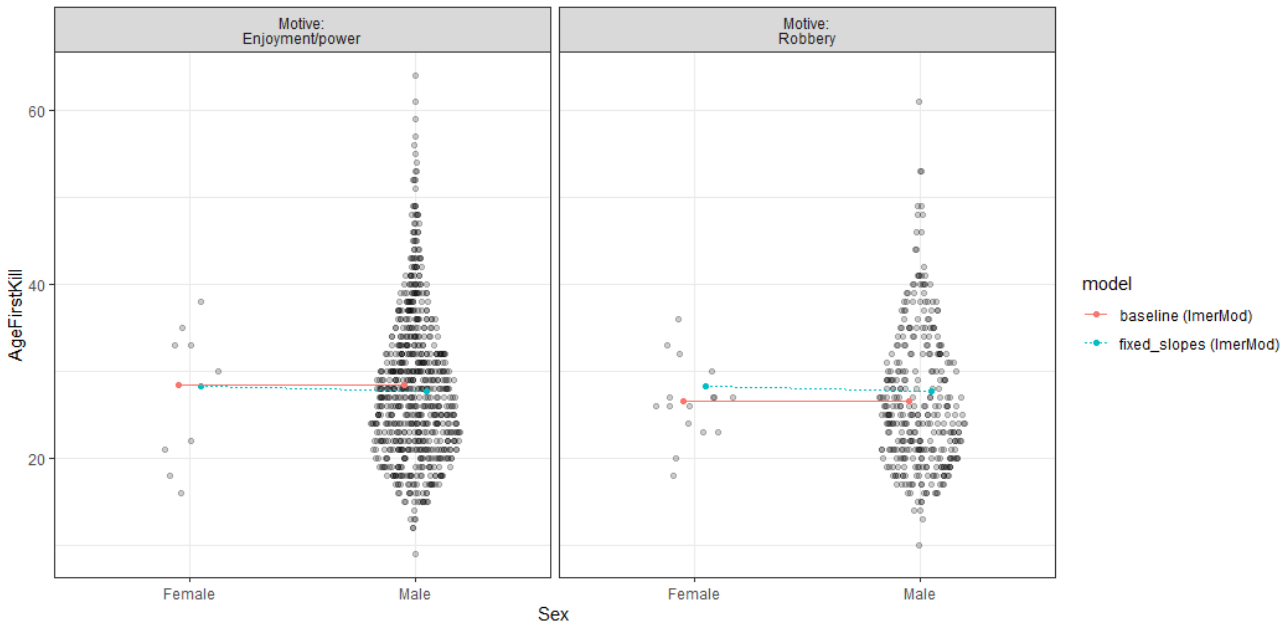


Figure 7: RIM Violin Plot - Unbalanced Covariate

The red line shows a VCM for **AgeFirstKill** and is refereed to as a baseline model. The blue line is the RIM in question as is refereed as a fixed_slopes model. The violin plots in figure **(7)** are an example of the large imbalance within the value of our covariate **Sex**. Throughout the entire data set we find a large dominant presence in male serial killer in comparison to female with (*Male*, *Female*) ~ (1745,150) individuals.

RIM derives its parameters in a similar fashion to that of the VCM of which, we have seen the mechanism behind its construction- see figure **(5)**. Large data imbalances results in large uncertainties in estimations of our level 1 and 2 residuals $\{e_{0ij}\}$ and $\{u_{0ij}\}$. For instance in relation to the RIM model seen in figure **(7)** , consider a new addition to the data for example, a female serial killer whom started their killings at the old age of 60.

This finding would impact our parameter estimates of β_0 and β_1 to a greater extent in comparison to a new male serial killer observation of the equivalent age. Imbalances in data, especially with small samples within clusters present as seen in this example brings uncertainty in our findings. For such an imbalanced feature input, partitioning the data into various clusters risks small sample sizes within specific groups. It seems applying a more balanced feature input of interest for instance, **Race**, as apposed to **Sex** may result in more significant findings.

4.3 Application: Balanced Covariate

Consider the following hypothesis:

1. The distribution of the age at first kill varies within the race of a serial killer. Black ethnic groups tend to start killings at an earlier age compared to white serial killers.

Our goal in this example is to apply equivalent analysis to that seen in section **(4.1)** however, analysis is now applied upon a more balanced feature input **Race** - A categorical variable with values (*White, Black, Hispanic, Asian, Native American*) ~ (1024, 750, 90, 10 ,13). Analogous to the variable **Sex** taking binary values (*Male, Female*), we impute **Race**, removing the minority classes *Hispanic*, *Asian* and *Native American*, treating them as missing values.

Justification stems from the small sample sizes present in the data for such classes, coupled with our goal of reducing down the feature input **Race** to that of a binary variable, similar to the variable **Sex** as analysed prior. From an imbalanced covariate **Sex** with values (*Male*, *Female*) ~ (1745,150), we are ready to consider the more balanced covariate **Race** with values (*White, Black*) ~ (1024, 750). We propose the model:

$$y_{ij} = (\beta_0 + u_{0j}) + \beta_1 x_{1ij} + e_{0ij}, \tag{10}$$

where y_{ij} denotes the **AgeFirstKill** of the i-th serial killer in the j-th cluster, **Motive**. This model is equivalent to **(8)** with only one covariate present x_{1ij}, the **Race** of the i-th serial killer in the j-th cluster. Let us check our model assumptions.

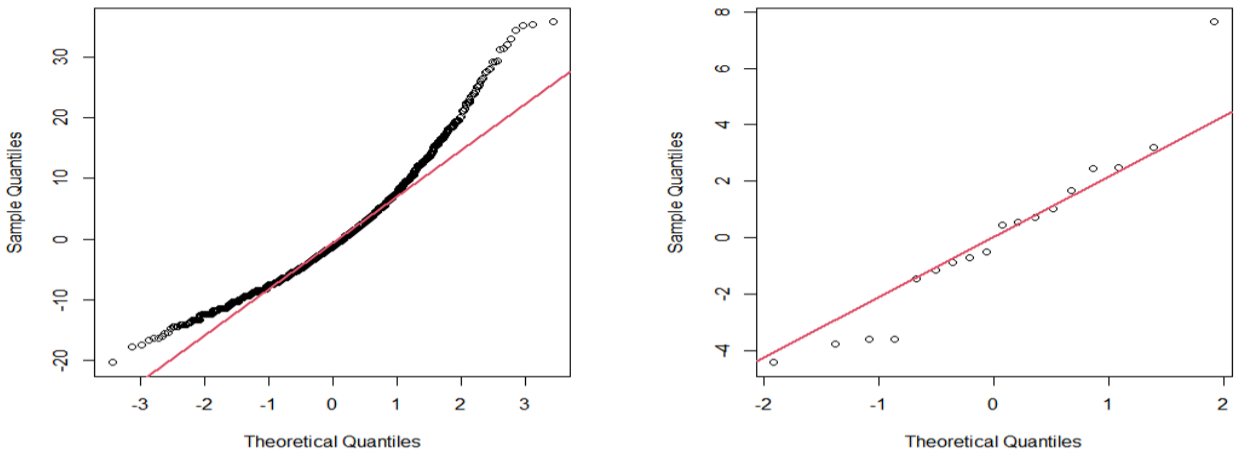


Figure 8: RIM (Race) Quantile-Quantile Plots Of Level 1 & 2 Residual Error

Left: Level 1 standard error residuals $\{e_{0ij}\}$ quantile-quantile plot. Right: Level 2 random error $\{u_{0j}\}$ quantile-quantile plot. 18 points are seen which are reference to the 18 unique values the variable **Motive** can take. Only a marginal difference is seen in the distribution of RIM residuals in comparison to VCM residuals - I refer you to figure **(4)** for implications.

	AgeFirstKill		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	26.39	24.60–28.18	<0.001
Race[White]	2.63	1.77–3.49	<0.001
Random Effects			
σ^2_{e0}	8.38^2		
$\sigma^2_{u0\text{Motive}}$	3.27^2		
ICC	0.13		
N_{Motive}	18		
Observations	1649		

Table 5: RIM Estimates For Model **(10)**, AgeFirstKill ~ Race + (1|Motive)

From figure **(8)**, it seems justifiable to assume the underlying distribution of this variable to be normal, thus the assumption of multivariate normality of our sample $\{y_{ij}\}$ required by RIM is met. Application of the random intercepts model yields parameter estimates $\hat{\beta}_0 = 26.39$ and $\hat{\beta}_1 = 2.63$ however, unlike in our previous RIM example in section **(4.1)** whereby we observed no meaningful impact of the gender of a killer on the age at first kill, we now do indeed see statistical significance present.

Black ethnic groups do indeed tend to start killings at an earlier age compared to white serial killers, with the data suggesting around two and a half years difference between the two groups. Our confidence interval $CI = [1.77, 3.49]$ spans a positive domain. Our p-value $p < 0.001$ is sufficiently small.

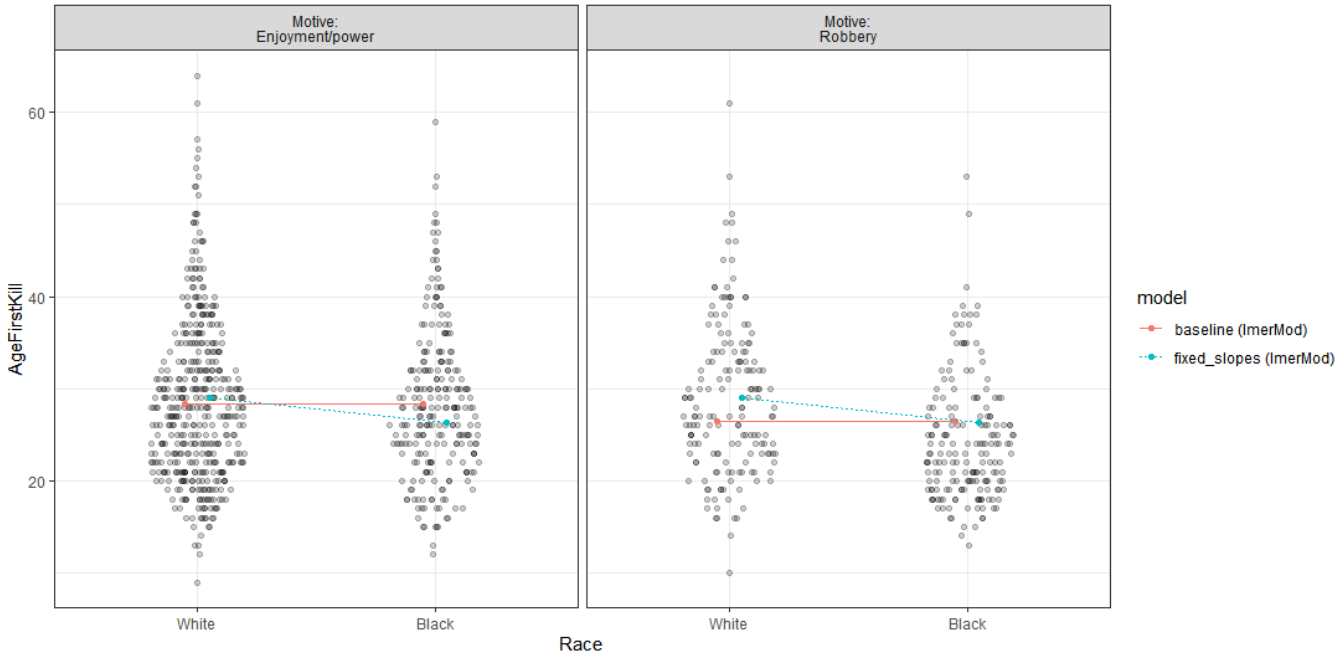


Figure 9: RIM Violin Plot - Balanced Covariate

Plot showing model **(10)** for the most prevalent motives *Enjoyment/power* and *Robbery*, equivalent to plots seen in figure **(7)**, but with the covariate in question being **Race**. As seen in figure **(9)**, a large sample is present between motives and between racial groups - issues of uncertainty now do not arise as discussed in the previous application. A natural continuation from a random intercept model is to consider relaxing the constraint of assuming the relationship between our response variable Y and feature input X for each cluster is equivalent i.e. to allow slopes to vary.

5 Random Slopes Model (RSM)

Here we introduce an extension to the level 2 RIM seen in section **(4)**. Unlike the random intercepts model, the **r**andom **s**lopes **m**odel (RSM) allows the inclusion possibility that the effect of a covariate on the response might also vary between clusters. A random slopes model assumes multivariate normality of our sample $\{y_{ij}\}$ such that

$$y_{ij} = (\beta_0 + u_{0j}) + (\beta_1 + u_{1j})x_{1ij} + \sum_{l=2}^{p} \beta_l x_{lij} + e_{0ij}, \tag{11}$$

where all random terms on the right hand side are mutually independent and normally distributed,

$$e_{0ij} \sim N(0, \sigma^2_{e0}), u_{0j} \sim N(0, \sigma^2_{u0}) \text{ and } u_{1j} \sim N(0, \sigma^2_{u1}).$$

Whilst these random terms are assumed to be mutually independent, an exception is made for a possible covariance between the random effects, $cov(u_{1j}, u_{0j}) = \sigma_{u01}$. We can think of u_{1j} as the level 2 random effect of covariate x_1 for group j. For the covariance between the random effects, a positive value indicates that clusters with high intercept residuals u_{0j} tend to have high slope residuals u_{1j}. Not only do we need to consider the sign of the covariance term, but the signs of the intercept β_0 and slope β_1 must be examined too.

For example, if both the slope and intercept are positive, a positive σ_{u01} suggests that clusters with large intercepts i.e., large $\beta_0 + u_{0j}$, on average, have steeper slopes- high $\beta_1 + u_{1j}$. The complement is true too- groups with low intercepts have flatter slopes than the average. This will result to a fanning out of the cluster respective regression lines when plotted together. The same reasoning can be made for negative values of σ_{u01}, with the outcome of regression line fanning inwards. We have $p+4$ unknown, constant parameters to estimate: fixed effects $\beta_0, ..., \beta_p$, and variance parameters σ^2_{e0}, σ^2_{u0} and σ^2_{u1} .

5.1 Application

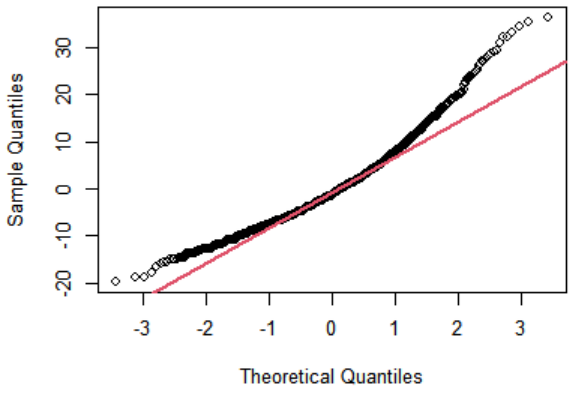
Recall the following hypotheses that were proposed in section **(4)**:

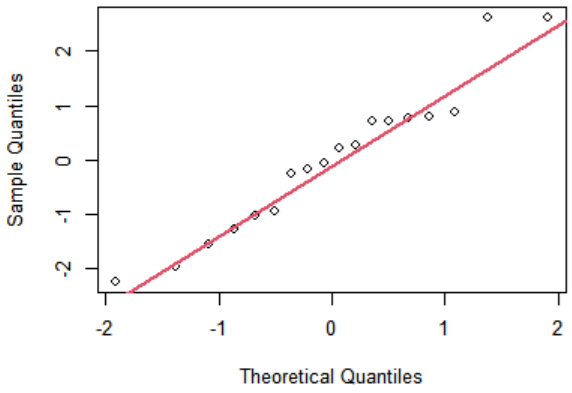
1. The distribution of the age at first kill varies within the race of a serial killer. Black ethnic groups tend to start killings at an earlier age compared to white serial killers.

In the previous section, we allowed for motive effects on the mean age a serial commits their first murder by allowing the intercept of the regression of **AgeFirstKill** on **Race** to vary across the motives. We assumed, however that changes in **AgeFirstKill** seen by **Race** are the same for all motives, i.e., the slope of the regression line were assumed fixed across motives. Using the random slopes model, we can consider the equivalent hypothesis proposed in the previous section, but instead of fixing the changes in **AgeFirstKill** seen by **Race** for all motives, we allow the intercept and slope to vary randomly across motives. Consider the model:

$$y_{ij} = (\beta_0 + u_{0j}) + (\beta_1 + u_{1j})x_{1ij} + e_{0ij}, \tag{12}$$

where y_{ij} denotes the **AgeFirstKill** of the i-th serial killer in the j-th cluster, **Motive**. This model is equivalent to **(11)** with only one covariate present x_{1ij}, the **Race** of the i-th serial killer in the j-th cluster.





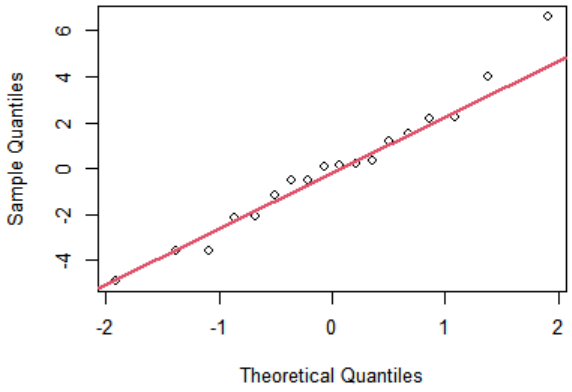


Figure 10: RSM Quantile-Quantile Plots Of Level 1 & 2 Residual Error
Top Left: Level 1 standard error residuals $\{e_{0ij}\}$ quantile-quantile plot. Top Right: Level 2 random individual error $\{u_{0j}\}$ quantile-quantile plot. 18 points are seen which are reference to the 18 unique values the variable **Motive** can take. Bottom Left: Level 2 random cluster error $\{u_{1j}\}$ quantile-quantile plot. 18 points are again seen with the same reasoning as the random individual error case. Only a marginal difference is seen in the distribution of RSM residuals in comparison to VCM residuals - I again refer you to figure **(4)** for implications.

	AgeFirstKill		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	25.39	24.03-26.75	<0.001
Race[White]	4.29	2.17-6.41	<0.001
Random Effects			
σ^2_{e0}	8.29^2		
$\sigma^2_{u0\text{Motive}}$	1.88^2		
$\sigma^2_{u1\text{Motive.RaceWhite}}$	3.39^2		
$\rho_{01\text{Motive}}$	0.15		
$ICC_{Intercept}$	0.04		
N_{Motive}	18		
Observations	1649		

Table 6: RSM Estimates For Model **(12)**, AgeFirstKill ~ Race + (1+Race|Motive)
From figure **(10)**, it seems justifiable to assume the underlying distribution of the residuals to be normal. Application of the random slopes model yields parameter estimates $\hat{\beta}_0 = 25.39$ and $\hat{\beta}_1 = 4.29$. There is significant evidence to suggest black ethnic groups do tend to start killings at an earlier age compared to white serial killers as hypothesised- positive domained $CI = [2.17, 6.41]$ and small p-value $p < 0.001$ present in our findings. Whilst fixing the effect of **Race** across motives gave the parameter $\hat{\beta}_1 = 2.63$ years in RIM, allowing variation within motives in this random slopes model has increased the impact of our feature input substantially to $\hat{\beta}_1 = 4.29$ years.

From table **(6)** we find both the slope, intercept and covariance $\rho_{01\text{Motive}} = 0.15$ take positive values. This suggests that clusters with large intercepts i.e., large $\beta_0 + u_{0j}$, on average, have steeper slopes- high $\beta_1 + u_{1j}$. Serial killers whom start killings at an older age on average, tend to be white. Intraclass correlation coefficients cannot be applied to random slope models due to the increase in complexity given by the level 2 random effect u_{1j} [6] however, it is useful

to define the statistic

$$ICC_{\text{Intercept}} = \frac{\sigma^2_{u0}}{\sigma^2_{u0} + \sigma^2_{u1} + \sigma^2_{e0}} = \frac{1.88^2}{1.88^2 + 3.39^2 + 8.29^2} = 0.04, \text{ as given in the table above.} \tag{13}$$

This is a measure of the variability seen at the intercept in relation to the total level 1 and 2 variance $\sigma^2_{u0} + \sigma^2_{u1} + \sigma^2_{e0}$. Coupled with a weak positive correlation is seen in regards to the intercept and slope of RSM with a correlation coefficient of $\rho_{01\text{Motive}} = 0.15$ present between parameters, it seems only a small amount of variability seen within serial killers age at first kill and race are explained at the motive level.

5.2 Limitations

Relaxing the constraint of assuming the relationship between our response variable Y and feature input X for each group has added an extra layer of complexity in the random slopes model in comparison to the random intercepts model. Is this complexity justified? From the example prior regarding how the distribution of the age at first kill varies within the race of a serial killer, we have seen a substantial increase in variation between race and age at first kill from RIM with race parameter $\hat{\beta}_1 = 2.63$, to the more complex RSM taking value $\hat{\beta}_1 = 4.29$ years. However, variation between motives seems to have fallen with intraclass correlation coefficient values of $ICC = 0.13$ and $ICC_{\text{Intercept}} = 0.04$ respectively. Whilst the latter statistic, the $ICC_{\text{Intercept}}$, is an adjusted version of the intraclass correlation coefficient, it still remains useful to compare both statistics [7]. Less variability between age at first kill and race is seen at the motive level. Is multilevel modeling justified to answer the hypothesis given in section **(4.3)** and **(5.1)**, and if so, which presented model RIM or RSM is best suited to answer our hypothesis?

6 Model Comparison

To determine whether multilevel modeling is justified to answer given hypotheses about serial killers, we should ask the obvious question that is, are the additional random terms really necessary? There might be some common sense reasons for including the random terms as mentioned during our model application examples for instance, the belief that the age at first kill of serial killers varies within motives nevertheless, we can also ask if the observed data supports this from an objective standpoint. So our focus becomes under which of these models, the GLM applied in section **(2)** against their respective multilevel model extensions: VCM, RIM or RSM is our observed sample "most likely" to have occurred?

6.1 Likelihood Ratio Test

The likelihood ratio test can be used to compare nested models that is, the terms present in one of the two compared models in question is a subset of the latter. This test is useful in determining if whether the addition of a feature input is useful in answering the hypothesis in question as compared to the exclusion of the feature. To determine the likelihood of our data under a given model, consider the joint probability density of our sample under that model, $f(y_{11}, y_{21}, ..., y_{n_m m})$, evaluated at the observed values of $y_{11}, y_{21}, ..., y_{n_m m}$, where n_m denotes the number of individuals situated in cluster $j = m$. The greater this density value is the more "likely" these observations are under that model. However, when computed at the observed values, $f()$ is no longer thought of as a function of the sample, instead, it is a function of the unknown model parameters. For example in a random intercepts model, the likelihood is defined as

$$L(\beta_0, ..., \beta_p, \sigma^2_{e0}, \sigma^2_{u0}) = f(y_{11}, y_{21}, ..., y_{n_m m}), \tag{14}$$

where our aim would be to maximise the likelihood subject to varying these unknown model parameters $\beta_0, ..., \beta_p, \sigma^2_{e0}, \sigma^2_{u0}$, with fixed values that accomplish such a maximisation taking the values $\hat{\beta}_0, ..., \hat{\beta}_p, \hat{\sigma}^2_{e0}, \hat{\sigma}^2_{u0}$ respectively. These values best explain the data under the framework of the model given and a formally referred to as maximum likelihood estimates (MLEs) of the parameters. As mentioned in section **(3.1)**, multilevel modeling uses a slightly more complicated parameter derivation method that is, iterative generalised least squares however, such a method converges to maximum likelihood methods and so applying likelihood ratio tests to multilevel models is justified [8]. Define our maximum likelihood estimate as

$$\widehat{L} = L(\hat{\beta}_0, ..., \hat{\beta}_p, \hat{\sigma}^2_{e0}, \hat{\sigma}^2_{u0}), \text{ for the random intercept model case.} \tag{15}$$

Then, to compare two nested models, A and B, where A is nested within B, we can consider their maximised likelihoods $\widehat{L}_A$ and $\widehat{L}_B$ by taking the likelihood ratio $\widehat{L}_B/\widehat{L}_A$. If such a ratio takes a value close to 1, then there is no meaningful reason to prefer the model B over A because choosing such a model does not increase the likelihood of the data by a substantial amount. Alternatively, if the ratio is "large" then it seems that the likelihood of the data is greater under B. A measurement of such model disparity can be calculated via the deviance statistic.

6.1.1 Deviance Statistic

The deviance statistic is defined as

$$D = 2log\frac{\widehat{L}_B}{\widehat{L}_A}, \text{where } \widehat{L}_A \text{ and } \widehat{L}_B \text{ are the maximum likelihood values of models A and B respectively.} \tag{16}$$

Under the assumption that model A is correct we can assume $D \sim$ approx χ^2_d, where d denotes the number of additional parameters introduced in model B as compared to A [8]. The larger the deviance statistic of our model comparison, the greater evidence in favour of preferring B over A. We test the the hypotheses:

$$\begin{cases} H_0: & \text{All additional parameters in model B that are not in A are zero, against the alternative} \\ H_1: & \text{at least one of the additional parametrs in model B is nonzero.} \end{cases}$$

Basing our test on the significance level $\alpha = 5\%$, as standard. A test in which we reject H_0 if and only if

$$D \geq \chi^2_d(\alpha) \tag{17}$$

is called a **likelihood ratio test** with significance level α for model B against that of model A. Note, $\chi^2_d(\alpha)$ is defined as the $(1-\alpha)$ quantile of the χ^2_d distribution i.e., the 95% quantile of the chi-squared distribution with d degrees of freedom. $\chi^2_d(\alpha)$ is the number such that the $P(D \geq \chi^2_d(\alpha)) = \alpha$. It is worth noting that the deviance statistic does not consider the likelihood of the models in question but the log-likelihood $D = 2log(\widehat{L}_B/\widehat{L}_A) = 2\left[log(\widehat{L}_B) - log(\widehat{L}_B)\right]$. Reasoning for this comes from the logarithm being a monotonically increasing function- parameter values that maximise the likelihood will indeed maximise the log-likelihood. Furthermore, computational efficiencies are brought to the table when considering logarithm-transformed function in comparison to applying computation to the raw values themselves [9].

6.1.2 Application

Model A	*Model B*
AgeFirstKill ~ 1	AgeFirstKill ~ 1 + (1|Motive)
AgeFirstKill ~ Sex	AgeFirstKill ~ Sex + (1|Motive)
AgeFirstKill ~ Race	AgeFirstKill ~ Race +(1|Motive)
AgeFirstKill ~ Race	AgeFirstKill ~ Race +(1+Race|Motive)
AgeFirstKill ~ Race +(1|Motive)	AgeFirstKill ~ Race +(1+Race|Motive)

Likelihood $log\widehat{L}_A$	*Likelihood* $log\widehat{L}_B$	*Deviance Statistic D*	*Quantile* $\chi^2_d(\alpha = 5\%)$	*Null Hypothesis* H_0
-6308.093	-6272.595	70.997	5.024	reject
-6304.245	-6272.426	63.637	5.024	reject
-5893.685	-5862.231	62.909	5.024	reject
-5893.685	-5849.818	87.734	7.378	reject
-5862.231	-5849.818	24.825	5.024	reject

Table 7: Likelihood Ratio Test Statistics For GLM, VCM, RIM & RSM

Table **(7)** shows the generalized linear models, variance components model, random intercept and random slopes models proposed and analysed in sections **(2-5)** respectively. The simpler model i.e., the model with fewer parameters is proposed as model A, and consists as a nested structure within the proposed, more complex model B. For example, the first row of table **(7)** compares the most simple GLM with no feature inputs **AgeFirstKill** ~ 1, to that of the multilevel model extension, the VCM given by **AgeFirstKill** ~ 1 + (1|**Motive**). Subsequently, the GLM with the inclusion of a feature that is, **Sex**, is compared to its multilevel counterpart, the RIM given by **AgeFirstKill ~ Sex +** (1|**Motive**).

Further model comparison are given, with the latter comparing whether the inclusion of the possibility that the effect of a covariate on the response might also vary between clusters provides greater "likelihood" in our findings- RIM vs RSM given by **AgeFirstKill ~ Race +** (1|**Motive**) against **AgeFirstKill ~ Race +** (**1+Race**|**Motive**). Likelihood ratio test statistics are seen for each model comparison. The maximum log-likelihood estimated value of the models $log\widehat{L}_A$ and $log\widehat{L}_B$, the deviance statistic D, the quantile threshold $\chi^2_d(\alpha = 5\%)$ and the result of the hypotheses proposed during a likelihood ratio test- see section **(6.1.1)**. For example, The deviance statistic in the first row of table **(7)** is

calculated using the definition provided in section **(6.1.1)** giving:

$$\begin{aligned}\text{Deviance Statistic } D &= 2\left[log\widehat{L}_{\text{A=GLM Base Case}} - log\widehat{L}_{\text{B=VCM}}\right] \\ &= 2\left[-6308.093 + 6272.595\right] \\ &= 70.997.\end{aligned}$$

All quantile thresholds $\chi^2_d(\alpha = 5\%)$ take equivalent values that of $\chi^2_{d=1}(\alpha = 5\%) = 5.024$ except for the penultimate model **AgeFirstKill ~ Race** against **AgeFirstKill ~ Race + (1+Race|Motive)**, taking quantile threshold value $\chi^2_{d=2}(\alpha = 5\%) = 7.378$. This is because all model comparison seen only consider an increase in complexity of one parameter, resulting in a net difference in degrees of freedom between model A and B as $d = 1$- the penultimate model jumps up by two. Throughout all model comparisons we reject the null hypothesis H_0 that is, all additional parameters in model B that are not in A are zero. Since our deviance statistic $D \geq \chi^2_d(\alpha)$ we can say that, under the assumption that model A is correct (i.e. the null hypothesis that there is no random serial killer effect) our deviance statistic is improbably large- less than 5% chance of being so large. Hence observing such a large statistic is strong enough evidence against our null hypothesis H_0 that we can reject it at the 5% significance level, in favour of the hypothesis that model B is preferred.

Thus we may conclude from likelihood ratio testing that multilevel modeling is indeed justified to answer given hypotheses about serial killers; additional random terms really can be deemed as necessary and provide more insight, more "likelihood", in our findings. Likelihood ratio testing is one of many tools that could be used to compare multilevel models. More comparison statistics should be considered in order to justify one model over another.

6.2 Further Model Comparison Metrics

So far we have focused on a maximum likelihood based model selection, a measure of how well the model is able to fit to the available data, when the parameters are chosen to make this fit as good as possible. A possible issue with this approach is that with more parameters we can almost always tweak the values to make the likelihood just a little higher. For instance, consider the RSM given by **AgeFirstKill ~ Race + (1+Race|Motive)**:

$$y_{ij} = (\beta_0 + u_{0j}) + (\beta_1 + u_{1j})x_{1ij} + e_{0ij}, \tag{18}$$

where y_{ij} denotes the **AgeFirstKill** of the i-th serial killer in the j-th cluster, **Motive**. During maximum likelihood based model selection such as the likelihood ratio test, we vary the parameters $\hat{\beta}_0, ..., \hat{\beta}_p, \hat{\sigma}^2_{e0}, \hat{\sigma}^2_{u0}, \hat{\sigma}^2_{u1}$ with the goal to maximise the "likelihood" of the observations occurring $\widehat{L} = L(\hat{\beta}_0, ..., \hat{\beta}_p, \hat{\sigma}^2_{e0}, \hat{\sigma}^2_{u0})$. Consider the RIM given by **AgeFirstKill ~ Race + (1|Motive)**:

$$y_{ij} = (\beta_0 + u_{0j}) + \beta_1 x_{1ij} + e_{0ij}, \tag{19}$$

where y_{ij} denotes the **AgeFirstKill** of the i-th serial killer in the j-th cluster, **Motive**. The RSM is equivalent to that of RIM with the inclusion of some perturbation in our covariate term via $u_{1j} \sim N(0, \sigma^2_{u1})$. We can always make the RSM consist of a greater (or at least equivalent) maximum likelihood just by first setting the level 2 residual$\hat{\sigma}^2_{u1}$ to zero and varying the value such that the likelihood increases. In table **(7)**, the fit for the RSM model is only marginally improved, in terms of likelihood, than the RIM with maximum log-likelihood values of -5862.231 and -5849.818 respectively. We might, qualitatively, think "that little bit of extra likelihood isn't worth having an extra parameter". How do we make that idea quantitative? One way is to apply a hypothesis test as seen in the likelihood ratio method. Another is to use a penalised likelihood that explicitly "punishes" the extra parameters. Usually this takes the form of an extra term, in the log-likelihood, proportional to the number of parameters p, with a penalisation strength λ:

$$logL_{\text{PENALISED}} = log\widehat{L} - \lambda p, \tag{20}$$

where $log\widehat{L}$ is the usual maximum log-likelihood. Unfortunately there is no absolute consensus in the statistical community on how to choose λ [5]. It may depend on your personal experience or needs regarding model complexity.

6.2.1 Akaike Information Criterion (AIC)

Probably the most prevalent penalised likelihood is known as the **A**kaike **I**nformation **C**riterion (AIC). This is defined as:

$$AIC = 2p - 2log\widehat{L}, \tag{21}$$

effectively setting λ, seen in equation **(20)**, to be equal to one, and the model that minimises the AIC is deemed to be the best [5]. Setting λ equal to one makes equation (19) proportional to the AIC definition. It is proportional due to the fact the AIC is not just some arbitrary formula. Instead the Akaike information criterion is derived from minimizing the Kullback–Leibler divergence, a type of "statistical distance" that measures how one probability distribution is different from a second. Further information can be found in Hox J's (et al) Handbook of Advanced Multilevel Analysis [5].

6.2.2 Bayesian Information Criterion (BIC)

Similar to AIC, the **B**ayesian **I**nformation Criterion (BIC) can be used as a measure regarding model selection, penalising the likelihood of the models not by setting λ equal to one as seen by the AIC metric, instead, $\lambda = log(n)/2$ with n denoting the number of observations of serial killers present hence:

$$BIC = plog(n) - 2log\widehat{L}, \tag{22}$$

where $log\hat{L}$ is the usual maximum log-likelihood and p the number of parameters. Unlike the Akaike information criterion that derives from a frequentist approach to comparing models, BIC is a metric derived from a Bayesian approach, where prior beliefs of the model in question our taken into account [5].

6.2.3 Application

Model	*AIC*	*AIC Rank*	*BIC*	*BIC Rank*	*ICC*	*ICC Rank*
AgeFirstKill ~ 1	12620.19	8	12631.14	8	N/a	N/a
AgeFirstKill ~ 1 + (1|Motive)	12551.19	6	12567.61	4	0.129	2
AgeFirstKill ~ Sex	12614.49	7	12630.91	7	N/a	N/a
AgeFirstKill ~ Sex + (1|Motive)	12552.85	5	12574.75	5	0.126	3
AgeFirstKill ~ Sex + (1+Sex|Motive)	12548.73	4	12581.58	6	$0.089_{\text{Intercept}}$	4
AgeFirstKill ~ Race	11793.37	3	11809.59	3	N/a	N/a
AgeFirstKill ~ Race + (1|Motive)	11732.46	2	11754.09	2	0.132	1
AgeFirstKill ~ Race + (1+Race|Motive)	11711.64	1	11744.08	1	$0.042_{\text{Intercept}}$	5

Table 8: Further Model Comparison Metrics For GLM, VCM, RIM & RSM

As seen in table **(8)**, model comparison metrics may be applied to our previously seen examples. The table shows the AIC, BIC and ICC values for the multilevel models and the generalized models seen in section **(2.1)**. The meaning behind intraclass correlation coefficient (ICC) values can be found in section **(3.1)** with an example of computation given. ICC values have been left regarding generalized linear models because the intraclass correlation coefficient is a measure of variability between individuals at a cluster level, thus making in not applicable in the GLM case. For RSM, an adjusted ICC value is given and is defined by equation **(13)**. An example of how the AIC value for the first model, **AgeFirstKill ~ 1** is as follows:

$$
\begin{aligned}
AIC_{\text{GLM Base Case}} &= 2p_{\text{GLM Base Case}} - 2log\widehat{L}_{\text{GLM Base Case}}, \\
&= 2(2) - 2(-6308.093), \\
&= 12620.19.
\end{aligned} \tag{23}
$$

A ranking between all cases can be seen taking values from 1, the model produces the most optimum performance metric relative to other case to 8, the model performs least. This ranking format can be used to compare desired cases for example, we may ask the following question:

- Does the effect of **Race** vary between serial killers clustered by **Motive**?

Whilst the previous model comparison method, the likelihood ratio test, suggested a random effect to be most applicable, i.e., the RSM **AgeFirstKill ~ Race +** (**1+Race**|**Motive**) has a greater preference at explaining the data observed in comparison the to the RIM **AgeFirstKill ~ Race +** (1|**Motive**); metrics found in table **(8)** suggest otherwise. Whilst the RSM does indeed have greater performing AIC and BIC values in comparison to the RIM, these differences are only marginal- (RIM, RSM) ~ (11732.46, 11711.64) for the Akaike information criterion, and (RIM, RSM) ~ (11754.09, 11744.08) for the Bayesian information criterion. A greater differentiation is prevalent regarding intraclass correlation coefficient values (RIM, RSM) ~ (0.132, 0.042) suggesting that the majority of variance explained is seen from the fixed effect, not the random effect of the models [8].

References

[1] Nicole H. Augustin, Erik-AndrÃ© Sauleau, and Simon N. Wood. On quantile quantile plots for generalized linear models. *Computational Statistics and Data Analysis*, 56(8), 2012.

[2] University Bristol. *Multilevel Modelling online course: LEMMA VLE Centre for Multilevel Modelling. (www.cmm.bris.ac.uk/lemma/)*, 2012.

[3] Julian J Faraway. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Chapman and Hall/CRC, 2016.

[4] Harvey Goldstein. *Multilevel statistical models*. John Wiley & Sons, 2011.

[5] Ellen L Hamaker, Pascal van Hattum, Rebecca M Kuiper, and Herbert Hoijtink. Model selection based on information criteria in multilevel modeling. *Handbook of advanced multilevel analysis*, pages 231–255, 2011.

[6] Joop J Hox, Mirjam Moerbeek, and Rens Van de Schoot. *Multilevel analysis: Techniques and applications*. Routledge, 2017.

[7] Shinichi Nakagawa, Paul CD Johnson, and Holger Schielzeth. The coefficient of determination r 2 and intraclass correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, 14(134):20170213, 2017.

[8] James L. Peugh. A practical guide to multilevel modeling. *Journal of School Psychology*, 48(1):85–112, 2010.

[9] JosÃ© C. Pinheiro and Douglas M. Bates. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, 4(1):12–35, 1995.

[10] Hugo Quené and Huub Van den Bergh. On multi-level modeling of data from repeated measures designs: A tutorial. *Speech communication*, 43(1-2):103–121, 2004.

[11] Mark J. Schervish. P values: What they are and what they are not. *The American Statistician*, 50(3):203–206, 1996.

[12] Holger Schielzeth. Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution*, 1(2):103–113, 2010.

[13] Samaradasa Weerahandi. Generalized confidence intervals. *Journal of the American Statistical Association*, 88(423):899–905, 1993.