Analysis of Longitudinal Data for Paediatric Brain Trauma

Submitted by:

Lisa Henderson

Supervisor:

Robin Henderson

Submitted in the support of:

Mathematics and Statistics MMathStat
May 2015

Abstract

This project describes the analysis of data on recovery of motor function for 74 paediatric study cases. Each child has between 1-20 measurements taken upon them after series injury has occurred. The aetiology of the injury is split into three categories: ABI, HYPOXIC and TBI.
After exploratory data analysis and multiple regression techniques, we examine the correlation and normality structure of this longitudinal data. General estimating equations and random effect techniques is used for modelling the data. We compare methods through simulation and misspecification and assess the accuracy of model-based prediction.

Contents

1 Introduction

1.1 Longitudinal data analysis

Longitudinal data analysis (LDA) has been a growing topic through statistics. Also known as panel data analysis, it concerns data with a certain structure. The characteristics of the data are that a series of measurements are taken upon several subjects. The reason that we wish to keep the structure of repeated measurements is due to the possibility of dependence between observations on the same subject.

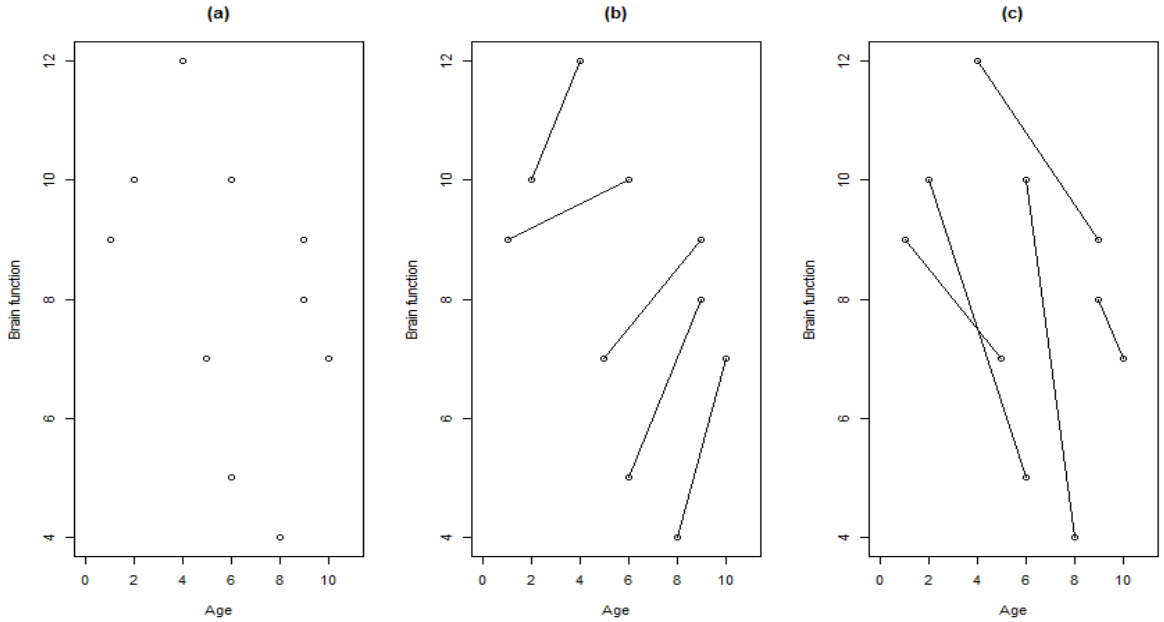


Figure 1: (a) Hypothetical data, (b) Incorporating a LDA structure and (c) Alternate LDA structure

Figure 1(a) shows some hypothetical data on children's brain function and age. It appears to suggest that the older a child is, the lower the brain function they have. Now we incorporate the fact that this data was collected on several children on a particular treatment with two measurements per child. Figure 1(b) still suggests that the older children have a lower brain function, but it now also shows that the children's brain function is increasing with time, possibly caused by a response to a particular treatment. Analysing Figure 1(c), there is now an alternate conclusion. Once again, it reiterates the point from Figure 1(a), yet there is a decreasing effect with respect to time, which could be down to the treatment failing over time. Hence we have three different explanations for the same set of data points.
It is clear to see from this simple example that studies that have repeated measurements taken on subjects need to incorporate this structure into their model. A main advantage of LDA is the benefit of being able to look at the change of an effect with respect to time. There are other methods that may be appropriate, such as time series analysis described

by Diggle (1990) [4]. The difference between the methods is that we usually use time series analysis for one, or a small number, of long series, whereas LDA is used for a large number of short series. The idea is that each patient follows a general pattern, but with some patient-specific differences. Diggle (2002) [5] explains this as "inferences can be made by borrowing strength across people", meaning that wrongly assumed model assumptions can be made to have a less severe impact on the LDA model than time series analysis.
Within this study, we wish to explore the effects of wrongly assumed models and the importance of modelling patient-specific effects. We will look at the use of two main methods of LDA (GEE and Random Effects) in the analysis of data on the brain function of children after injury. We will consider the fit of various models and then explore in more detail using simulation and prediction methods.

1.2 Gross Motor Function Measure (GMFM-66)

Kelly et al (2014) [8] explore the Gross Motor Function Measure for 74 patients recovering from severe brain injury. The paediatric study consists of patients ranging from the age of 0.3 months to 17 years. The possible injuries are categorised into three main groups: Acquired Brain Injury (ABI), Traumatic Brain Injury (TBI) and Hypoxic Brain Injury (HYPOXIC), which is the starvation of oxygen to the brain.
ABI injuries occur after birth; they are not hereditary. ABI's are a result of a cause that damages the brain such as a stroke, hemorrhage, infection or toxic substances. TBI injuries are such that the patient has undergone a jolt, causing the brain to hit against the skull. This could be the result caused by accidents such as a car accident, fall or an assault. Both ABI and TBI may then undergo further damage from consequent bleeding, therefore increasing the pressure surrounding the brain. The third injury is a little different. HYPOXIC is the starvation of oxygen to the brain caused from, for example, drowning, respiratory or cardiac arrest, asphyxiation or suffocation. This can result in irreparable brain cell damage [16]. Geddes (2001) [7] found that "global hypoxic damage was the most common histological finding"amongst 37 infants (9 months old or less) when exploring the neuropathology of severe brain injuries to infants that result in death. All in all, the injuries sustained in our data collection are often critical.
The result of these injuries can affect children in a variety of ways. Physical issues can occur which include muscle weakness, spasticity and restricted range of movement. It can also affect them cognitively such as problems with memory, concentration, problem solving ability and self monitoring [16]. An area of concern within the early rehabilitaion period is the child's ability to re-obtain use of motor functional skills. In this study, we will assess this gross motor function recovery using the Gross Motor Function Measure (GMFM-66) scores.
GMFM-66 is a subset of measurements from the GMFM-88. This is a test of 88 individual gross motor function attributes in 5 areas (lying and rolling, crawling and kneeling, sitting, standing and walking, running and jumping). They are scored on a 4 point scoring system (0 to 3). The patient undergoes each test and is rated on how well they can carry out each task with 0 being that they did not initiate the task. A percentage

score is then calculated from this. The GMFM-66 is calculated through the use of Gross Motor Ability Estimator (GMAE) from these measurements [11]. GMFM-66 is primarily used to tailor the care of cerebral palsy patients to assess their mobility rather than brain function and so worries over the ability to capture the complexity of the brain may arise. Using conclusions drawn from Linder-Lucht et al [10], who concluded that GMFM was sufficiently sensitive to change when analysing paediatric TBI injuries, we believe it can be a correct way to measure brain function.
The main aims of Kelly et al (2014) [8] were to explore

- the use of GMFM-66 as a measure of recovery in paediatric brain injuries
- the association between aetiology (*cause of the brain injury*) and gross motor recovery
- the effect of age on the gross motor recovery

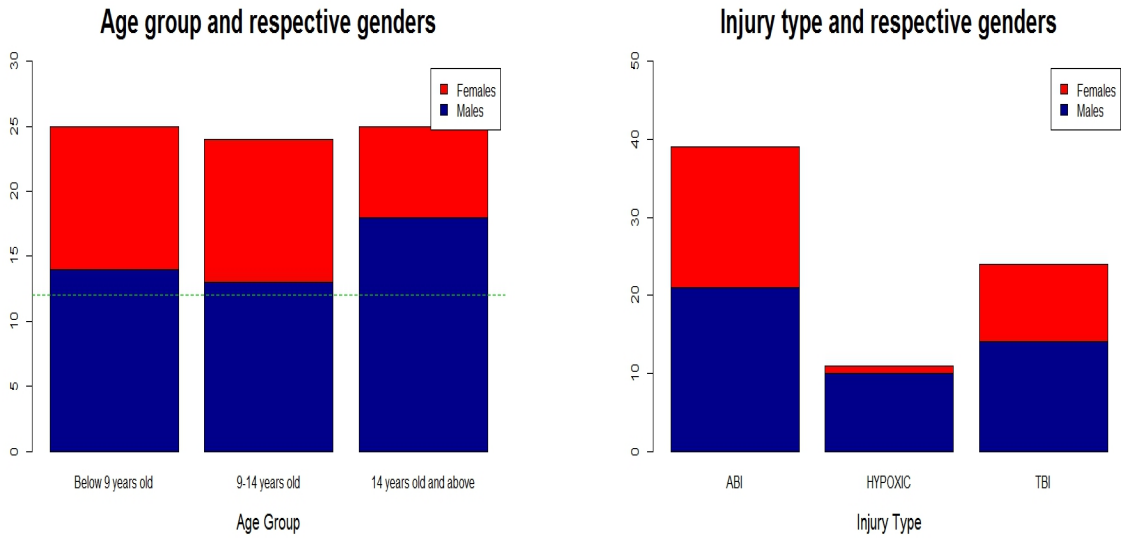
The aim of this project is to address the latter two points via longitudinal data analysis. Some issues will be explored along the way such as the standard assumptions of normality, problems over within-patient correlation and the effects of patient specific errors.

2 Exploratory Data Analysis (EDA)

The raw data can portray a lot of information. Through EDA we will explore what the data tells us to give us a rough idea of what we should be expecting. For instance, we can draw basic conclusions like that suggested from Figure 1(a) discussed in the Introduction, through linear regression in Section 2.3.

2.1 Introductory analysis of Kelly et (2014) [8]

As stated, we will consider the data on 74 children; the ages ranging from 0.3 months to 17 years. To look more closely, we have split up the ages into three groups: Younger (below 9 years old); Medium (9-14 years old) and Old (above 14 years old) of sizes 25, 24 and 25 respectively. There are more boys than girls within the older group; shown in Figure 2(a). Note that the dotted line represents a 50% gender split.
As in the age group categories, Figure 2(b) suggests that there is a similar proportion of male and females in ABI and TBI injury types (58% and 53% males respectively), but having a significantly different gender proportion in the HYPOXIC group (90% males). In addition, the number of patients within each injury type is rather unequal: we have 39 ABI patients, 24 TBI patients and 11 HYPOXIC patients. We only have one female HYPOXIC patient in the study which may lead to inaccuracy in our modelling (later) as there may be too few females within the HYPOXIC group to allow sensible inference.



(a) Gender and age groups

(b) Gender and types of injuries

Figure 2: Gender distribution throughout age groups/injury type

We also find that there is a difference in the mean age per each injury group. The difference in the HYPOXIC group may lead to future transformations being used to reduce some of the variation. Overall, the mean age of the patients in this study is 11 years old, matching up with ABI and TBI (12 and 11 years respectively). However, the mean age of the HYPOXIC group is 6 years old which differs from the other two injury types once again. Hence if we see that age is significant in our model, it may be confounded with the effect of injury type. For instance, if the younger group is found to recover slowest, is this down to them being younger or down to more of them having a HYPOXIC injury?

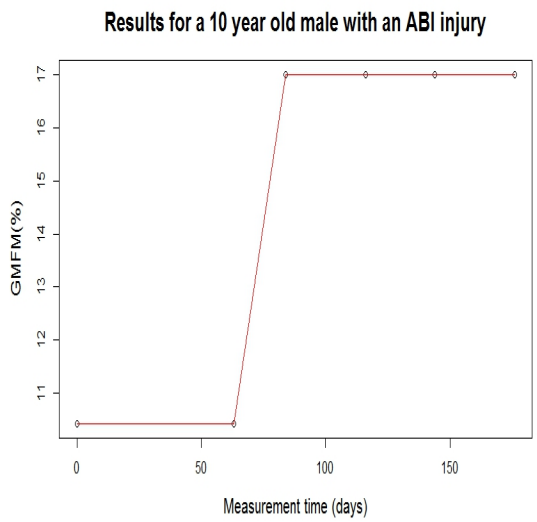
Measurements per child	ABI	TBI	HYPOXIC	TOTAL
Minimum	1	1	1	1
Maximum	10	8	20	20
Median	3	3	3	6
Mean (1 s.f.)	3	4	6	3
Total	133	86	68	287

Table 1: Number of serial measurements on patients in each injury type group

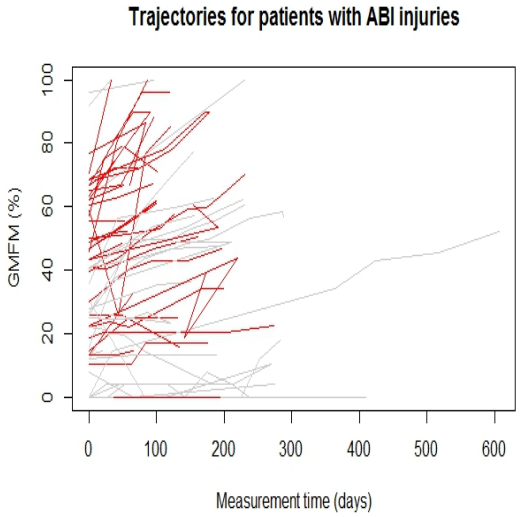
Each patient had repeated GMFM measurements carried out over a space of time, typically 126 days (max period of 606 days for a TBI patient). Table 1 shows that typically ABI and TBI patients had 3 and 4 measurements taken, whereas HYPOXIC had slightly more (6 GMFM measurements). We have 13 patients who only had one measurement taken, which will not help identify any recovery rate trajectories as there was no follow up after their initital GMFM score was taken.

2.2 Typical trajectories

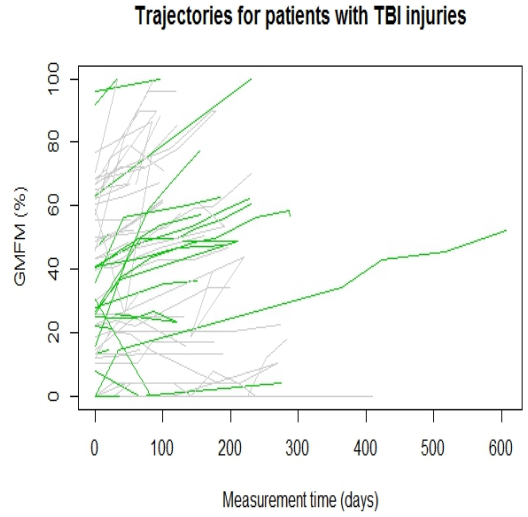
What does a GMFM measurement trajectory typically look like? We illustrate this using a 10 year old male who has sustained an injury in the ABI category, as shown in Figure 3(a). We see that he started with around 10% brain capability, but recovered to a score of 17%. It is important to note that it is only 17%, as it shows that this little boy is still very poorly. His measurements were taken over a period of around 180 days, consisting of 6 GMFM measurements in total. We can see a distinct s-shape curve which is typical of the patients in the study.



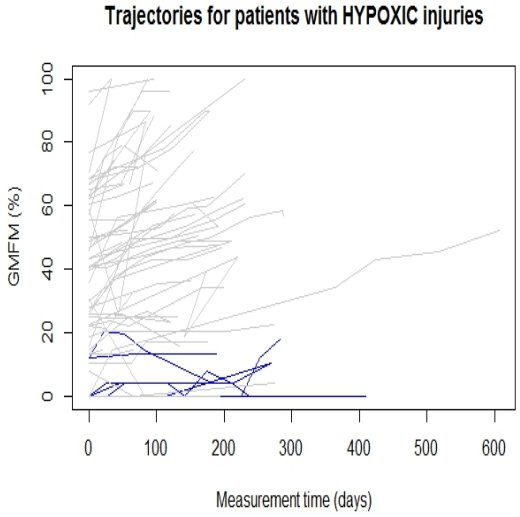
(a) 10 year old, ABI injured, male patient



(b) ABI injured patients



(c) TBI injured patients



(d) HYPOXIC injured patients

Figure 3: Example trajectory (a) and typical trajectories for each injury type (b-d)

In Figure 3(b-d), we show all patient trajectories, where red represents ABI injuries, green represents TBI injuries, and blue represents HYPOXIC injuries. When comparing all of the patients with regards to their given injury types, there are some fairly clear characteristics.
The ABI injured patients generally have positive slopes for their trajectories (signifying the patients get better) and all of a similar length (similar measurement period). The TBI injured children tend to have a slightly lower GMFM score, but still have a positive slope (recover) in general. They also appear to have a slightly longer measurement period on average, noting one particular child having been measured 606 days.
HYPOXIC is very different to the above two injuries. There are fewer trajectories (fewer patients in the study) for this injury type and the average is considerably lower on the GMFM percentage. They are all below 20%, suggesting that the children are seriously ill, and have not recovered anywhere near their full brain capacity. We can also note that the slopes are a lot less positive, more flat, suggesting that children do not recover much brain function with this particular injury type.
When referring to the recovery, we see that each child starts and ends at different GMFM percentages, hence we spot the potential need for patient-specific modelling. We can see, just by looking at the plotted data that there are lots of factors to consider, such as within-patient correlation. It is evident that the injury type that the child sustained has an effect on, not only the general GMFM score, but also the slope/recovery rates for the children.

2.3 Multiple linear regression

The main aim throughout this report is to model the brain recovery and to see what affects the GMFM scores after the monitoring period begins. In order to do this, we need to look at which variables may be important, and the general relationship they have with GMFM scores. As a starting point, we look at a regression analysis, not as a final model, but to get a brief understanding of the effect aforementioned.
To start our regression, the model that we begin with is:

$$GMFM = \beta_0 + \beta_1 * \text{TBI} + \beta_2 * \text{HYPOXIC} + \beta_3 * \text{Age} + \beta_4 * \text{Time3} + \beta_5 * \text{Female} + \epsilon \quad (1)$$

where β_i are estimated coefficients for each parameter and ϵ is the error term following $\epsilon \sim N(0, \sigma^2)$. We include the factor Injury Type which separates into ABI, TBI and HYPOXIC injuries, the factor Gender (male and female) and the variables Age and Time. In later Chapters, it was useful to consider age as a factor variable and so we may refer to Age as Age Group with the Younger Group being of ages less than 9 years old, the Older Group being more than 14 years old and the Medium Group being in between.
Within our regression, we see fit to transform the variable Time by taking the square root which will be discussed in Section 3.1. Note that in this, and future chapters Time3 now refers to $\sqrt{\text{Time}}$.
The baseline for this model is a male child who has sustained an ABI injury (and the

Age Group baseline is the Medium Group when used). We have used ordinary least squares to derive

$$\hat{\beta} = (\mathbf{X^T X})^{-1} * (\mathbf{X^T Y}). \tag{2}$$

Equation (2) is used to estimate the β_i parameters in multiple linear regression, as it is an unbiased estimator for when errors are assumed to follow a normal distribution and are independent. We fitted this as a linear model and examined Output 1. This showed that Age and Time3 both had very large, insignificant p-values, with Age having the largest p-value (0.8335). However, when fitting Equation (1) with Age Group rather than Age, Time3 was the most insignificant, but once again both still needed to be removed. This model is a poor fit as only $R^2_{adj} = 33.29\%$ of the variation is explained.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  42.94529    4.39359   9.775   <2e-16 ***
TBI          -8.09243    3.32852  -2.431   0.0157 *
HYPOXIC     -39.11297    4.10886  -9.519   <2e-16 ***
female        7.50817    3.15262   2.382   0.0179 *
age           0.06544    0.31095   0.210   0.8335
time3         0.10913    0.25583   0.427   0.6700
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 23.78 on 281 degrees of freedom
Multiple R-squared:  0.3446,Adjusted R-squared:  0.3329
F-statistic: 29.55 on 5 and 281 DF,  p-value: < 2.2e-16
```

Output 1: Regression model output, Equation (1)

We proceed with removing any insignificant variables. Our final multiple linear regression model looks like:

$$GMFM = 44.413 - 7.862 * TBI - 39.104 * HYPOXIC + 7.490 * Female \tag{3}$$

Output (2) shows the model fit after removing Age and Time3. The type of injury is highly significant, especially in regards to HYPOXIC with a p-values of order 10^{-16}, suggesting that the type of injury that the patient had, severely affects the GMFM measurements i.e. the brain recovery rate. TBI patients, in comparison to the GMFM scores for an ABI patient, generally have a lower score by around -7.86, with HYPOXIC being even lower. The β parameter estimate for this is -39.104 which is in agreement with what was suggested by the trajectories shown in Figure 3(d). The average initial GMFM score is 44.413, but this also suggests the inappropriateness of this model as the initial patient scores vary considerably for each patient from 0.1% to 99.9%. This is not captured via the intercept here (β_0). Once again, very little of the variability in the model is explained.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    44.413      2.431  18.273   <2e-16 ***
TBI            -7.862      3.282  -2.395   0.0173 *
HYPOXIC       -39.104      3.692 -10.592   <2e-16 ***
female          7.490      3.134   2.390   0.0175 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 23.71 on 283 degrees of freedom
Multiple R-squared:  0.344,Adjusted R-squared:  0.3371
F-statistic: 49.47 on 3 and 283 DF,  p-value: < 2.2e-16
```

Output 2: Final regression model output

Within multiple linear regression, we have two main assumptions about the errors. Firstly, the errors follow a normal distribution and the second being that the errors are independent (and hence uncorrelated). This can be assessed briefly with residual and QQ plots, which can also pin-point any high leverage or outlying points in our patient scores.

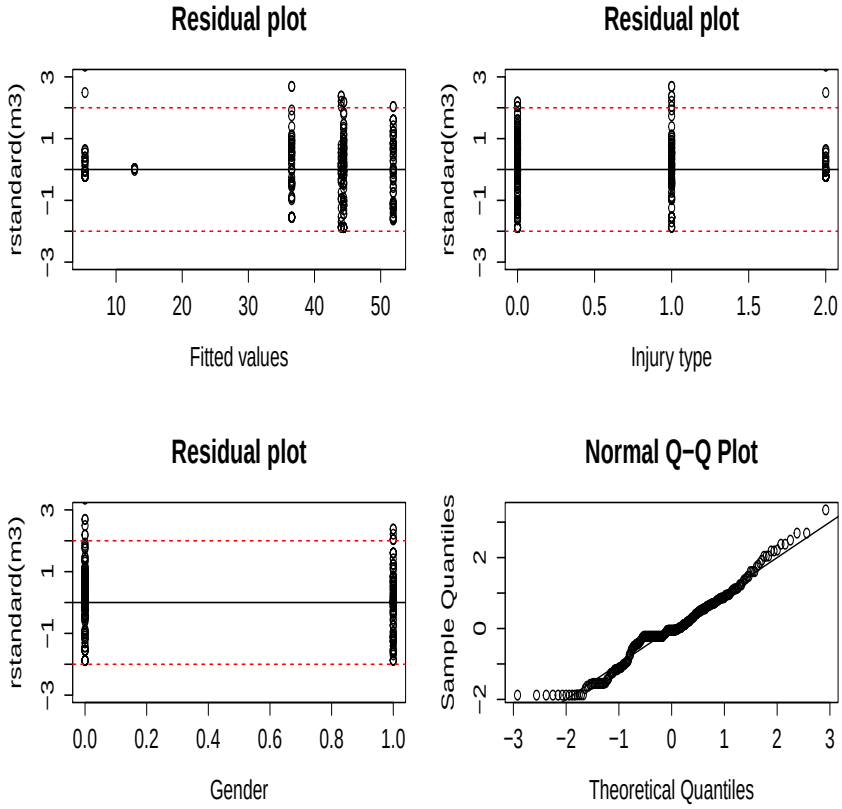


Figure 4: Residual and Q-Q plots of the final model

We need to check the residuals to assess normality. In the standardized residual plots, we hope to see a random scatter with few outliers being outside the region of ± 2, although we can expect approximately 5% of the data points to be outliers. Figure 4

shows that there are 6 main areas (two at around 45), which is to be expected as these represent the male and female counterparts for each injury type. There are quite a few outliers throughout each plot (those which are outside the ± 2 interval).
Looking at the Q-Q plot, it is in agreement with the idea that the model assumptions are possibly broken. The majority of the data does not follow a $N(0, \sigma^2)$ distribution and visibly shows there are a few ripples, suggesting that we have a departure from normality. This could be down to successive observations not being independent from each other, in particular within-patient correlation could be the cause of this. At the tails of the data, the data is flattened clearly suggesting that data, once again, does not follow a $N(0, \sigma^2)$ distribution. It also suggests possible influential points which may mislead conclusions on the overall pattern of the data. For instance an influential point that has a particularly high GMFM score, at say 500 days, may suggest that the general slope of the regression is more positive (than if it were ignored) leading to higher values when using the model to predict the GMFM scores.
We can assess whether there are any leverage points or outliers in our data set. In Figure 5 we see that there are five high leverage points. All of these are measurements taken from one HYPOXIC female. However, this patient is the only HYPOXIC female in our study and this is likely to be the reason that these points are very influential. There is a possible outlier in the Cook distance plot, but as we expect around 5% of the data to be outliers this is perhaps reasonable. To check, the model was re-evaluated without this outlier and also without the HYPOXIC female, but neither proved to make any significant change to justify removing them.

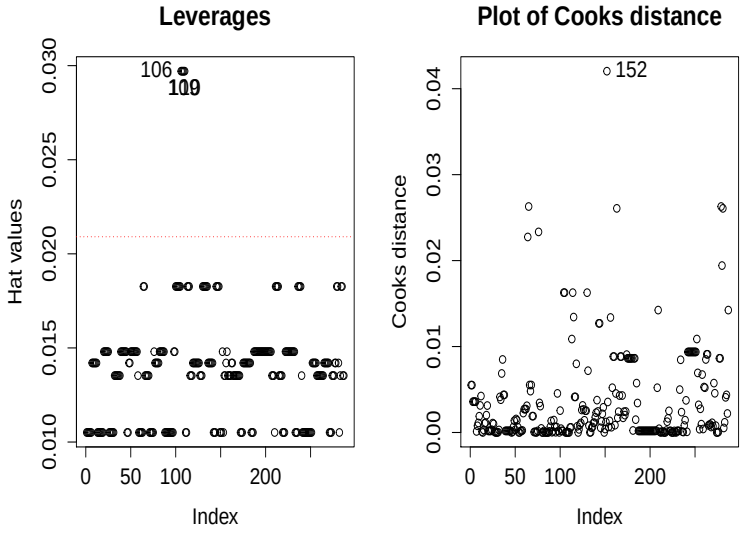


Figure 5: Leverage and Cooks Distance plots of the final model

We also applied regression analysis including interaction terms, however none of these proved to have a significant effect and we concluded that they were not relevant with this model type. Details are omitted.

2.3.1 Conclusion

From the regression analysis we see that the type of injury and the gender of the patient are highly significant in the recovery of their brain function. However, we must note that the assumptions that we have made in this section are unrealistic. The errors are distinctly not normal, and so we need to search for a possible transformation of the GMFM which will be evaluated in Chapter 3. The errors do not appear to be independent, as the QQ plot in Figure 4 show a big rippling effect possibly from the successive observations not being independent. This may be due to possible correlation which will be assessed in Chapter 4.
The final multiple linear regression model in Equation (3) has a very low R^2 value of 34.46%. Therefore very little of the variability in the data is explained from this linear model. In addition, the single intercept from the model does not look appropriate for our model suggesting that we need to incorporate different start and end points for each patient. From this we assume that multiple linear regression is unsuitable, as previously thought.

3 Non-Normality

When using multiple linear regression (MLR), we have the assumption that $\epsilon \sim N(0, \sigma^2)$, i.e. the errors follow a Normal distribution. We saw in Section 2.3 that this was not the case and in this chapter we will try to test the validity of this conclusion. The non-normality of the residual errors has two issues associated with it. Firstly, the tests and confidence intervals carried out in Multiple Linear Regression become invalid, and secondly, the least squares estimates (β coefficient estimates) might not be optimal. Therefore it is important to search for a transformation to try and overcome these problems. If the residuals do not appear to be normally distributed then we can address this in two ways:

1. Seek a transform for the covariates
2. Seek a transform for the response variable Y, GMFM

Hence, we will search for an appropriate transformation if needed. If normality is not the case and no transformation can be found, it will confirm our conclusion that the data cannot be modelled by MLR. However, normality is not expected here because of the s-shape curve trajectories; the curvature changes and it is not monotone (always has a negative or a positive slope).

3.1 Transforming covariates

Although the covariates do not need to be normally distributed, it may be useful to transform them. This can be down to a multitude of reasons such as:

1. Adjust for non normality in residuals
2. Make the variance of the response variable and the residuals homogeneous

3. Simplify the relationship between the response variable and the explanatory variables

Our main concern here is Point 1. Transforming the variable hopefully makes them more symmetric, in the aim to cancel out any skew that the variation in the variable causes the errors to have. The two variables that we consider are Age and Time (we cannot transform Injury Type or Gender as these are factors). The three typical transformations are: taking logs, square rooting or cube rooting.
The first step that we take is to analyse the boxplots for the given variables shown in Figure 15 and 16 in Appendix A. We found that Age looked most symmetric when left alone. The alterations made the variable either more skewed or created possible outliers. We concluded that this should be left without any editing.
Analysing the Time (in days since injury) we saw that the original variable was not very symmetric and appeared to have many outliers. The logarithmic transformation was not appropriate here as we have initial measurements at Time equal to zero. Both the square and cube root transformations do not make the variable very symmetric, but eliminates the worry over the outliers aforementioned. We concluded that the square root and cube root transformations may aid our search for symmetry.
The next step was to analyse the effect that they would make on the model residuals and ran three multiple linear regressions based on Equation (1). One with the Time covariate, one with the $\sqrt{Time}$ and one with the $\sqrt[3]{Time}$ variable. It is important to note that there is no specific need to have the variables being symmetric; we are only doing this to help assess the residuals and model fit.
After running the regression we can compare the most appropriate variable by looking at R^2, R^2_{adj} and the p-value testing significance to see if it should be included in the linear regression model. We want R^2 and R^2_{adj} to be large, noting that R^2_{adj} is a better value to assess the model with. Iit is the adjusted R^2 value accounting for the number of variables in the model. Ideally we would like the p-value to reflect that the Time variable is significant.

Regression including:	R^2	R^2_{adj}	P-Value
Time	0.3444	0.3328	0.7492
$\sqrt{Time}$	0.3446	0.3329	0.6700
$\sqrt[3]{Time}$	0.3447	0.3330	0.6486

Table 2: Table for three regressions assessing the effect of transforming the Time variable

Table (2) shows that transforming time has a small effect on R^2 and R^2_{adj} with all three regressions having a very low value. However, the best model shown is when Time is transformed by the cube root. It is the transformation as to which the p-value is closest to significance. Nevertheless it is still insignificant suggesting that it is not needed in this model. It has helped create a slightly better model, but unfortunately it has not helped enough to make the errors follow a normal distribution. Figure (14) in Appendix (A) shows the QQ-plot of the residuals from the model containing the cube root of Time. It was very clear that the errors are still not normal from excessive rippling and flattened

tails.
Another way in which we can test the best transformation for time is to use AIC, BIC and the log likelihoods. (*Note that AIC and BIC will be discussed in depth in Section 6.4.1*). A multiple linear model is created and includes all of the variables, but without any time variable. We then formulate three full models, but differing with respect to the time variable that is used. We first assess which of the three time models have the lowest AIC or BIC values and then compare the first model and with each of the three time models separately, testing the log likelihood ratio statistic (2∗(-log likelihood for the full model + log likelihood for model without time)) against a χ^2_p distribution where p is the difference in degrees of freedom.
Known as the likelihood ratio test, this assesses whether it it is better to include the time variable or remove it. Here a significant p-value would suggest to include the time variable. We would then pick the transformed (or not) version of time that gives the most significant p-value when tested against the first model. The results are not shown for simplicity, but the log likelihood test, AIC and BIC all agree that $\sqrt{\text{Time}}$ is the best version of the time variable. However the log likelihood test did show, in all three cases, that removing the time variable was better in the case of linear regression.
We will continue with transforming our time variable with the square root, from the assessment of the last three tests. Note that similar tests were carried out to assess the best transformation of Time in later models resulting in $\sqrt{\text{Time}}$ = Time3 being used (details omitted).
Although transforming time has (*slightly*) improved the amount of explained variation in our model, normality of the residuals has not been amended. Our next port of call is to look at transforming the GMFM scores.

3.2 Normal transformations

The model that that we are analysing is based upon the following equation.

$$Y = X\underline{\beta} + \underline{\epsilon}$$

We assume that each $\epsilon_i \sim N(0, \sigma^2)$ distribution, which causes us to assume that Y (in our case GMFM) follows a normal distribution. This is why we look for a transformation of the GMFM scores in order to have normal distributed errors in our model. Transforming our Y variable also helps us with Point 2 and 3 mentioned in Section 3.1.
We begin by examining the boxplots of GMFM and possible transformations, shown in Figure (6). This is to give us an indicator of the kind of transformation that would be suitable. The simple transformations that were analysed were once again: square root, cube root and log. In Figure 6, upon looking at the original GMFM measurements, we can see that the boxplot is clearly not symmetric with a large upper tail. Similarly the logarithmic transformation has a rather large lower tail. This implies that the errors will not be normal, as we have seen, and that a transformation is needed. The square root and cube root are more appropriate, the best here being the cube root. However, the mean is a sightly high on the boxplot for the cube root transformation.

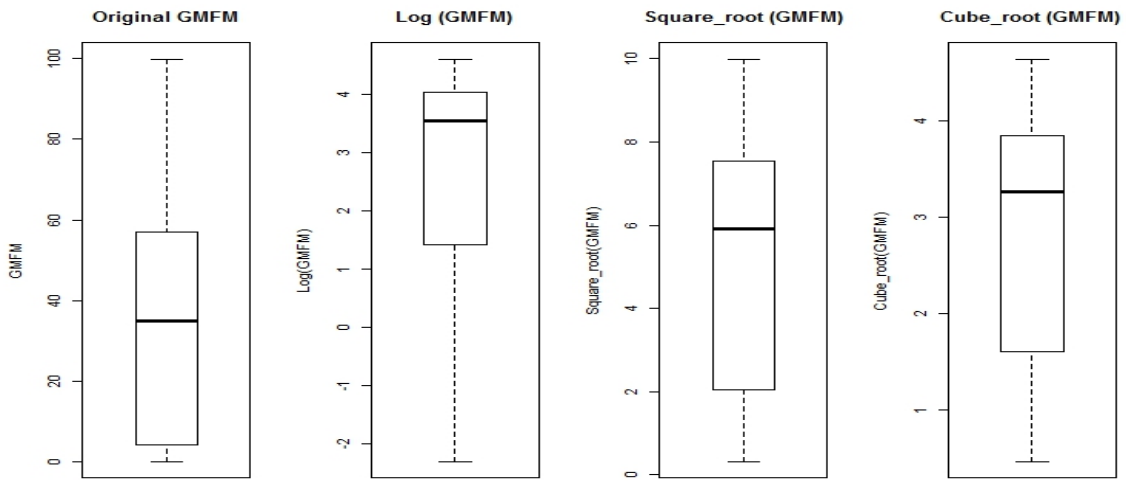


Figure 6: Boxplots of the GMFM transformations: square root, cube root and log

Referring back to Output 1, we notice that although both TBI and HYPOXIC are significant, HYPOXIC is much more so. This leads us to think that there may be something with this factor variable that is causing non-normality. Looking at Figure 7, we analyse the GMFM boxplots for symmetric variability once again, but with respect to the three types of injury that the patient may have sustained. It is clear that the HYPOXIC group is distinctively different to the other types of injury. We examined the three types of transformations and it was obvious that although the scores for ABI and TBI are similar, HYPOXIC GMFM scores remains highly different in all 4 cases (the original alongside the 3 possible transformations). We conclude that although transforming GMFM helped make the variability of HYPOXIC GMFM score slightly more symmetric, the variation still appears to differ from that of ABI and TBI groups. It is likely that one of the reasons for non-normality could be due to HYPOXIC GMFM measurements being highly different to those of ABI and TBI with non-symmetric variability.
Another regression was carried out using the different transformations to assess if they improved the amount of variability in the residuals that is explained by the model. From Figures 6 and 7, we expect that if there is an improvement, it will be with the square or cube root.

	GMFM	$\sqrt{GMFM}$	$\sqrt[3]{GMFM}$	$\log GMFM$
R^2	0.3444	0.4222	0.4348	0.4200
R^2_{adj}	0.3328	0.4120	0.4248	0.4096

Table 3: Table for three regressions assessing the effect of transforming the GMFM response variable following the model given including Injury Type, Gender, Time and Age

Table 3 shows that the best transformation in terms of explaining variability is the

cube root with R^2_{adj} marginally improving by 0.092. However the QQ-plots still implied non-normality.

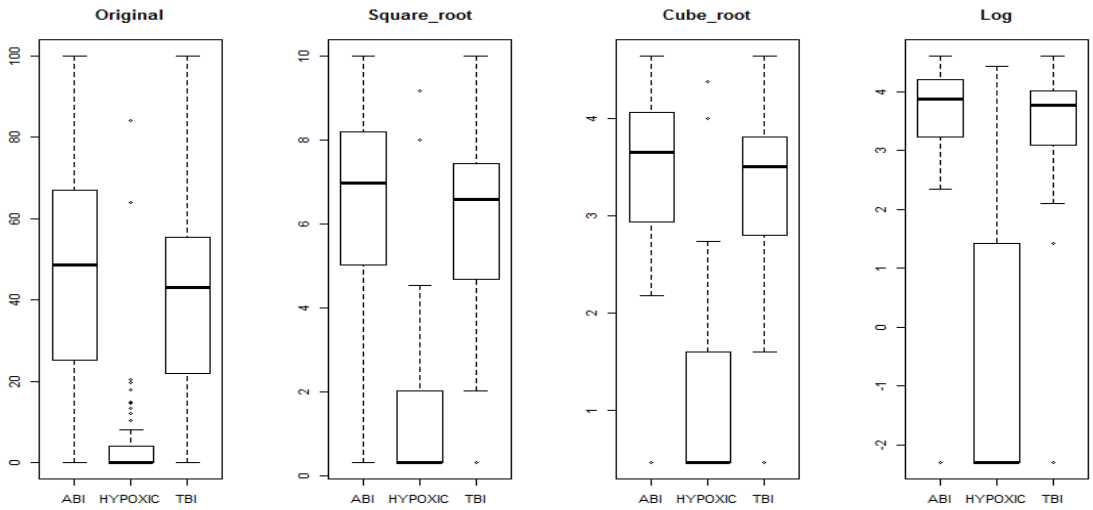


Figure 7: Boxplots of GMFM split into injury types (log(GMFM), square root(GMFM), cube root(GMFM)

We carried out a full regression anslysis using the transformation of GMFM, removing insignificant variables. This left the Injury Type, Gender and Age being the significant factors. Due to insignificance we removed Time (transformed or otherwise) which left the model with $R^2 = 0.4346$ and $R^2_{adj} = 0.4266$. Hence this was the best model found in any linear regression carried out (also assessing interactions).
To conclude, although using $\sqrt[3]{GMFM}$ improves the model, we still do not appear to have normally distributed errors.

3.2.1 Shapiro-Wilk test

Using boxplots are just a guide and assessment is a little subjective, so we turn to a different method. One form of transformation that we can consider is the Box-Cox transformation suggested by Box Cox (1964) [1]. This includes the transformation functions mentioned in Chapter 3.2, alongside many more. It gives us a larger variety of transformations to choose from. It is defined as follows:

$$y = \frac{GMFM^{\lambda} - 1}{\lambda}, \lambda \neq 0 \tag{4}$$

$$y = \log(GMFM), \lambda = 0 \tag{5}$$

We see that in Equation (5) that we get $\log(GMFM)$ when $\lambda = 0$, but with different vaues for λ we also get the other simple transformations mentioned above. It was also suggested by Box Cox (1964) [1] that Equation (5) be redefined to $y = \log(GMFM + \lambda_2)$ when $\lambda = 0$. The reason being is that these values would be greater than zero and we

could still look use it as only having one λ as a parameter in the model. However, we will stick with the first proposal. Using the Box-Cox function gives us a large range of possible transformations to use to coerce the errors to follow a normal distribution.
After a value of λ is chosen, how can we see how close the residuals are to a Normal distribution? We can use a test known as the Shapiro-Wilk test.
The Shapiro-Wilk test is on whether a sample $(x_1, \ldots, x_n)$ follows normality; here our sample would be the fitted residuals. The null hypothesis is H_0 : The sample data follows a normal distribution with the alternate being $H_1 : H_0$ not true. The test statistic used is:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

where $x_{(i)}$ is the ith ordered statistic with $x_{(1)}$ being the smallest. The constants $(a_1, \ldots, a_n)$ are obtained as:

$$(a_1, \ldots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{\frac{1}{2}}}$$

Here m is a vector of $E[x_i]$'s and V is the covariance matrix of the order statistics (ordered x_i's) [13].
With our particular data set, we would suggest that if we get a p-value (with our test statistic) that is less than 5% then we would say that there is enough evidence to suggest that our residuals are not normal. However, if the p-value is greater than 5% then we say that we don't have enough evidence to suggest a departure from normality.

3.3 Finding the optimal transformation

In order to pick the best version of the Box-Cox transformation, we need to choose the optimal value of λ. We consider two methods.

3.3.1 Method 1

One method that can be used to find the best value of λ is an iterative approach. We can perform a Shapiro-Wilk test on a range of possible $\lambda's$ and search for the one which gives the least significant p-value i.e. that which is closest to Normal. After doing so, the least significant p-value, that has been found, occurs when $\lambda = 0.84$.

λ Value	0.31	0.37	0.42	
P-Value	$1.657 * 10^{-10}$	$1.185 * 10^{-9}$	$8.108 * 10^{-9}$	
λ Value	0.82	0.83	0.84	0.85
P-Value	$2.492 * 10^{-4}$	$2.496 - 10^{-4}$	$2.523 * 10^{-4}$	$2.2518 * 10^{-4}$

Table 4: Shapiro-Wilk Hypothesis tests for Box-Cox transformations for Equation(6)

If this transformation was chosen, a regression analysis would be performed on the

following model:

$$\frac{GMFM^{0.84}-1}{0.84} = \beta_0+\beta_1*\text{TBI}+\beta_2*\text{HYPOXIC}+\beta_3*\text{Age}+\beta_4*\text{Time}+\beta_5*\text{Female}+\epsilon \quad (6)$$

Upon analysis of the residuals, shown in Figure 8, they still appear to be non-normal. Table 4 shows that the p-value for this λ is 0.0002523. It is still very significant at all levels. Therefore when $\lambda = 0.84$, using the Shapiro-Wilk test, we reject normality. Note that all of the $p - value \ll 0.05$ and so we reject normality for those Box-Cox transformations.

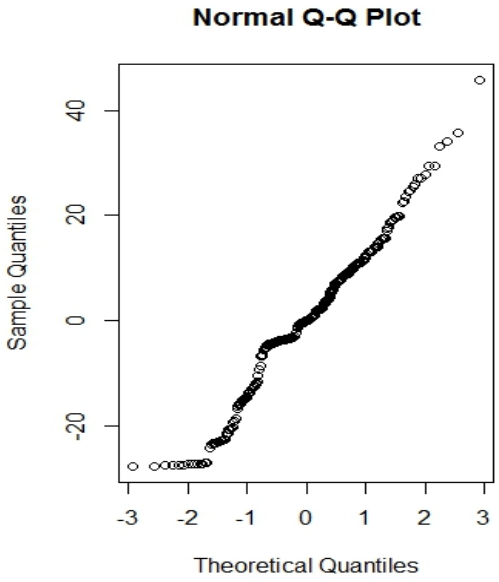


Figure 8: The QQ plot of the residuals using $\lambda = 0.84$

We evaluated this model again for the transformed variable Time. Using $\sqrt{\text{Time}}$ gave a p-value of 0.00027 with $\lambda = 0.84$ being the best, alongside $\sqrt[3]{\text{Time}}$ gave a p-value of 0.0002436 with $\lambda = 0.83$ being the best. From this we conclude that the best transformation would still be $\lambda = 0.84$ based on the model in Equation (6) as this gave the least significant p-value in the Shapiro-Wilk test. We still reject normality.

3.3.2 Method 2

An alternative approach in which we choose the optimal value of λ is via maximum likelihood estimation. Now using the Box-Cox transformation means that we can assume that the transformed GMFM follows a normal distribution. i.e. we can use the p.d.f. of a multivariate normal distribution. Let $\mathbf{G} = GMFM$, $\mathbf{Y}$ be the *transformed* GMFM in Equation (4) and $\mathbf{V}$ is the variance matrix. Under the assumption $E[Y] = \mathbf{X}\beta$ then

$$f(\mathbf{Y}|\beta, \mathbf{V}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{V}|^{\frac{1}{2}}} \exp(-\frac{1}{2}(\mathbf{Y}-\mathbf{X}\beta)^T\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\beta))$$

However we want this in terms of $\mathbf{G}$, not $\mathbf{Y}$, the transformed $\mathbf{G}$. Hence we multiply this by the Jacobian to find the p.d.f. of $\mathbf{G}$ which gives us the likelihood for the whole data. If we assume $\mathbf{V} = \sigma^2 I$ then

$$L(\beta, \sigma^{\mathbf{2}}, \lambda|\mathbf{Y}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp(-\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)) \prod_{i=1}^{n} g_i^{\lambda-1}$$

is our likelihood. This leads to the log likelihood function:

$$\mathcal{L}(\beta, \sigma^{\mathbf{2}}, \lambda|\mathbf{Y}) = \frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) + (\lambda - 1)\Sigma_{i=1}^{n} g_i.$$

Differentiating with respect to β and σ^2 leads to the estimators:

$$\hat{\beta} = (\mathbf{X^T X})^{-1}\mathbf{XY}$$

$$\sigma^2 = \frac{\mathbf{Y^T}(\mathbf{I} - \mathbf{X}(\mathbf{X^T X})^{-1}\mathbf{X^T})\mathbf{Y}}{n}.$$

We can then substitute these back into the log likelihood function and differentiate with respect to λ. This leads us to the following:

$$\frac{\partial \ln \mathcal{L}(\lambda|\mathbf{Y})}{\partial\lambda} = -\frac{\frac{\partial \mathbf{Y^T}}{\partial\lambda}(\mathbf{I} - \mathbf{X}(\mathbf{X^T X})^{-1}\mathbf{X^T})\mathbf{Y}}{\hat{\sigma^2}} + \Sigma_{i=1}^{n} \ln g_i$$

which cannot be solved directly when set to zero. Hence we use numerical searches to find the optimal λ. Newton Raphson is widely used here, but we calculate multiple values of the likelihood, setting λ to various values using the R command, *boxcox*, in the MASS library [15].

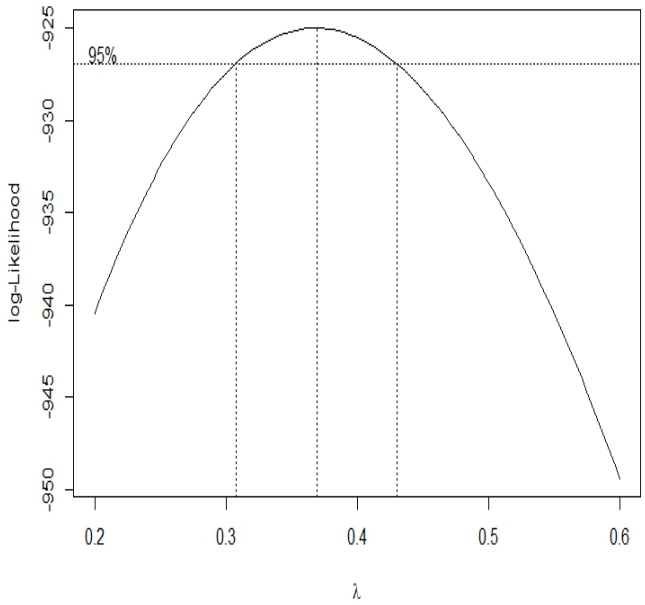


Figure 9: The log-likelihood plotted against the λ values

We then use a numerical search method to find the optimum value of λ which maximises the log-likelihood. Figure 9 shows the likelihood plotted against various values of λ within a range of $0.2 < \lambda < 0.6$. We can see that within our 95% confidence interval, we could choose an approximate value where $0.31 < \lambda < 0.42$ so that we have the best chance of transforming the GMFM measurements to be normal. We have found that the optimum value is actually $\lambda = 0.37$ $(2d.p.)$.
We can assess these values obtained for λ with Shapiro-Wilk test to see if normality in the residuals occur. Again the null hypothesis is that "the data follows a normal distribution". Looking once more at Table 4 we see that the 95% limits, alongside the optimal value of λ, all have extremely significant p-values and hence we reject any assumption of normality for the transformed data.

3.4 Conclusion

There is no requirement for Method 1 and Method 2 to give the same optimal value as they are very different methods. Method 1 is used as an assessment once the transformation has been implemented, using the formed residuals. However, Method 2 uses all of the data to contribute to the likelihood which in turn is used to estimate the best transformation. Both methods have failed to find a transformation that works well as both of their optimal transformations produce residual errors that still fail the Shapiro-Wilk test as their p-values are highly significant.
As a result, we conclude that we reject the Box-Cox transformation for all values of λ, because the optimum values of λ failed to give an appropriate transformation. Due to this, we are going to assume that we cannot transform the GMFM measurements in order make them follow a Normal distribution and will look for alternative methods to model the data.

4 Correlation

From multiple linear regression, we saw that there may be correlation in our data due to the ripples in the QQ plot. This means that the assumption that the errors were independent from each other may not be true. We even assumed non-correlated errors in Section 3.3.2, when deriving the maximum likelihood estimator for λ. We need to assess if we do indeed have the suspected correlation.

4.1 Correlation between variables

We have calculated the correlation between explanatory variables which are shown in Table 5, noting that the variable Injury Type is omitted due to it being a three-level factor. All of the variables have a relatively low correlation, which is good for our investigations as it infers we will not need to remove one of the variables as of yet. We note that some variables are positively correlated, for example Gender and Age have a correlation of 0.18. This suggests that females are relatively older. Problems arise when

variables are highly correlated as this may imply that they measure similar attributes, however the correlations are still quite low for all variables. Overall, the explanatory variables do not appear to be correlated to any worrying level.

	Time	Female	Age
Time	1.00	-0.07	0.04
Female	-0.07	1.00	0.18
Age	0.04	0.18	1.00

Table 5: Correlations between Variables

4.2 Within-patient correlation

Correlation is an important thing to consider. Our data is on the GMFM measurements which is a way to measure the activity level of the brain on the children in our study. If we think about this medically, it is likely that we will have within-patient correlation. What this means is that a measurement taken on one child may be more similar (correlated) to another measurement taken on the same child, than that taken on a different child. If there is no within-patient correlation, then we can just model all of the GMFM scores altogether. However, if there is this correlation, then we need to incorporate this into the model and allow for subject-specific effects. How can we see if this is the case this? We can study a *variogram*, but first we will explain the theory behind it. If we say that Y_1 and Y_2 are two GMFM measurements on the same child separated by time d, then:

$$\mathrm{var}(\mathrm{Y}_1 - \mathrm{Y}_2) = \mathrm{var}(\mathrm{Y}_1) + \mathrm{var}(\mathrm{Y}_2) - 2\mathrm{cov}(\mathrm{Y}_1, \mathrm{Y}_2). \tag{7}$$

Rearranging gives

$$\frac{1}{2}\mathrm{var}(\mathrm{Y}_1 - \mathrm{Y}_2) = \frac{1}{2}\mathrm{var}(\mathrm{Y}_1) + \frac{1}{2}\mathrm{var}(\mathrm{Y}_2) - \mathrm{cov}(\mathrm{Y}_1, \mathrm{Y}_2). \tag{8}$$

Now setting $\mathrm{var}(Y_1\ ) = \mathrm{var}(Y_2) = \sigma^2$

$$\frac{1}{2}\mathrm{var}(\mathrm{Y}_1 - \mathrm{Y}_2) = \sigma^2 - \mathrm{cov}(\mathrm{Y}_1, \mathrm{Y}_2). \tag{9}$$

Aside, we know that

$$\begin{aligned} \mathrm{cov}(\mathrm{Y}_1, \mathrm{Y}_2) &= \mathrm{corr}(\mathrm{Y}_1, \mathrm{Y}_2) \times \sqrt{\mathrm{var}(\mathrm{Y}_1)\mathrm{var}(\mathrm{Y}_2)} && (10)\\ &= \mathrm{corr}(\mathrm{Y}_1, \mathrm{Y}_2) \times \sigma^2 && (11)\\ &= \rho(d) \times \sigma^2, && (12) \end{aligned}$$

where $\rho(d)$ is the correlation at a separation distance d. Substituting this back into Equation (9) :

$$\frac{1}{2}\mathrm{var}(Y_1 - Y_2) = \sigma^2(1 - \rho(d)). \tag{13}$$

Equation (13) is the basis for the variogram. If we plot $\frac{1}{2}(Y_{ij} - Y_{ik})^2$ against the separation distance we can assess if we have within-patient correlation. (NB: $Y_{ij} - Y_{jk} =$

$GMFM_{ij} - GMFM_{jk}$ where we are taking all pairs of successive GMFM observations; patient i with $j^{th}$and k^{th} observations). However, what should we expect to see if we do indeed have this within-patient correlation?
Let us consider two situations using Equation (13). Firstly, imagine if we took two GMFM measurements at exactly the same time (d=0) on the same child, they would be perfectly correlated ($\rho(d) = 1$) and hence our variance would be zero. Secondly, imagine if we took two GMFM measurements on the same child, but taken very far apart in time ($d \to \infty$), it would be unlikely that they would be correlated ($\rho(d) \to 0$). Hence our variance would tend to some constant representing σ^2. This means that if we have correlation between successive observations on the same child, we might expect to see a curved slope starting at (0,0) and tending upward towards $\frac{1}{2}(GMFM_{ij} - GMFM_{jk})^2 = \sigma^2$.

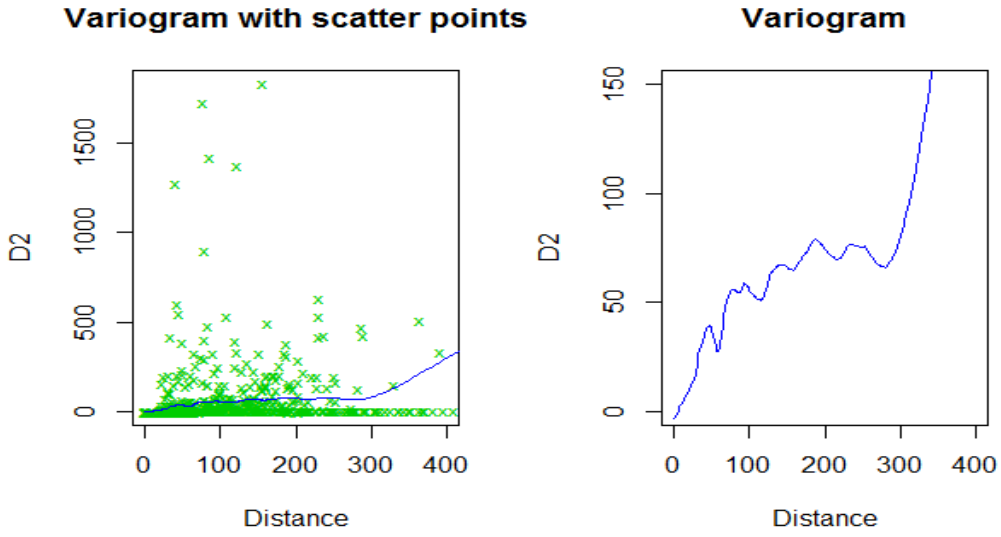


Figure 10: Plotting $D2 = \frac{1}{2}(GMFM_{ij} - GMFM_{ik})^2$ vs distance,d, where the blue line represents the smoothed curve

Figure 10 shows the plotted distances used to calculate the variogram with a blue curve. This blue curve is the variogram which is formed by using a smoothing command, *supsmu* [14], which fits Friedman's supersmoother developed by Friedman (1982) [6]. A smoother works on the principal of removing variation whilst still capturing the characteristics of the data, generally scatter plot data. A smoother works on bivariate data to create a model shown below:

$$y_i = s(x_i) + r_i$$

for $i = 1, \ldots, n$. $S(.)$ is the smoothing function with r_i being the residuals. The smoothing fuction removes some of the variability via removing some of the residuals. The smoother is taken to be "to be the average of the responses for those observations with predictor values in a neighborhood N of x_i ". The span defines how many points in the neighbourhood of the point to be predicted is used in the estimation of it. The *supsmu* R command [14] chooses the best of three smoothers and uses cross validation to pick

the best one. Cross validation is the prediction of a model which leaves the last point out in order to assess the fit of the prediction.
Referring back to Figure 10, we have two plots. The first showing the distances plotted alongside the smoother, and the second being only the smoother on a scaled axis to assess the trend better. The plotted variogram suggests that we do indeed have within-patient correlation. It suggests that $\sigma^2 \approx 70$ as this is where our graph tends to. The sharp upward turn after around $d = 300$ is of little concern as it is most likely due to outliers near the larger d values. It suggests that we should maybe incorporate more GMFM measurements that have a larger separation distance.

4.3 Conclusion

To conclude, there is correlation on successive observations taken on the same child. This needs to be incorporated and allowed for within our model. The model needs to allow for within-patient correlation, but also to allow us to look at the separate trajectories for each child. Variation due to each child needs to be assessed.

5 General Estimating Equations

We have found in Chapters 3 and 4 that we have two problems to deal with: we have non-normality and we have within-patient correlation on successive observations in our data. It is clear that multiple linear regression is not an option for us meaning our $\hat{\beta}$ given in Equation (2) is not appropriate. As a result, this chapter will explore one possible model based on the General Estimating Equation method, which we will refer to as GEE, to overcome these two issues.
GEE allows the assumption of non-normality alongside allowing for different within-patient correlation structures. Some of the commonly used structures are: the independence working assumption (IWA), exchangeable, unstructured and AR(M) structure. These will be discussed in Section 5.5.

5.1 A new $\hat{\beta}$ estimator

5.1.1 Derivation

The GEE model is based upon using a different estimating equation to that of Equation (2). This estimate was derived from $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, with $\epsilon \sim \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I})$. If we take $\mathbf{Y} \sim \mathbf{N}(\mathbf{X}\beta, \mathbf{V})$ where $\mathbf{V}$ is the variance matrix a new estimator can be formed.
We begin with analysing the multivariate normal distribution, where $\mathbf{Y}$ represents the GMFM scores and $\mathbf{X}$ represents a vector of the explanatory variables. The density is

$$f(\mathbf{Y}|\beta, \mathbf{V}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{V}|^{\frac{1}{2}}} exp(-\frac{1}{2}(\mathbf{Y} - \mathbf{X}\beta)^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\beta)).$$

Following on from this we can differentiate the log likelihood with respect to β to get

$$\frac{\partial log(L(\mathbf{Y}|\beta,\mathbf{V}))}{\partial \beta} = \mathbf{X^T V^{-1}}(\mathbf{Y} - \mathbf{X}\beta).$$

Solving for β leads to

$$\mathbf{X^T V^{-1}}(\mathbf{Y} - \mathbf{X}\beta) = 0. \tag{14}$$

Now although this equation, Equation (14), has been derived from a multivariate normal distribution, we could use it as a starting point without assuming normality. Hence if we derive a $\hat{\beta}$ estimator from Equation (14), then that too doesn't depend upon normality. When rearranged, it gives:

$$\hat{\beta} = (\mathbf{X^T V^{-1} X})^{-1}(\mathbf{X^T V^{-1} Y}). \tag{15}$$

This new $\hat{\beta}$ estimator now differs by incorporating variance structure. This means that we can change the correlation structure in the variance matrix.
Equation (14) is very important and it is this that is known as the *General Estimating Equation* (GEE). The GEE is unbiased as when taking expectation we see that the

$$E(\mathbf{Y} - \mathbf{X}\beta) = E(\mathbf{Y}) - E(\mathbf{X}\beta) = 0,$$

using $E(\mathbf{Y}) = \mathbf{X}\beta$. In addition

$$E(\hat{\beta}) = (\mathbf{X^T V^{-1} X})^{-1}\mathbf{X^T V^{-1}} E(\mathbf{Y}) = (\mathbf{X^T V^{-1} X})^{-1}\mathbf{X^T V^{-1} X}\beta) = \beta.$$

As a result, our estimator given in Equation (15) is unbiased.

5.1.2 Block diagonal variance matrix

One of the things that we would like to see the GEE incorporate is within-patient correlation. Setting $\mathbf{V}$ to be block diagonal, means that there is no correlation between patients but allows for correlation between measurements taken on the same patient. This is how Chapter 4 suggested our data was structured. Using this structure, we can re-write Equation (14) in a different form.
First we will start with Equation (14) re-written as:

$$0 = \begin{pmatrix} x_1^T & x_2^T & \dots & x_n^T \end{pmatrix} \begin{pmatrix} v_1^{-1} & 0 & \dots \\ 0 & v_2^{-1} & \dots \\ \vdots & \vdots & \ddots \end{pmatrix} \left(\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \right),$$

where x_i is a explanatory variable matrix for patient i (i.e the time, age, injury type, gender for each measurement on child i); v_i is the variance matrix for patient i; y_i is the

vector of the GMFM measurements for patient i and β_j is the parameter coefficients. This would then give:

$$0 = \begin{pmatrix} x_1^T v_1^{-1} & x_2^T v_2^{-1} & \dots & x_n^T v_n^{-1} \end{pmatrix} \left(\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \right).$$

Multiplying out gives:

$$0 = \begin{pmatrix} x_1^T v_1^{-1} y_1 + x_2^T v_2^{-1} y_2 + \dots & x_n^T v_n^{-1} y_n \end{pmatrix} - \begin{pmatrix} x_1^T v_1^{-1} x_1 \beta_1 + x_2^T v_2^{-1} x_2 \beta_1 + \dots & x_n^T v_n^{-1} x_n \beta_n \end{pmatrix},$$

and then swapping terms around gives

$$0 = \begin{pmatrix} x_1^T v_1^{-1} y_1 - x_1^T v_1^{-1} x_1 \beta_1 \end{pmatrix} + \begin{pmatrix} x_2^T v_2^{-1} y_2 - x_2^T v_2^{-1} x_2 \beta_1 \end{pmatrix} + \ldots + \begin{pmatrix} x_n^T v_n^{-1} y_n - x_n^T v_n^{-1} x_n \beta_n \end{pmatrix}.$$

Which can be finally written as:

$$0 = \sum_i \mathbf{x_i^T V_i}^{-1} (\mathbf{y_i} - \mathbf{x_i}\beta). \tag{16}$$

This means that Equation (14) can be written as Equation (16) when $\mathbf{V}$ is block diagonal. Now the formulation of Equation (16) means that we can analyse each GMFM trajectory for individual children separately and then sum them together for the overall model. This may be appropriate for our model because we suggested in Section 2.3.1 that we have a highly varied intercept for each child and so we may need to incorporate separate intercepts for each child. This will be easier to analyse with individual child trajectories. Using this block diagonal structure assumes that we know the value for $\mathbf{V}$, but in reality we do not know this. We have chosen to assume a value of our choice to solve this. In practice we can choose a structure for $\mathbf{V}$ that is appropriate to our data. The choosing of this is discussed in Section 5.4.
Hence, we can use a new, unbiased $\hat{\beta}$ estimator. This allows non-normality and the choice of a different correlation structure.

5.2 Finding the GEE based model

We can run a similar process to multiple linear regression with our GEE estimator, but there are a couple of differences that affect the analysis of the model. First, as there is no assumption of normality we cannot calculate the maximum likelihood estimates here as we do not know the distribution. Therefore we cannot compare models using the likelihood methods such as Akaike information criteria or Bayesian information criterion and so we have to look for an alternate. Non-normally, and misspecified correlation structures, also mean that the standard variance estimator for $\hat{\beta}$ may not be appropriate.

Instead, a robust version can be obtained. Hence the model output will contain two versions of standard errors based upon the naive variance (S^{-1}). One being

$$\mathrm{Var}_{\mathrm{Naive}}(\hat{\beta}) = \left(\Sigma_i \mathbf{x_i^T \hat{V}_i}^{-1} \mathbf{x_i}\right)^{-1} = S^{-1} \tag{17}$$

and the other being

$$\text{Var}_{\text{Robust}}(\hat{\beta}) = S^{-1}HS^{-1}, \tag{18}$$

where

$$H = \Sigma_i \mathbf{x_i^T} \hat{\mathbf{V}}_\mathbf{i}^{-1} (\mathbf{y_i} - \mathbf{X_i}\hat{\beta})(\mathbf{y_i} - \mathbf{X_i}\hat{\beta})^T \hat{\mathbf{V}}_\mathbf{i}^{-1} \mathbf{x_i^T}.$$

We must note that if our chosen correlation structure, defined in $\hat{\mathbf{V}}_\mathbf{i}$, is the true correlation for our data then $\hat{\mathbf{V}}_\mathbf{i} = (\mathbf{y_i} - \mathbf{X_i}\hat{\beta})(\mathbf{y_i} - \mathbf{X_i}\hat{\beta})^T$ which causes $H = S$. This means that when the true correlation is specified, the Naive and Robust variance will be equivalent. We will look out for this later when choosing an appropriate structure.
The GEE method estimates the robust variances via an iterative method by treating the $\mathbf{V_i}$ as another parameter to approximate. A benefit of using the robust variances is that they still provide a consistent estimator for the variance, even when correlation is misspecified. Conversely, the issue with robust standard errors is that they tend to be larger and cause larger confidence intervals. Therefore our $\hat{\beta}$ estimates tend to be less certain.

5.3 GEE model assuming IWA

We first choose to substitute $\mathbf{V_i} = \mathbf{I_i}$ into Equation (16). This still gives us an unbiased estimate because when taking the expectation with $\mathbf{I_i}$ substituted into Equation (14), we expect zero. Using the identity matrix, $\mathbf{I_i}$, is known as the *Independent Working Assumption (IWA)*. Note that this will give us the same structure as in multiple linear regression with no within-patient correlation. However, we do this to get a feel for the data based upon using GEE. When using the IWA, we should expect to see the robust standard errors being larger, especially because we do not believe IWA is the best correlation structure as it does not incorporate the suspected within-patient correlation. Modelling our paediatric data, we carry out the regression analysis and analyse the model containing the covariates: Time3 ($\sqrt{Time}$), TBI, HYPOXIC, Gender and Age. The baseline here is a male patient, in the middle age group who has sustained an ABI injury. (*Note that we also analysed Time and Agegroup in place of Time3 and Age, but the results showed non-significance anyway*).
We see the following regression output:

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	40.11	3.70	10.85	6.03	6.65
Time3	0.347	0.62	0.56	0.68	0.51
HYPOXIC	-39.71	3.88	-10.25	6.17	-6.43
TBI	-8.10	3.35	-2.44	7.44	-1.10
Gender	7.71	3.14	2.46	6.27	1.23
Older	5.00	3.46	1.44	7.99	0.63
Younger	5.07	3.52	1.44	5.17	0.98

Output 3: Initial GEE regression output

Output (3) shows that, as expected, the robust standard errors are indeed larger than the naive standard errors, hence the 95% confidence interval for our β estimates

are wider than with the naive model. The β estimates are fairly similar to those given in Output (1), however we have larger variation.
After removing non-significant explanatory variables, we obtain a predictive model that only includes the injury types. Output 4 shows that HYPOXIC is significantly different from ABI, with a β estimate of -41.65 This means that HYPOXIC is over 40 GMFM scores lower than ABI on average. TBI is not that significant. It also has a lowering effect on the GMFM measurements, but less so than the effect of HYPOXIC. This agrees with our earlier judgments from Figure 6.

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	47.51	2.07	22.92	5.13	9.26
HYPOXIC	-41.65	3.56	-11.69	5.71	-7.29
TBI	-8.09	3.31	-2.44	6.96	-1.16

Output 4: Final GEE regression output

Our common intercept is 47.51%, with a high robust variance of 5.13^2. This large variance is expected due to the vast range of GMFM scores each patient began with. It is interesting to note that the time since injury, or the age has not had a significant effect on the brain functioning capability which seems to go against the medical intuition.

5.4 Possible correlation structures

We said that one of the main benefits of modelling with the GEE is that we can change our variance matrix without making the coefficient estimations biased. We previously looked at the IWA, but what about alternate correlations? Four of the main structures that are commonly used are:

- independence working assumption (IWA)
- exchangeable
- unstructured
- autoregressive, AR(M)

We keep the block diagonal form ($\mathbf{V_i}$ for $i = 1, \dots n$ down the diagonal, zero's everywhere else), but change the construction of $\mathbf{V_i}$ where $\mathbf{V_i}$ is the variance matrix for patient i. The following variance matrices are the correlation structure for patient i assuming (a) IWA (b) exchangeable (c) unstructured and (d) Autoregressive.

$$(a)\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & & \ddots & \vdots & \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} \quad (b)\begin{bmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \rho & \rho & 1 & \dots & \rho \\ \vdots & & \ddots & \vdots & \\ \rho & \rho & \rho & \dots & 1 \end{bmatrix} \quad (c)\begin{bmatrix} 1 & \rho_{1,2} & \rho_{1,3} & \dots & \rho_{1,m} \\ \rho_{1,2} & 1 & \rho_{2,3} & \dots & \rho_{2,m} \\ \rho_{1,3} & \rho_{2,3} & 1 & \dots & \rho_{3,m} \\ \vdots & & \ddots & \vdots & \\ \rho_{1,m} & \rho_{2,m} & \rho_{3,m} & \dots & 1 \end{bmatrix}$$

$$(d)\begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{|t_1-t_m|} \\ \rho & 1 & \rho & \dots & \rho^{|t_2-t_m|} \\ \rho^2 & \rho & 1 & \dots & \rho^{|t_3-t_m|} \\ \vdots & & \ddots & \vdots & \\ \rho^{|t_n-t_1|} & \rho^{|t_n-t_2|} & \rho^{|t_n-t_3|} & \dots & 1 \end{bmatrix}$$

IWA is equivalent to suggesting that there is no correlation between any measurements taken on child i. The exchangeable correlation structure is also known as compound symmetry or sphericity. This allows for any two GMFM measurements on patient i to have the same constant correlation, ρ. The unstructured correlation structure is different again. It allows for a different correlation for any two measurements taken on child i. However this allows for high correlation between two measurements taken close together in time, which decreases with respect to increasing time distance apart.
The autocorrelation structure allows for decreasing correlation as the time period between two observations increases and, out of all four correlations, our intuition tells us that the autoregressive structure would be most suitable to match the conclusions given the analysis the variogram. Unfortunately we will not be considering the AR(M) model due to some patients only having one measurement taken on them which causes non-convergence in the calculation of AR(M) correlations.
We expect that IWA is inappropriate for our data as we believe we need within-patient correlation. The unstructured correlation is a little too unrestricted for our liking and so we think that the exchangeable correlation may be most suitable here.

5.5 GMFM correlation structure

The results from fitting a GEE based model with IWA, exchangeable and unstructured correlation lead us to believe that two different models were most appropriate to fit the data dependent on the structure chosen.
As we have seen, IWA gave a model that only included the injury types in Output 4. Outputs 5 and 6 show the output for the best model suited to the data with correlations following an exchangeable and unstructured pattern respectively. Both of these gave the same model where we find that alongside the injuries being significant, time and an interaction term between these are now significant. (Note that TBI is not significantly different from the ABI group, but as a whole the Injury Type factor is significant and as it is included inside this factor, TBI remains in the model).

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	44.74	3.95	11.34	3.27	13.69
HYPOXIC	-27.86	8.28	-3.36	8.54	-3.26
TBI	-10.42	6.37	-1.64	7.02	-1.48
Time3	1.07	0.24	4.52	0.20	5.31
HYPOXIC:Time3	-1.16	0.37	-3.13	0.25	-4.56
TBI:Time3	0.18	0.34	0.53	0.36	0.50

Output 5: Best GEE model with exchangeable correlation structure output

The exchangeable and unstructured model both give similar estimates, with exchangeable being generally higher except for the HYPOXIC term. Comparing these with the IWA estimates, we see that the intercept estimate is higher still and with HYPOXIC estimate being -41.51 rather than just under 30 for the exchangeable and unstructured estimates. This is to be expected as by incorporating correlation, we are reducing some unknown information about the model i.e. removing some of the error in the model (if the correlation is true). Note also that for TBI, it is the IWA and unstructured models that have similar estimates with the exchangeable estimate being higher by around a value of 2.

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	43.36	3.55	12.20	4.20	10.32
HYPOXIC	-29.74	7.40	-4.02	8.49	-3.50
TBI	-8.32	5.83	-1.43	6.90	-1.21
Time3	0.73	0.34	2.16	0.26	2.78
HYPOXIC:Time3	-1.37	0.60	-2.29	0.51	-2.65
TBI:Time3	0.08	0.48	0.16	0.43	0.18

Output 6: Best GEE model with unstructured correlation structure output

We are pleased to see that there is a significant interaction term including time because we noted in Figure 3 that we had an s-shape curve. This means that it is highly likely that the effect of a covariate, here being HYPOXIC and TBI injuries, is additive with respect to time. Here in Outputs 5 and 6 we see that this is the case. Hence our three proposed models are:

$$(IWA) \quad GMFM = 47.51 - 41.65 * \text{HYPOXIC} - 8.09 * \text{TBI} + \epsilon$$

$$\begin{aligned}(Exchangeable) \quad GMFM = {} & 44.74 - 27.86 * \text{HYPOXIC} - 10.42 * \text{TBI} + 1.07 * \text{Time3} \\ & -1.16 * (\text{HYPOXIC} * \text{Time3}) + 0.18 * (\text{TBI} * \text{Time3}) + \epsilon\end{aligned}$$

$$\begin{aligned}(Unstructured) \quad GMFM = {} & 43.36 - 29.74 * \text{HYPOXIC} - 8.32 * \text{TBI} + 0.73 * \text{Time3} \\ & -1.37 * (\text{HYPOXIC} * \text{Time3}) + 0.08 * (\text{TBI} * \text{Time3}) + \epsilon\end{aligned}$$

As we can see they differ from each other so how can we tell which one is most appropriate? For likelihood-based methods, we would use something called the AIC or BIC discussed in Chapter 6. However "for non likelihood- based methods, such as GEE, there is a lack of literature on model selection"Pan (2001) [12].
We look at a different approach by comparing the variances of the $\hat{\beta}$ estimates. We said that when $\hat{\mathbf{V}}_\mathbf{i} = (\mathbf{y_i} - \mathbf{X_i}\hat{\beta})(\mathbf{y_i} - \mathbf{X_i}\hat{\beta})^T$ then $H = S$ and hence Equation 18 became Equation (17) i.e. the robust variance became the Naive variance. This occurs when the true correlation has been specified. Hence, to compare models with the *s*ame parameters but different structures, we look for the smallest value of total of (naive variance - robust variance).
If we compare two models with different numbers of parameters in this way, it doesn't

account for the number of parameters in the model and so this method is only appropriate to compare the same models but with different correlation structures.
We have two sets of potential model parameters, and three possible correlations. Hence we will find the best correlation structure for a model (a) only containing HYPOXIC and TBI and (b) containing HYPOXIC, TBI, Time3 and HYPOXIC*Time3. Table 6 shows the results. We see that for both models it is clearly evident that the closest to the true correlation structure is exchangeable particularly in model (a).

	IWA	Exchangeable	Unstructured
(a)	2284.04	270.26	1309.95
(b)	351.33	119.34	507.56

Table 6: GEE model comparisons based upon the difference between the naive and robust variances

The exchangeable correlation is the best one with either set of parameters. The combination of Injury Type, Time3 and Time3*Injury Type and exchangeable correlation has the smallest difference between Robust and Naive variances out of all combinations in Table 6 and so we conclude that this is the best GEE model. We note this model predicts a the working correlation with $\rho = 0.776$. This means that observations on the same individual have quite a high correlation as expected.

6 Random Effects Modelling

6.1 Introduction

The models that have been looked at in previous sections consider population average interpretations of the sample data. Our next step is to include within-patient error alongside large between-patient differences. We will do this by incorporating a patient-specific error (random effect) alongside the previous parameters (fixed effects).
Note that by incorporating patient-specific error, we eliminate some of the potential recall bias. This is where patients have already sustained an injury and are more likely to have the injury again. For instance, if they had previously sustained a head injury, and then a second injury occurred, they may be more likely to have worse GMFM scores then the average population and they would bias the results. However, patient-specific error helps eliminate this.
In this section we will address this issue by looking at a mixed effects linear model by incorporating random effects alongside the fixed effects.

6.2 Random effects

Random effects modelling is based on the idea that the parameters, β will be different for each patient. The effects that cause this, the random effects, are thought to be unobservable, i.e. we cannot measure the actual value of these effects. We can think of

this as a conditional distribution:

$$f(y) = \int_u f(y|u)g(u)du$$

Here the random effect distribution is denoted by $g(u)$, and our conditional data distribution is denoted by $f(y|u)$. For our data $f(y|u)$ are the GMFM measurements given a patient-specific effect. The general model that we will consider will be

$$GMFM_i = X_i\beta + Z_iU_i + \epsilon_i$$

where

- X_i is a matrix of the covariates possibly including Time3, Injury Type, Female and Age group.
- β is the vector of parameters
- Z_i is the matrix of the random effect covariates
- U_i is a vector of the unobserved, subject specific random effect parameter estimates of Z_i, with zero mean and covariance matrix $\mathbf{V}$

The fixed effects part of this model is $X_i\beta$ and the random effects part is denoted by Z_iU_i. The difference between the behaviour of β and that of U_i is that β is the coefficient value on average, which is constant with respect to the patient. However, U_i is a value of a coefficient that is specific for a particular subject and remains constant for observations on the same individual, but changes when a measurement is taken on a different individual.
When there are both fixed and random effects, the model is known as a mixed effects model. In our case, a possible model for patient i is a so-called intercept and slope model:

$$\begin{aligned} GMFM_i &= \beta_0 + \beta_1 * Time3_i + \beta_2 * Female_i + \beta_3 * TBI_i \\ &\quad +\beta_4 * HYPOXIC_i + \beta_5 * YoungGroup_i + \beta_6 * OlderGroup_i \\ &\quad +U_{i0} + U_{i1} * Time3_i + \epsilon_i \end{aligned} \tag{19}$$

Here the baseline is a male in the Medium Age group who has sustained an ABI injury. We are most interested in including a random intercept and a random slope, which will be seen in Section 6.4.2. Equation (19) assumes that one of the covariates is Time3, which allows for a general linear trend. The subject specific trends are incorporated into the model through the random effects. The random effects intercept allows for the patient to have a different mean level whereas the random effects Time3 covariate allows for a different slope because it allows each patient's recovery rate to differ.
We will assume $U_{i0} \sim N(0, \sigma_0^2)$ and $U_{i1} \sim N(0, \sigma_0^2)$ such that $Cov(U_{i0}, U_{i1}) = \sigma_{01}$. This means that the variance and covariance of two measurements on child i is defined as:

$$Var(Y_{ij}) = \sigma_0^2 + \sigma_1^2 t_{ij}^2 + 2\sigma_{01}t_{ij} + \sigma^2$$

$$Cov(Y_{ij}, Y_{ik}) = \sigma_0^2 + \sigma_1^2 t_{ij}t_{ik} + \sigma_{01}(t_{ij} + t_{ik})$$

where σ^2 is equivalent to var(ϵ_i).

6.3 What should we expect?

Figure 12 (b), shows that there is quite a lot of variety in the trajectories for each patient. The intercepts are highly varied with no obvious consistent value on the graph. For example, the range of the initial GMFM measurements is from 0.1% to 99.9% which is quite a large! Where does this variation come from? Looking at Figure (11), we see that there is a large variation in the initial GMFM measurements within ABI and TBI patients. For instance, for ABI injured patients have initial measurements (intercept) that range from 0.1% to 80.93%. We also note that their is a lot of variation in each of the age groups and both genders. This suggests that as large variation in the initial measurements can be seen in all of variables, it may be due to another cause; a patient specific effect. This suggests that we may need to include a random intercept.

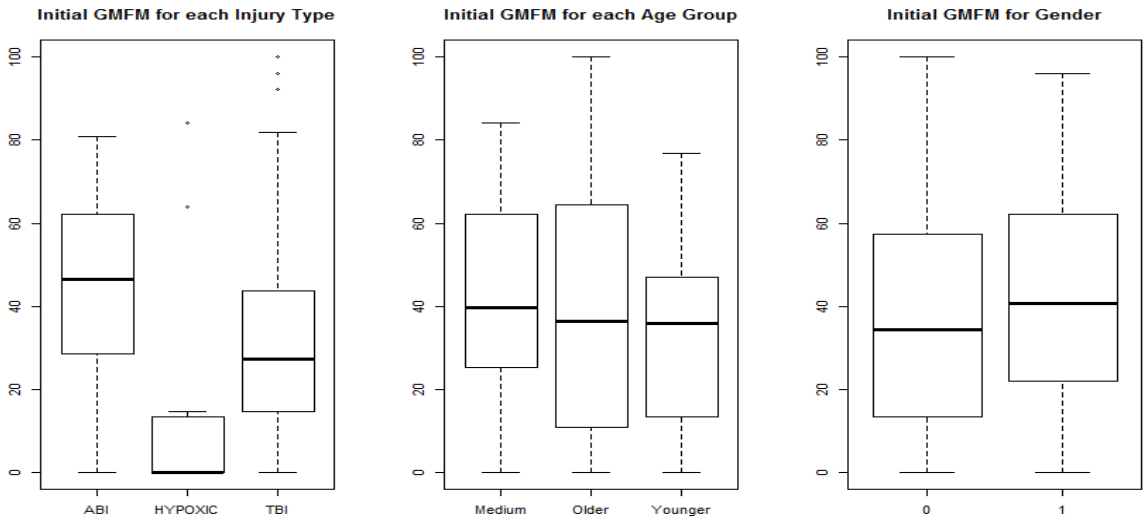


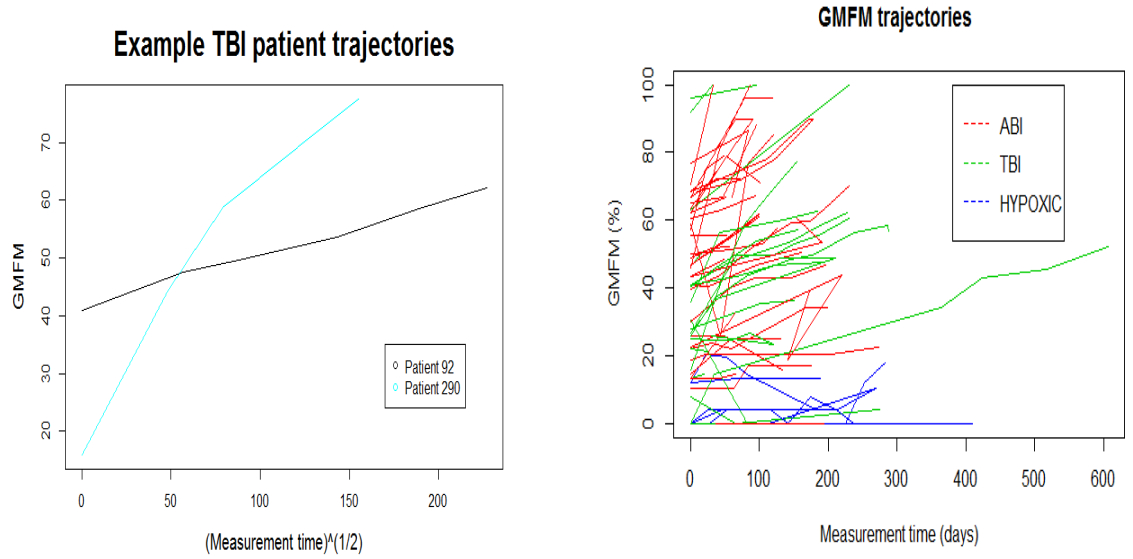
Figure 11: Boxplots for the initial GMFM measurements, taken at Time= 0, with respect to the different factors. Note that for the Gender boxplot, 0 refers to males and 1 refers to females.

It is expected that different children will recover, increasing their brain activity at different rates, dependent on the type of injury that they sustained. Moreover, it is likely that each patient will have different recovery rates over time compared to the average rate with some recovering faster than others, and vice versa. We can see this illustrated by Figure 12 (a).

We have two patients who are both male, of similar age and have both have sustained a TBI injury. It is clear that they both have different slopes in their trajectories, and so have different recovery rates. A patient-specific slope is likely to be needed, and we consider including a random effect slope with respect to time in our model.

We contemplate what covariates may be important in our model in the fixed effects part. From the GEE results, we found that only the type of injury that the patient had was significant, and hence we presume this will be needed in our fixed effects part of the model. We may or may not require gender in the model as well; it was significant in the multiple linear regression, but not very significant when incorporating robust standard

errors with the GEE estimator.
Therefore we expect that a patient-specific intercepts and slopes will be significant in our mixed effects models. We think that the model should contain similar fixed effects to those included in the GEE models.



(a) Patient 92 is a 16.8 year old TBI male; Patient 290 is a 15.8 year old TBI male

(b) GMFM trajectories for all patients

Figure 12: GMFM trajectories for (a) Two typical TBI patients and (b) All GMFM data

6.4 Results

We begin by looking at a model as in Equation (19), which incorporates the possible fixed effects, but also consider a possible intercept and time as part of the random effects. Note that we also looked into interaction terms in the fixed effects, but they were found to be non-significant and will not be discussed further here.

6.4.1 Finding the best model

We compare models using three measures:

1. The log likelihood of the model (and restricted log likelihood REML)
2. The Akaike information criterion (AIC)
3. The Bayesian information criterion (BIC)

We explore fitting a mixed effects model through the maximisation of the log likelihood. We must begin with a full model, or at least include all possible fixed variables (remembering to back-test). We first assess the appropriate random effect parameters

using the likelihood ratio test. Once the random effects have been selected, we then fit the model with these and analyse the significant fixed effects. We seek the best maximised likelihood. We have look at using the the REML method. The reason for this is that using the maximum likelihood (ML) method is biased for random effects due to it ignoring the fact that a random intercept and random slope being estimated [17]. The issue with the maximum likelihood method is that these are within the parameter estimates and so REML uses the same estimation process, but removes the parameter estimator from the equation via multiplying by orthogonal matrices. This leads to an unbiased estimator. However, the REML method should be used mainly to compare the random effects part of the model. The ML method is most suited to choosing the fixed effects.

Note that when there is a relatively small number of variables in the model compared to the number of observations, the estimates from both ML and REML should be similar. This is the case with our data and so we explain our analysis using the ML method. It gives very similar findings to the REML method. (*Note that the REML estimates have been omitted, but conclusions used*).

A problem with using the log likelihoods as a measure of fit is that it will increase just from the act of adding parameters, even if they are not significant. Two ways to solve this problem are to use AIC and BIC which incur a penalty for the number of parameters in the model.

AIC is calculated by $AIC = -2 * \hat{L}(\hat{\theta}) + 2 * K$, where $\hat{L}(\hat{\theta})$ is the maximised log likelihood and that K is the number of parameters. BIC is calculated by $BIC = -2 * \hat{L}(\hat{\theta}) + K * log(n)$, with n being the number of data points observed.

We compare the fit of different models by looking for the lowest AIC or BIC values. Both of the REML and ML methods lead to the same conclusion in our case. However take care with the choice of REML and ML methods in AIC and BIC comparison. This is because the AIC and BIC calculated through REML are not comparable with the AIC and BIC calculated through ML [17]. Hence, we cannot compare a model's AIC calculated with REML with another model's AIC calculated with ML.

There is some controversy over whether using AIC or BIC is better. AIC selects the optimal model with the least mean squared error acting upon the assumption that the TRUE model is not an option [3].This makes sense because there may be other variables that affect a child's brain activity rate that have not been measured and so it is possible that we do not have the TRUE model as a possibility. However BIC assumes that the TRUE model is a possibility, but also penalizes more than the AIC for the number of parameters used in the model.

AIC and BIC both give good estimates in finding the most appropriate model, although under different assumptions. Kuha (2004) [9] states "they will identify good models for observed data, but both criteria can still fail in this respect, ... an approach of using the two criteria together (as well as significance tests and perhaps other indices of model fit) has been advocated here."

It is from this that we will draw our conclusions using the log likelihood, AIC and BIC including t-tests on the fixed effect parameters to pick the optimal model. Note that we ideally want to see the largest log likelihood, and smallest AIC and BIC scores.

6.4.2 Model results

Table 7 shows results for a variety of models that have been considered. Models 1-3 include all fixed effect covariates. The best log likelihood occurs when both random effects are included in the model (Model 3) which is in agreement with the AIC and BIC values. For example, if we only included a random slope associated with time, the AIC (2538.664) and BIC (2571.000) scores are the worst out of all models in Table 7, and then when we include the random intercept term, we see a dramatic improvement of 369.301 in the AIC and 361.383 BIC values (looking at model 3). There is a similar effect when comparing model 2 (only including the random intercept) and model 3. Hence we conclude that both the random intercept and slope need to be included.
Models 3-5 compare the best fit after removing insignificant variables. The Gender and Age Group factors were found to be insignificant, whereas the Time Since Injury and Type of Injury where highly significant. According to the log likelihood we should choose Model 3 as it has the smallest value, however this doesn't account for the number of variables included in the model. More variables included in the model will always improve the log likelihood, even if a variable is insignificant, as it will account for some, however small, variation in the residuals. We see that the AIC and BIC values are in agreement in choosing Model 5 with AIC improving by 5.704 and BIC by 16.682 as the most suitable choice. We note that the loglikelihood has only changed slightly by 0.147 for the worse by removing the insignificant variables. Therefore, we agree with AIC and BIC and choose Model 5.

	β_0	β_1	β_2	β_3	β_4	β_5	β_6	$U_{i,0}$	$U_{i,1}$	ϵ	log likelihood	AIC	BIC
1	✓	✓	✓	✓	✓	✓	✓		✓	✓	-1260.03	2538.66	2571.00
2	✓	✓	✓	✓	✓	✓	✓	✓		✓	-1127.35	2272.70	2305.64
3	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**-1073.68**	2169.36	2209.61
4	✓	✓		✓	✓	✓	✓	✓	✓	✓	-1073.68	2167.37	2203.97
5	✓	✓		✓	✓			✓	✓	✓	-1073.83	**2163.66**	**2192.94**
6	✓	✓		✓	✓				✓	✓	-1260.58	2533.15	2555.11
7	✓	✓		✓	✓			✓		✓	-1127.77	2267.55	2289.50

Table 7: Comparisons between various mixed effect models. The blue variables represent the fixed effects, and the purple variables represent the random effects. The bold values highlight the best model. The coefficients $\beta_0, \ldots, \beta_6$ correspond to the fixed effect terms in Equation (19) with $U_{i,0}$ and $U_{i,1}$ corresponding to a random intercept and slope respectively.

Models 6 and 7 compares the effect of removing one of the random effects from the model, after removing the insignificant fixed variables. It is clear that these are worse models. Other models that have been analysed were: including interaction terms in the fixed effects; replacing the factor Age Group with the actual Ages and modelling only a random effects model. None of these proved to be a better fit and so we conclude that model 5 is the best solution in Equation 20 we can see model 5 with the β parameter

estimates.

$$\begin{aligned} GMFM &= 43.153 + 1.138 * \text{Time3} - 7.796 * \text{TBI} + -21.649 * \text{HYPOXIC} \\ &\quad + U_0 + U_1 * \text{Time3} + \epsilon \end{aligned} \tag{20}$$

Output 7 shows the regression output for the model shown in Equation (20). We see that all of the variables are highly significant apart from TBI as it appears that TBI doesn't significantly differ from ABI injuries, unlike HYPOXIC. This makes sense as, medically speaking, HYPOXIC is a different type of injury to the other two. HYPOXIC has an estimated coefficient of -21.65 which means that a HYPOXIC patient will have a GMFM score lower, than ABI, by this value on average. This agrees with the trajectories shown in Figure 3 (d).
We see that HYPOXIC and TBI both have standard errors similar to the robust versions in the GEE models with exchangeable and unstructured correlation. The estimates of these and Time 3 follow similar trends, although the random effects estimates are slightly lower here.
The average affect of Time3 is estimated to be 1.14, with a small standard error. However we see that the patient-specific effect of Time 3 has a standard deviation of 1.21 and so the recovery rate differs depending on the patient. The average initial GMFM measured is given by the estimated intercept which is 43.15, however the patient-specific initial measurement has a large standard deviation of 24.36. This looks to be about right when analysing Figure 12 (b) as, like aforementioned, there is a large range in initial measurements. This tells us that each patient can start with a large range of GMFM scores.

```
Random effects:
 Formula: ~1 + time3 | id
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev   Corr
(Intercept) 24.56   (Intr)
Time3        1.21    0.42
Residual     4.85

Fixed effects: GMAE ~ injtype + time3
             Value  Std.Error  DF  t-value p-value
(Intercept)  43.15    3.95     212   10.92  0.00
HYPOXIC     -21.65    8.18      71   -2.65  0.01
TBI          -7.80    6.25      71   -1.25  0.22
Time3         1.14    0.17     212    6.52  0.00
 Correlation:
               (Intr)  HYPOX    TBI
HYPOXIC        -0.462
TBI            -0.604  0.294
Time3           0.197  0.013  0.017
```

Output 7: Final Mixed Effects Model regression output

We see that the correlation between the random effects is quite prominent and suggests that a patient starting with a higher GMFM score will generally recover faster (increased slope) because we have a positive correlation of 0.42. We can also see the correlation between the fixed effects. The main ones to point out here is that the types of injury are negatively correlated with the intercept. We accept this as an appropriate mixed effects model.

7 Misspecification

Misspecification is always a worry for statisticians. We want to explore the effect of misspecification in GEE and mixed effect modelling on our data. The reason we do this is to see what we should expect if we have misspecified our model. One of the ways in which we can compare the different models and their effect is to fit the different models to a new set of data with similar characteristics as our original data.
We believe that the best model so far is the mixed effects model given in Equation (20) with a random intercept and random slope. We simulate 1000 data sets based on having similar characteristics defined by this model. We can compare the estimates and standard errors given from modelling these data sets to those given by the model in Equation (20) to assess the effectiveness of the fit for different models.
We estimate $\hat{\beta}_0 : \hat{\beta}_3$ (which is for the fixed Intercept, HYPOXIC, TBI and Time3 respectively) with their standard errors 1000 times. In addition, for the GEE model we have Robust standard errors, and in addition for the random effects models we have the standard deviations associated with a random intercept (σ_0) or slope with respect to time (σ_1) (or both depending on the model) with the correlation between them and a residual error (the variability not associated to either of the random effects σ_ϵ) for each of the 1000 simulated data sets.
In Output 8 you can see the results when finally three GEE models with IWA (like multiple linear regression), exchangeable or unstructured correlation and the mixed effects models incorporating a random intercept, a random slope or both (the latter being the true model). The mean estimate is the average of the estimated $\hat{\beta}_j$ for $j = 0, 1, 2, 3$. We can compare this against the true value which is defined by the characteristics we simulated the data for.

GEE with independence working assumption

	β_0	β_1	β_2	β_3
True	43.152	-21.675	-7.796	1.138
Mean estimate	42.952	-21.968	-7.361	1.141
Empirical SD	5.579	13.695	9.372	0.468
Mean naive estimated SE	3.316	4.541	4.134	0.317
Mean robust estimated SE	5.252	11.975	8.651	0.415
Coverage with naive SE	0.897	0.743	0.791	0.896
Coverage with robust SE	0.959	0.943	0.953	0.947

GEE with exchangeable correlation structure

	β_0	β_1	β_2	β_3
True	43.152	-21.675	-7.796	1.138
Mean estimate	43.021	-22.086	-7.447	1.142
Empirical SD	4.414	10.048	7.649	0.225
Mean naive estimated SE	4.513	9.445	7.183	0.148
Mean robust estimated SE	4.326	9.449	7.297	0.216
Coverage with naive SE	0.984	0.968	0.963	0.877
Coverage with robust SE	0.969	0.962	0.959	0.964

GEE with unstructured correlation

	β_0	β_1	β_2	β_3
True	43.152	-21.675	-7.796	1.138
Mean estimate	42.984	-21.828	-7.321	1.137
Empirical SD	6.280	15.151	9.584	0.515
Mean naive estimated SE	4.194	6.663	6.475	0.254
Mean robust estimated SE	16.741	20.219	11.829	3.032
Coverage with naive SE	0.947	0.796	0.918	0.959
Coverage with robust SE	0.967	0.919	0.957	0.966

Random effects with intercept and slope

	β_0	β_1	β_2	β_3
True	43.152	-21.675	-7.796	1.138
Mean estimate	43.037	-21.924	-7.603	1.144
Empirical SD	3.878	8.206	6.308	0.169
Mean estimated SE	3.829	7.883	6.030	0.171
Coverage	0.974	0.970	0.960	0.971

	σ_0	σ_1	ρ	σ_e
True	24.556	1.212	0.423	4.850
Mean estimate	24.025	1.197	0.429	4.851
Empirical SD	2.143	0.135	0.136	0.280

Random effects with intercept only

	β_0	β_1	β_2	β_3
True	43.152	-21.675	-7.796	1.138
Mean estimate	43.018	-22.098	-7.434	1.142
Empirical SD	4.410	10.047	7.611	0.224
Mean estimated SE	4.527	9.577	7.277	0.101
Coverage	0.978	0.968	0.954	0.809

	σ_0	σ_1	ρ	σ_e
True	24.556	1.212	0.423	4.85
Mean estimate	27.573	-	-	8.115
Empirical SD	2.42	-	-	0.676

Random effects with slope only

	β_0	β_1	β_2	β_3
True	43.152	-21.675	-7.796	1.138
Mean estimate	43.033	-22.001	-7.511	1.144
Empirical SD	4.165	9.301	6.866	0.235
Mean estimated SE	2.250	4.300	3.467	0.448
Coverage	0.875	0.820	0.826	0.998

	σ_0	σ_1	ρ	σ_e
True	24.556	1.212	0.423	4.85
Mean estimate	-	2.951	-	15.848
Empirical SD	-	0.302	-	1.301

Output 8: Three GEE models with different correlations alongside three random effect models with random effect parameters outputs.

Different estimators can effect the bias of the estimate, as studied in Buddelmeyer(2008) [2] upon panel data (longitudinal data) and explored the bias upon fixed effect estimates. They used the mean absolute bias (mean absolute error) and the the root mean square error to assess the degree of bias. These, however, can only be calculated to assess the error of a particular covariate, but not inter covariate bias due to it's scale dependencies. For instance, it calculates the bias/error between β_0 and $\hat{\beta}_0$. Hence we have used these to assess the bias in the fixed effects, alongside the standard error and correlation estimates for the random effects to compare our models (Calculations omitted).
The mean standard error comes from calculating the theoretical variance for that $\hat{\beta}_j$ for each of the 1000 runs, and then calculating the average over these and square rooting it. However, the empirical standard deviation measures the dispersion of the 1000 parameter estimates. We want the mean standard error to be as close to the empirical SD as possible.
Another way to assess the fit of the model is to look at coverage. For each data set, we get the standard deviation for, say $\hat{\beta}_j$ and an estimate for it, which means that we can calculate a confidence interval. It would be useful to see if, for each of the 1000 data sets, the true β_j falls within the confidence interval. Coverage looks at how many times this occurs. We would ideally like the coverage to be around 95% but good coverage is estimated to be within (93.6% , 96.4%).
This is calculated by looking at 95% confidence intervals. Lets say that N is the number of times that a 100% confidence interval includes the true value, for say β_j. For 1000 simulations, this follows a binomial distribution $\mathrm{N} \sim (1000, \mathrm{p})$ where p is the probability for this occurring, hence for a 95% confidence interval we set p=0.95. Now as coverage is a proportion, we want to analyse $\frac{\mathrm{N}}{1000}$ which will vary according to

$$\mathrm{var}\left(\frac{\mathrm{N}}{1000}\right) = \frac{\mathrm{p} * (1 - \mathrm{p})}{1000} = 0.00689^2$$

Hence if we want our true value to be within the calculated confidence interval, for each

of the 1000 models, for 95% of the time then coverage should be $0.95 \pm 1.96 * 0.00689 = (0.936, 0.964)$. Consequently a good fitting model will have:

1. the estimated values to be similar to the true values (little to no bias)
2. the empirical standard deviation being similar to the mean standard errors
3. the coverage being within (0.936,0.964)

7.1 Analysing the models

We shall explore the differing GEE models. We explore the GEE model by changing the correlation three times, but we effectively have six combinations that we can try out due to the fact that we can use the mean naive standard errors or that of the robust. For any of the correlation structures with the GEE model we see that there is no real bias that occurs in the estimates. This is expected as the fixed effects are correctly specified in the GEE.
Fow IWA, the robust standard errors are a little low compared to the empircal SD's but give good coverage throughout. However, the naive standard errors are much too low and hence give poor coverage.The GEE with IWA and naive errors is essentially the multiple linear regression model and this poor fit is expected due to the normality and independent error assumptions for being broken as seen in earlier sections.
The exchangeable correlation leads to lower empirical SDs than for the IWA as it incorporates a better correlation more suited to the data set. Overall we see good coverage and standard errors (both naive and robust) in this model, except for low mean standard errors (and low coverage) for the $\hat{\beta}_3$ which is associated with time. This particular correlation structure is similar to incorporating something similar to a random intercept as it allows for a common correlation between GMFM measurements on the same patient, but allows this to be a different common correlation when compared to other patients. Hence, reduced variation in the parameters may be expected as the exchangeable correlation explains some of the variation, and this could be the reason for the poor coverage for $\hat{\beta}_3$. The other small issue to note is that the mean naive and robust standard errors are similar true correlation isn't correctly specified.
For the unstructured correlation, we see the highest empirical SDs for the GEE models. This is due to the lack of structure in the specified correlation. There is reasonable coverage under the robust standard errors, but the mean standard error estimates are much too high. On the other hand we have the reverse with the mean naive standard errors being much too low with poor coverage. Hence we believe that the exchangeable correlation and robust variances is the most appropriate GEE model of the three considered.
The mixed effects models have, in addition to the $\hat{\beta}$ estimates, the standard deviation (and correlation in the true model) for the random effects, but with only one set of mean standard estimates to compare to the empirical ones. Comparing the mixed effects models we see that once again we have no real bias. The model incorporating both random effects is the true model and as such we see that the mean standard errors match quite

nicely to the empirical standard deviation and provide good coverage. Note that the coverage is a little high, but is nothing to cause concern. For the random effects, we do not obtain standard errors for the estimated $\hat{\sigma_0}, \hat{\sigma_1}, \hat{\sigma_\epsilon}$ or $\hat{\rho}$. Consequently we cannot calculate coverage for these and can only comment upon the empirical standard deviation. We see that the empirical SD's look quite acceptable here, and no bias is shown for the estimates.
If we now look at only incorporating the random intercept we see similar empirical standard deviation and mean standard errors with good coverage for the fixed effects, except for $\hat{\beta_3}$ which is too low. Now this is due to the fact that we have removed the random slope which is associated with Time3 and so the fixed effect of Time3 is affected. Also note that by doing so we have moved the uncertainty from the random slope and shifted it onto the random intercept and residual error which both have increased error estimates. We can see that the fixed effects of this model are similar to the GEE model with exchangeable correlation as expected from Zuur (2009) [17].
The final model removes the random intercept, but incorporates the random slope associated with Time3. Although the estimates are fine for the fixed effects, we have considerably low coverage due to the small mean standard errors for $\hat{\beta_0} : \hat{\beta_2}$ Conversely, $\hat{\beta_3}$ has a high mean estimated error and high coverage. We also see that by removing the random intercept, the uncertainty has been shifted onto the remaining random effects noting that the estimate for $\hat{\sigma_1}$ has doubled with the residual deviation more than tripling!
In conclusion we see that there is never any real bias in the estimates of the fixed effects, but it is the standard deviation (and as a result coverage) that is affected by misspecification. We see that under-specified correlation predicts low standard errors, whereas allowing for too much correlation (unstructured) does the reverse and captures the true value, but down to having larger standard errors. We note that mis-specifying random effects has a large effect upon the remaining effects. Removing the slope has an effect, but we see that by mis-specifying the random slope has considerable affects upon the ability to capture the remaing true values. We saw in Chapter 2 that we had a large range of initial GMFM measurements for each patient, and these simulations were based upon having this feature. Hence it is to no surprise that by ignoring the random intercept for each patient has a large impact on the coverage and standard errors in the model.
The biggest impact on the model appears to be via using mis-specified random effects although using naive standard errors over robust also appear to lower coverage. The best model at predicting the true model was the random effects model that incoporated both random terms. Subsequently, we continue to include the fixed intercept, HYPOXIC, TBI and Time3 and a random intercept and slope within our model for the GMFM measurements.

8 Prediction

There are two main aspects as to why we want to model data. In previous sections it has been to explore the effect of different covariates upon the GMFM measurements. For

instance we have found that females tend to have a slightly higher GMFM score then males, but not significantly different. Or perhaps the more prominent relationship being that the HYPOXIC injuries significantly reduce the GMFM scores when compared to ABI or TBI injury cases. The other reason that we model data is to perhaps predict cases that have not been seen yet. This may be the next patient score who is in the trial, or perhaps to predict the GMFM scoring for a new patient altogether. We look at the first case of predicting a patient's next score.

One issue that we have is that we currently don't have a new patient's GMFM scores and so how can we test prediction as we have used all of the data to create the model? We have chosen to use cross validation which removes a point from the data and uses the rest of the data to predict this value. In our case, we choose to remove the last point from each of the patients, create a model with the remaining data, and predict the point we removed. We can compare the actual and predicted points to assess the effectiveness of our model. We calculate cross validation via:

$$CV = \frac{1}{n}\Sigma_{i=1}^{n}(f_i - \hat{f}_i)^2$$

[17], where f_i is the true value of the last GMFM measurement for child i and $\hat{f}_i$ is the predicted value of this point and n being the number of predicted values. In this case, $n = 47$ due to having to remove 27 patients with only one or two measurements being taken as we cannot predict a random slope with this lower number of measurements. The value of the cross validation term is a measure of the error associated between the actual and predicted GMFM scores and so we wish for this to be minimal. Note that $\sqrt{CV}$ gives the root mean square error (RMSE), and as such we analyse the RMSE in our analysis.

We decided to base our predictions on the random effects model as it has been thought to be the best model so far. We carried out three versions of the model to predict the removed GMFM scores. Firstly a model with no covariates or random slopes was used to predict the removed points as a baseline. The other two models included the type of injury, Time3, but the latter included a random intercept and slope associated with Time 3 as well. Figure 13 shows the predicted points from the models with a smoother calculated as described in Section 4.2. We hope to see a linear pattern to the data following a $Y = X$ manner. We see that with no covariates we get one predicted value for all of the points. This is clearly not the case and misses the range of GMFM values there are.

The fixed effects begin to show a better variety of values. Looking at the smoother, there is a linear style relationship (*not quite* $Y = X$), but the GMFM values that are above the actual value of 60% are poorly predicted. Note also that there are a few points of actual value of 0:20% that are predicted to be 40+%. This suggests that extreme GMFM values are harder to predict when incorporating only fixed effects. It is strange to see that the predicted values only occur between 20:60%. This is because the lowest predicted value occurs with a HYPOXIC injured patient and so GMFM$\approx$ $43.153 - 21.649 + 1.138 * Time3 = (21.504 + 1.138 * Time3)\% \approx 21.504\%$ at a early time period. However the highest comes from a TBI injured patient with a measurement

later on in time.

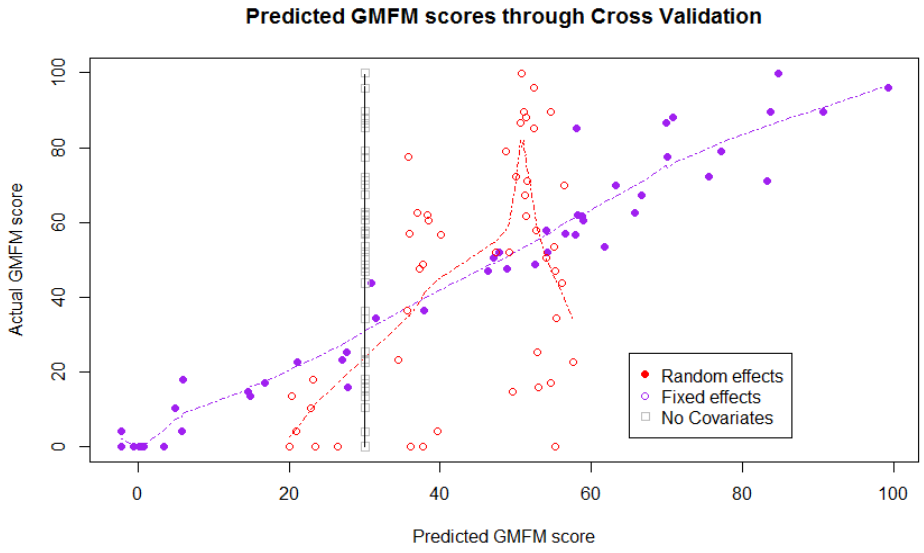


Figure 13: Plotting predicted GMFM vs actual GMFM measurements

We see the best relationship between the actual and predicted GMFM values when looking at the mixed effects model. The smoother shows that the values approximately follow a linear path and so the predicted values are similar to the actual GMFM scores.

Model	Root mean square error
No covariates	30.53
Only fixed effects	26.70
Fixed and random effects	7.43

Table 8: The root mean square error for the prediction of the last GMFM score for each patient (not including those patients with only 1 or 2 measurements)

Table 8 shows that the RMSE and re-iterates these points. We see a large variation in the no covariate model, and even a large variation appears in the fixed effects model. The root mean square error dramatically drops by around 19.27 from the fixed effects to the same model but including the random effects. This shows that the random intercept and slope are vital in the calculation of prediction in this longitudinal data. In conclusion, we see that the patient-specific random effects play a big role in the prediction and modelling of brain recovery. It also re-iterates the point that the mixed effects model is the most appropriate model to this data that we have looked at so far.

9 Conclusion

We have explored the GEE model with differing correlation assumptions and found that a suitable correlation would be one that follows the exchangeable structure. The random effects model suggests that we have an average and a patient-specific brain recovery rate,

with the average being specifically important as we saw when removing the intercept in Chapter 7. However, when it comes down to it, we conclude that the random effects model is most suited to modelling the paediatric data.
We saw that during prediction, it was vital to incorporate the random, patient-specific, effects in the model, or else the predicted values were far away from the observed values. The random effects are highly important in the accuracy of predicting future GMFM scores for children.
From this modelling, we have seen that the type of injury that a child has along with the time since injury is a big factor when predicting the brain recovery. One aim of the study was to find any relationship between aetiology and gross motor recovery, which we did. We saw that the HYPOXIC injured children were much more poorly and had very different GMFM scores than the ABI and TBI children. These injuries were described in Section 1.2, and HYPOXIC did appear to be of quite a different causation than ABI and TBI which agrees with out findings.
A second aim of the study was to explore the relationship of Age on the gross motor recovery. We however found no significant association between the two thus far. Although in general we found that the children recovered with respect to time, but HYPOXIC once again less so. We did find that females generally had higher GMFM values than males, but there was no significant difference and so there is no real evidence of this. However, as mentioned in Section 2.1, there was only one HYPOXIC female and so it may just be that there isn't enough females in the study to allow for sensible inferencing upon the significance of gender.
We believe that there is future work that can be done to develop the study further. Some of these are:

1. We could look at using alternate transformations of the response variable in the random effects model. Although non-normality could not be found from transforming GMFM, it might be that the parameters have a better relationship with a transformed GMFM. If we used f(GMFM), it would be interesting to see if this yielded better prediction results and also to explore the effects of misspecification with respect to ill-transformed variables.

2. We have only looked at the linear model, however it was noted that the trajectories were often particularly s-shaped. We would need to explore adding curvature into the model. This may affect the significance of various variables, or perhaps just improve prediction.

3. We believe that there is further work on prediction.

   (a) We looked at the model's ability to predict the last GMFM score for the patient, however it may be that it is much better at predicting GMFM values within the data range.

   (b) We could look at incorporating random change points to aid prediction. For instance, we could incorporate a random change point for the moment recovery begins, or perhaps begins to converge to a set gross motor recovery.

4. We could explore the effect of dropout. Each patient follow-up is of a different length. The patients with incomplete follow-up or drop-out can cause bias as an analysis that ignores this may not represent the original target population. For instance, we had to remove 27 out of 74 children during prediction as they had only one (13 children) or two (14 children) measurements taken. This is around a third of our data set. We should consider that the information we have lost from a lack of measurements upon those patients could have change our findings if follow-up was longer.

References

[1] Box, George EP, and David R. Cox. (1964) An analysis of transformations. Journal of the Royal Statistical Society. Series B (Methodological): 211-252.

[2] Buddelmeyer, Hielke, et al. (2008) Fixed effects bias in panel data estimators. Available at SSRN 1136288

[3] Burnham, Kenneth P., and David R. Anderson.(2004) Multimodel inference understanding AIC and BIC in model selection. Sociological methods and research 33.2: 270, 299, 261-304;

[4] Diggle P, Zeger, Scott L. Time Series, (1991) A Biostatistical Introduction. Clarendon Press, Oxford, 1990. No. of pages: xi+ 257. Page: 501.

[5] Diggle, Peter, et al. (2002) Analysis of Longitudinal Data. Oxford University Press, second edition.

[6] Friedman, Jerome H., and Werner Stuetzle. (1982) Smoothing of scatterplots. No. Orion-003. Stanford university CA project Orion.

[7] Geddes, J. F., et al. (2001) Neuropathology of Inflicted Head Injury in Children II. Microscopic Brain Injury in Infants. Brain 124.7: 1299-1306.

[8] Kelly, G., Mobbs, S., Pritkin, J.N., Mayston, M., Mather, M., Rosenbaum, P., Henderson, R. and Forsyth, R. (2015). Gross Motor Function Measure (GMFM-66) trajectories in children recovering after severe acquired brain injury. Developmental Medicine and Child Neurology 57, 241-247.

[9] Kuha, Jouni. (2004) AIC and BIC comparisons of assumptions and performance. Sociological Methods and Research 33.2: 188-229.

[10] Linder-Lutch M, Other V, Walther M, Vry J, Michaelis U, Stein S, et al. (2007) Validation of the Gross Motor Function Measure for use in children and adolescents with traumatic brain injuries. Paediatrics. 120(4):e880-6

[11] Motorgrowth.canchild.ca (2015). Overview-CanChild. [online] Available at: http://motorgrowth.canchild.ca/en/GMFM/overview.asp [Accessed 29 Apr. 2015].

[12] Pan, Wei. (2001) Akaike's information criterion in generalized estimating equations. Biometrics 57.1: 120-125.

[13] Shapiro, S. S. and Wilk, M. B. (1965) An analysis of variance test for normality (complete samples); Biometrika, 52, pages 591-611.

[14] Stat.ethz.ch, (2015). R: Friedman's SuperSmoother. [online] Available at: https://stat.ethz.ch/R-manual/R-patched/library/stats/html/supsmu.html [Accessed 29 Apr. 2015].

[15] Venables, W. N. and Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

[16] VLI Ltd, U. (2015). Headway-Hypoxic/anoxic brain injury. [online] Headway.org.uk. Available at: https://www.headway.org.uk/Hypoxic-anoxic-brain-injury.aspx [Accessed 29 Apr. 2015].

[17] Zuur, Alain, et al. (2009) Mixed effects models and extensions in ecology with R. Springer Science and Business Media.

Appendices

A Transforming Age and Time

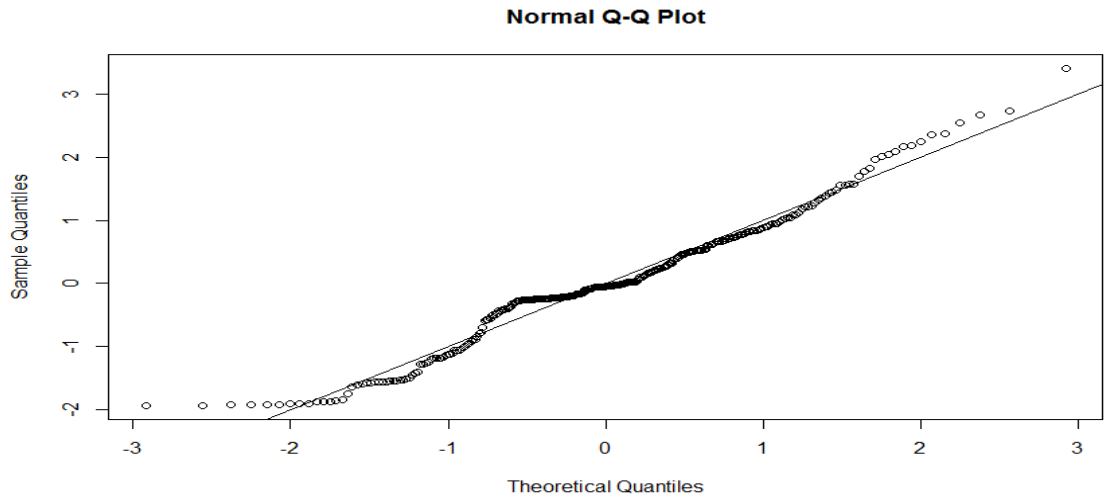


Figure 14: QQ plot for the regression containing the transformed Time variable

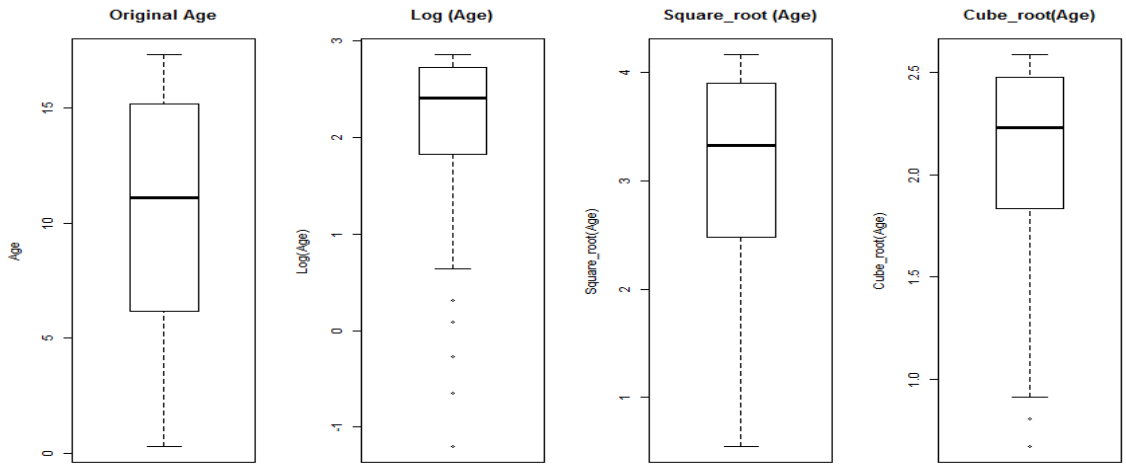


Figure 15: Boxplots of Age with three transformations (log, square root, cube root)

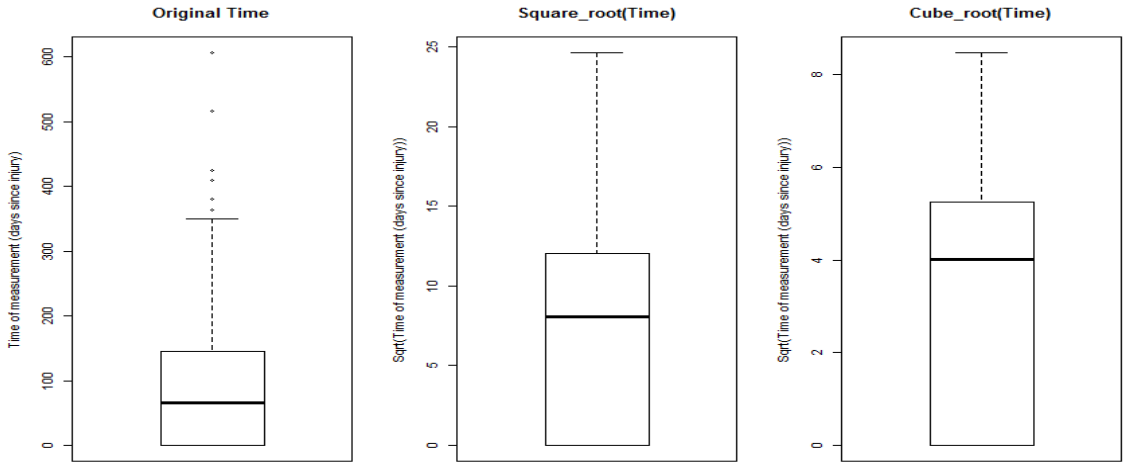


Figure 16: Boxplots of Time with two transformations (square root, cube root), no log transformation was used due to some measurements being at time zero