THE UNIVERSITY OF NEW SOUTH WALES

School of Mathematics and Statistics

MATH5885 Longitudinal Data Analysis

Term 2, 2019

Project

- Due Sunday 4 August (end of Week 9) via Moodle.

  Your report should consist of two parts.

  1. **Part 1: Report** (in a file named `zID-MATH5885-Project-Report.pdf`: maximum of 12 pages typed in minimum 12 pt font, single line spacing with minimum 2.5cm margins, single sided) which should include mathematical summaries of the models fit, essential R code and output only, any essential tabular and graphical output with a narrative about how you arrived at key modelling decisions, and your summary of findings or conclusions. You should also describe any model deficiencies and suggest possible remedies. Further details below.
  2. **Part 2: Appendices** (in a file named `zID-MATH5885-Project-Appendix.pdf`) containing R code and any additional graphs and tables properly labelled so that the main report can cross reference these and so that I can quickly locate the relevant R code and additional tables and graphs should that be needed. This is not a defacto extension to your report. Your Part 1 Report should stand on its own and be readable without reference to the Appendix.

- If you are not skilled at producing typeset reports, then neatly handwritten reports are acceptable provided the specifications on font size, margins, line spacing etc described above are reasonably conformed to.

- For upload to Moodle you should use naming conventions given above with your student ID replacing zID.

- Please include this cover sheet with your submitted Part 1 Report with your name and student number filled in.

  Declaration:
  I declare that this assessment item is my own work, except where acknowledged, and has not been submitted for academic credit elsewhere.

  I certify that I have read and understood the University Rules in respect of Student Academic Misconduct.

Signed:
Date: 4 AUG 2019
Name: SAM MASON
Student number: 1058231

Introduction

This report summarises the findings of a brief investigation into a small sample of the Multicenter AIDS Cohort Study dataset. It is known that HIV destroys CD4 cells as the disease progresses and the aim of this project was to attempt to develop a feasible model relating the evolution over time of CD4 cell counts as a function of a number of other pre-specified covariates.

The dataset consists of longitudinally collected observations on 369 subjects, resulting in a total of 2376 observations of CD4 cell counts. The covariates include time, subject age, cigarette use, CESD score (a depression indicator), drug use and sexual activity. A detailed data dictionary was supplied with this project specification and is included in the appendix document.

Note that this is simply an exercise in modelling the transformed (square root) response variable with respect to these covariates and it is not the intention of this work to imply any causality.

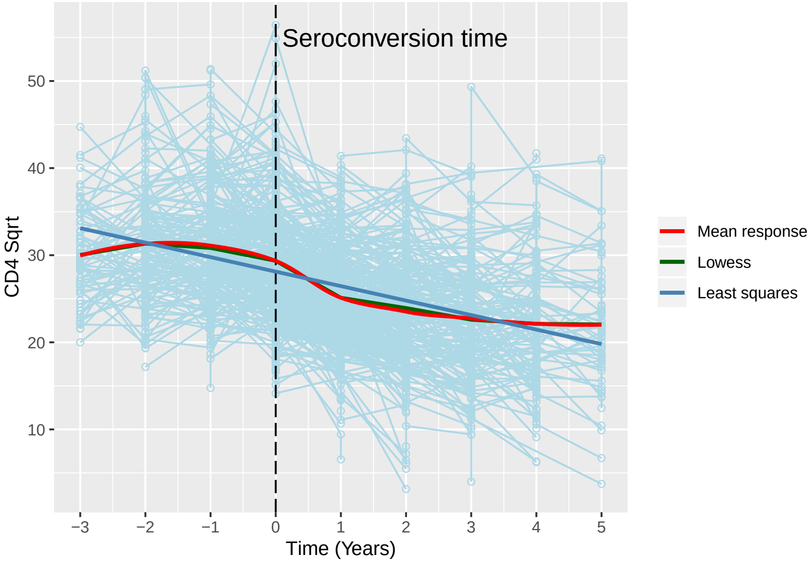
Exploratory Data Analysis

Table 1 below shows the variation of subjects and observations over the course of the study and we note a significant drop off by lag 5. Thus we need to be careful interpreting any mean response trends at high lags and use robust techniques when assessing trends. There is a lot of missing data and the dataset is highly unbalanced.

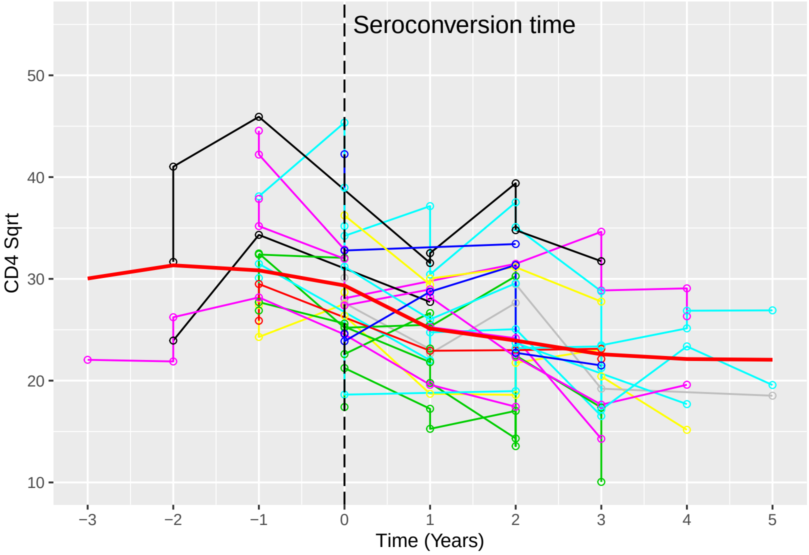
Table 1: Study Observations Profile

Year	-3	-2	-1	0	1	2	3	4	5
Num.Observations	71	198	315	529	431	346	254	163	69
Num.Subjects	70	133	211	307	279	226	167	109	51

Inspection of the mean response profiles shows a slight increase in CD4 cell levels just prior to seroconversion followed by an overall decrease with time post-baseline. This rate of decrease appears to be quite steep over the first year, followed by a slower decline from years one to three and flattening out to a shallow decrease rate thereafter. There is some visual evidence of a possible rise in mean response at year five consistent with less observations at these higher lags as subjects become lost to follow up.

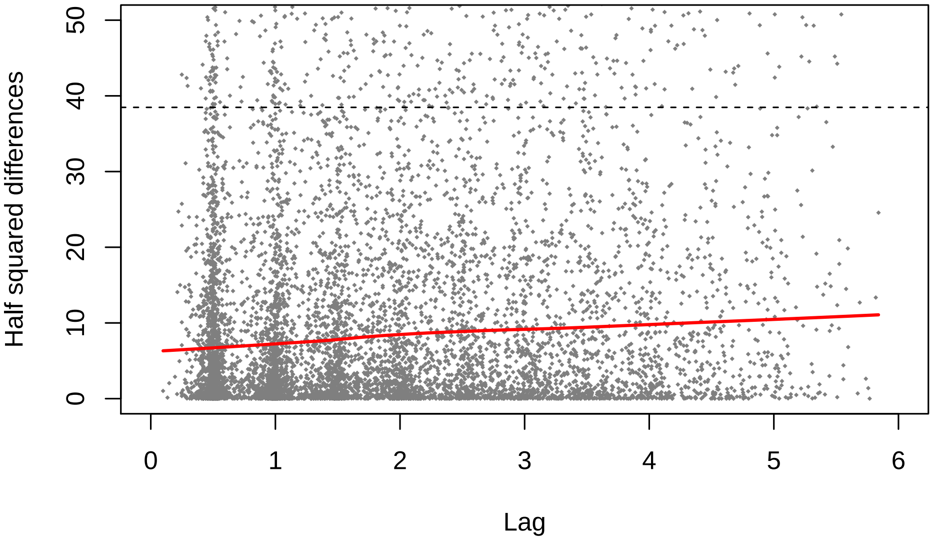


While there is a significant amount of missing data in this study, between-individual heterogeneity is still evident as subjects with higher CD4 levels tend to remain above average while those with low cell counts tend to remain below the overall average. Within-subject variation appears somewhat random but post-baseline, there is some evidence of persistent serial correlation. A time plot for a selected random sample of subjects is shown below where the population mean response curve is shown in red. In this plot the within-subject variability and between subject heterogeneity is much clearer.



To attempt to get a feel for the correlation in this data we can plot a sample variogram of the residuals from a simple spline fit and we see some evidence of persistent serial correlation even at large lags that appears to be decaying fairly slowly.

```
cd4_df$resid <- resid(smooth.spline(cd4_df$yr, cd4_df$CD4sqrt))
```

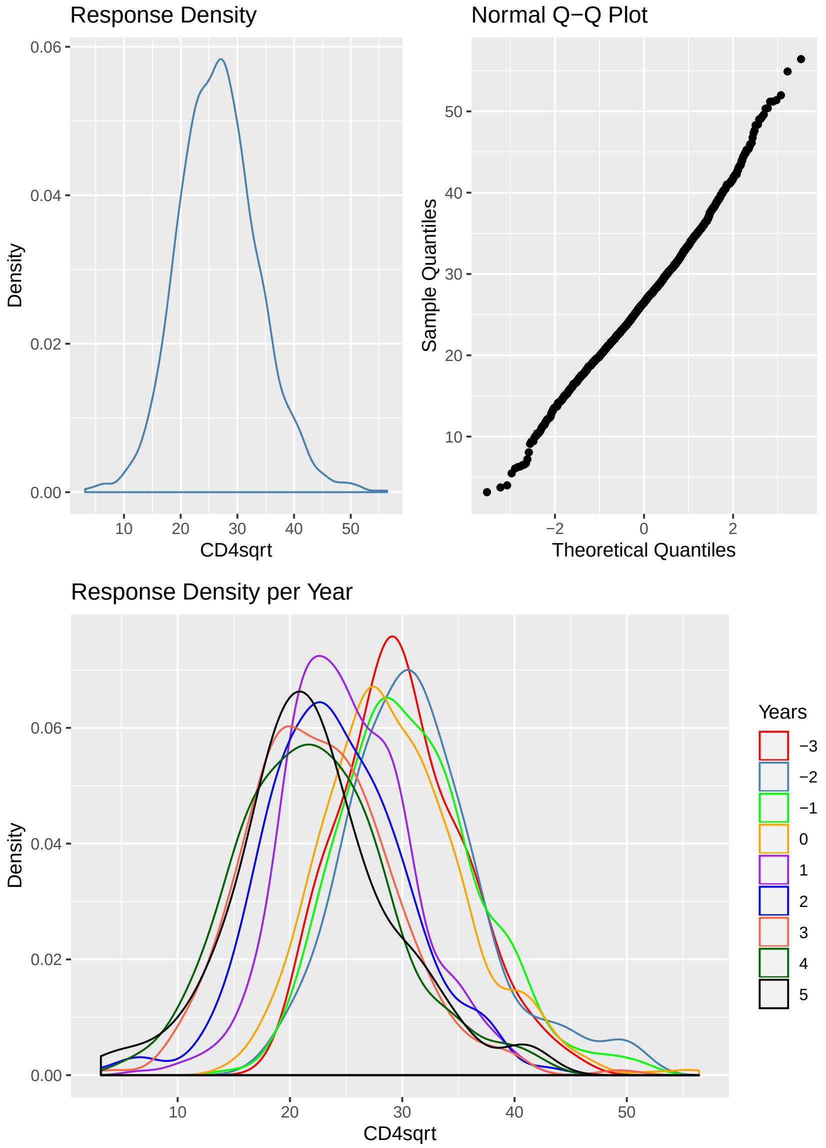


As can be seen in both the variogram above and the variances of the residuals in Table 2 below, the variance is not really constant and exhibits a general increase with lag. The correlation is persistent and exhibits very slow decay - almost constant with higher values out at large lags where there is much less data. The full correlation matrix is available in the appendix document.

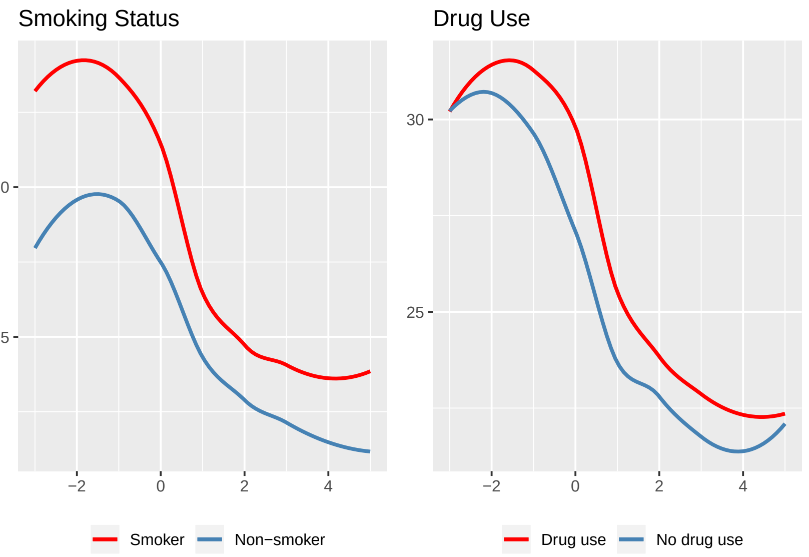
Table 2: Variances of Residuals

	resid.-3	resid.-2	resid.-1	resid.0	resid.1	resid.2	resid.3	resid.4	resid.5
Variances	27.98521	39.50176	40.56666	41.65118	28.78789	40.44546	37.82702	44.86887	53.37218

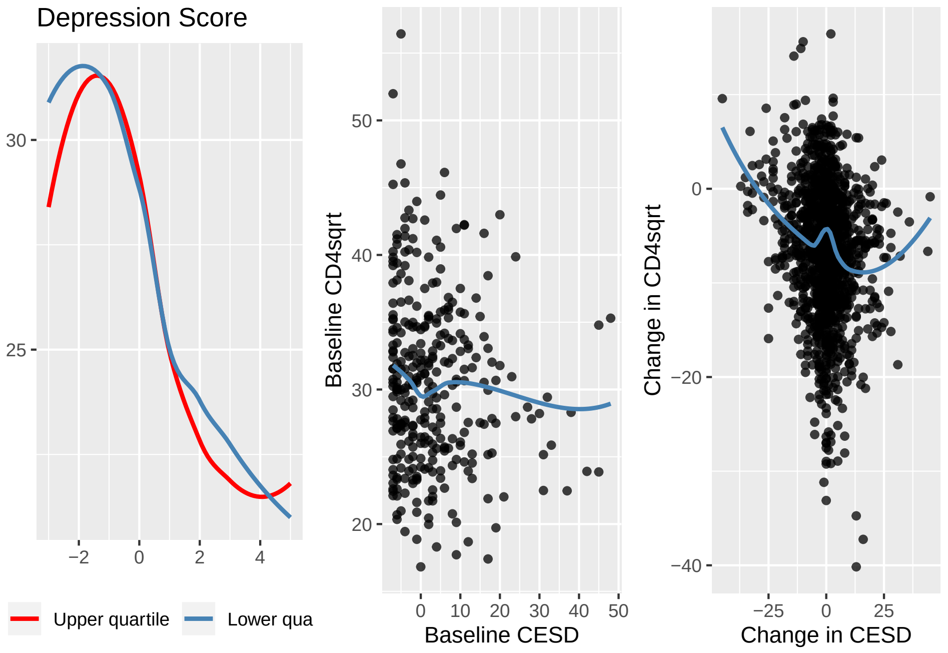
As has been shown in supplied work (CD4InitialAnalysis.R), the distribution of the response variable (the sqrt of CD4 cell counts) is approximately normal. Here we confirm this and also confirm that this holds roughly true for each year in the study and that therefore we can assume normality of residuals when modelling.



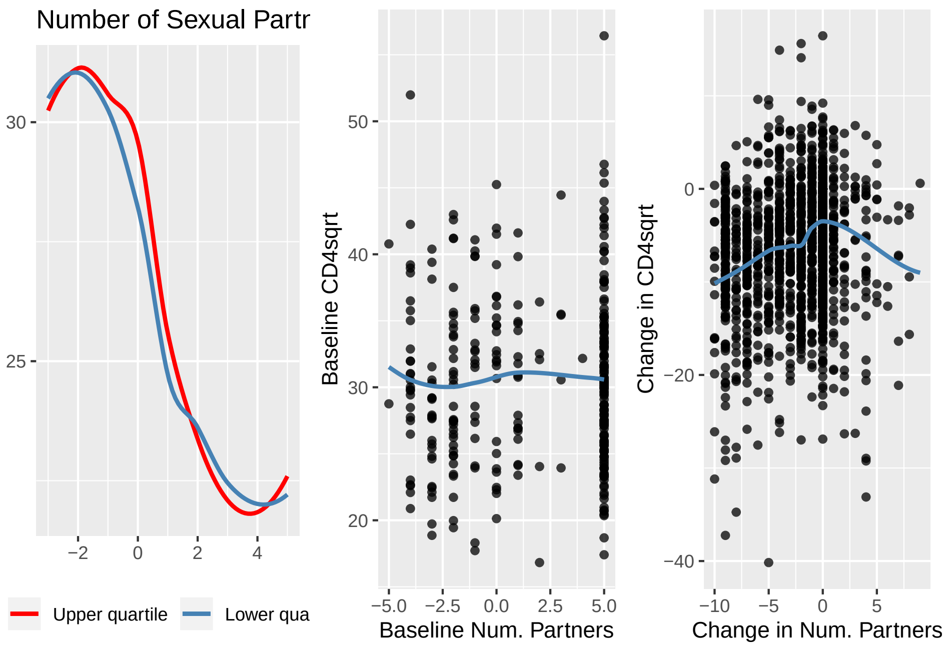
Next, we explore the relationship between four key covariates and the evolution of the response over time. To start with, note that the 'Packs' covariate has been used to group subjects into 'Smokers' and 'Non-Smokers' (zero Packs per day). As seen below there appears to be a fairly consistent group effect across time for Smokers/Non-Smokers as well as for recreational drug use.



For CESD depression scores, there does not appear to be any cross-sectional effect but we might expect depression scores to be associated with a decrease in CD4 counts longitudinally:



For number of sexual partners, there does not appear to be any cross-sectional effect. There is some longitudinal effect but it is difficult to interpret given that this covariate has been centered:



Similar plots and analysis for other covariates are detailed in the appendix document.

Model Formulation

The EDA process has resulted in a number of findings/indicators regarding approaches to modelling:

- CD4 cell counts exhibit three key changes in rate of decrease from seroconversion at times t=0, t=1 and t=3 years. This suggests a linear piecewise model with knots at these locations might be effective.
- The unbalanced nature of the data limits the covariance structures we can consider to account for serial correlation - specifically, the persistent correlation and slow decay over time suggests parametric representations such as exponential or Gaussian to model this.
- The large range of subject responses (i.e. between subject heterogeneity) suggests considering random effects for subject-specific intercepts and possibly for each slope segment in a piecewise model due to the high within-subject variability.
- As discussed above there is some evidence of a possible group effect for both Smokers vs Non-Smokers as well as Drug Use vs No Drug Use in the relevant stratified mean response profiles.
- The Smoker group effect (if it exists) appears more pronounced than any group effect relating to drug use.

Based on these, the modelling process proceeds as follows:

- Propose and fit a number of preliminary linear models ignoring correlation to get a sense of coefficients and basic significance.
- Choose one of these as provisional "maximal" model for the mean and holding that fixed, fit and assess a series of candidate covariance structures.
- Choose a suitable covariance structure and then re-fit and re-assess variables in the model.
- Explore random effects for intercepts and slope terms.
- Finalise model and assess the variability of the response the model is able to represent.

To begin, we fit two simple linear models, ignoring the fact that we have correlated errors and inspect regression coefficients and significances. Note that a full group interaction effect for the Smokers group is modelled here as this is not a randomised controlled trial and therefore there is no reason to assume apriori that these groups have the same intercepts at baseline (and in fact our EDA suggests that they do not).

```
lm.basic1.fit <- lm(CD4sqrt ~ Time + smoker + Age + Drugs + Cesd + Sex, data = cd4_df)
lm.basic2.smkr.int.fit <- lm(CD4sqrt ~ Time*smoker + Age + Drugs + Cesd + Sex, data = cd4_df)
```

```
anova(lm.basic1.fit, lm.basic2.smkr.int.fit)
```

```
## Analysis of Variance Table
##
## Model 1: CD4sqrt ~ Time + smoker + Age + Drugs + Cesd + Sex
## Model 2: CD4sqrt ~ Time * smoker + Age + Drugs + Cesd + Sex
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1   2369 89026
## 2   2368 88560  1    465.67 12.451 0.0004256 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The anova test strongly rejects the null hypothesis that the interaction terms are not required. Note that generally speaking the Age and Sex covariates are not significant (p-vals of 0.12 and 0.72) and that there is strong evidence of a group effect for Smokers (p-val 0.0004) which aligns with expectations from EDA.

The negative sign of the CESD coefficient suggests that increases in depression scores are associated with decreases in CD4 cell counts (as was also suggested in the EDA process).

The third preliminary model is piecewise linear with knots at times t=0, t=1 and t=3:

```
lm.basic3.3knots.fit <- lm(CD4sqrt ~ Time*smoker + Time0 + Time1 + Time3 +
                               smoker.Time0 + smoker.Time1 + smoker.Time3 +
                               Age + Drugs + Cesd + Sex, data = cd4_df)
```

At this point we have three fixed effects models, all of which ignore correlation, with AIC values of 15368.25, 15357.79 and 15303.75 respectively. In all three models, neither Age nor Sex covariates appear significant. A plot of the residuals for the 3-knot model is shown below and surprisingly shows a fairly good model fit.

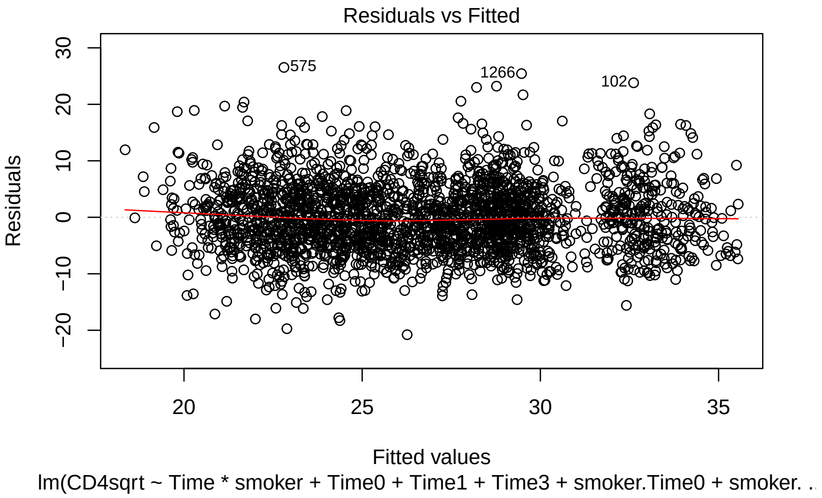


Table 3 below shows the results of using the piecewise linear model as our 'maximal' model while exploring various candidate covariance structures.

Table 3: Covariance Structures Summary

Variance.Model	Nugget	AIC
Compound symmetry	NA	14444.61
Exponential Decay	No	14454.11
Exponential Decay	Yes	14264.99
Gaussian Decay	No	14863.40
Gaussian Decay	Yes	14277.90

Likelihood ratio tests were performed against the compound symmetry model for both exponential and Gaussian models. In both cases, the test statistics are large (181.6 and 168.7 respectively) with p-values <

0.0001 and thus we can reject the null hypothesis and conclude that we are justified in adopting these simpler covariance structures.

Based on the AIC values in Table 3, we choose the exponential model with a nugget effect for our covariance structure as it has the lowest AIC value (14264.99).

The next step is to refit this model using maximum likelihood and re-assess the fixed effects currently in this model through a sequence of drop-refit cycles. Table 4 shows the model fit improvements made at each step when this is done.

Table 4: Fixed Effects Simplification Steps

Action	AIC
Refit by ML	14239.49
Remove Age	14237.69
Remove smoker.Time0	14235.75
Remove smoker.Time1	14235.28
Remove smoker.Time3	14237.96

Note that once the Age covariate is removed from the model, the presence or absence of the three smoker.Time variables makes little difference to the overall model fit. In fact, when each of these steps are performed in the sequence above, both the smoker.Time0 and smoker.Time1 variables are removed from the model but the smoker.Time3 variable is retained (at a 5% level). However, the penalty for removing this last term is small for a gain in model simplicity in terms of interpretability and therefore this has been removed from the model as well.

Note that it is also of interest that the number of sexual partners covariate (Sex) has remained in the model in contrast to earlier expectations in this document. At this point the current model, refit with REML, is shown below.

```
gls.exp.fit <- gls(CD4sqrt ~ Time*smoker + Time0 + Time1 + Time3 + Drugs + Cesd + Sex,
                   correlation = corExp(form = ~ Time | ID, nugget = T, value=c(2, 0.1)), data = cd4_df)
AIC(gls.exp.fit)
```

```
## [1] 14259.54
```

The final step is the introduction of any random effects. As stated at the start of this analysis, it is expected that a random effect on the individual intercept might be beneficial given the high between-subject heterogeneity. Further, the high within-subject variability may be well represented with random effects on subject-specific slopes.

Shown below are three candidate mixed effects models. The first includes a random effect on the intercept term, the second on both the intercept term and the overall slope and the last more complex model attempts random effects on each of the slope terms of the original 3-knot piecewise linear model.

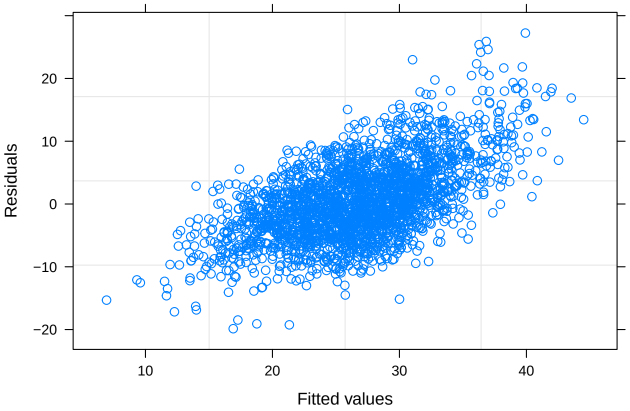
```
me.exp.fit <- lme(CD4sqrt ~ Time*smoker + Time0 + Time1 + Time3 +
                    Drugs + Cesd + Sex, data = cd4_df,
                  random = ~ 1 | ID,
                  correlation = corExp(form = ~ Time | ID, nugget = T, value=c(2, 0.1)),
                  control = lmeControl(opt = 'optim', maxIter = 200))

me.exp2.fit <- lme(CD4sqrt ~ Time*smoker + Time0 + Time1 + Time3 +
                    Drugs + Cesd + Sex, data = cd4_df,
                  random = ~ Time | ID,
                  correlation = corExp(form = ~ Time | ID, nugget = T, value=c(2, 0.1)),
                  control = lmeControl(opt = 'optim', maxIter = 200))

me.exp3.fit <- lme(CD4sqrt ~ Time*smoker + Time0 + Time1 + Time3 +
                     Drugs + Cesd + Sex, data = cd4_df,
                   random = ~ Time + Time0 + Time1 + Time3 | ID,
                   correlation = corExp(form = ~ Time | ID, nugget = T, value=c(2, 0.1)),
                   control = lmeControl(opt = 'optim', maxIter = 200))
```

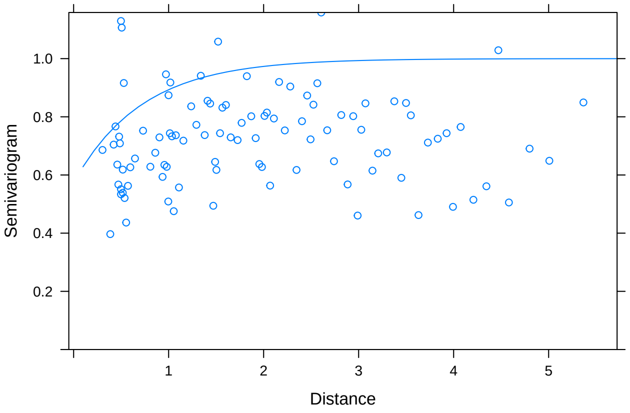
When each of these models were fit, the AIC values were 14261.61, 14230.02 and 14247.21 respectively. Thus, introducing random effect terms on both the slope and intercept provides a better overall fit than either the third complex mixed effects model or the first intercept only model.

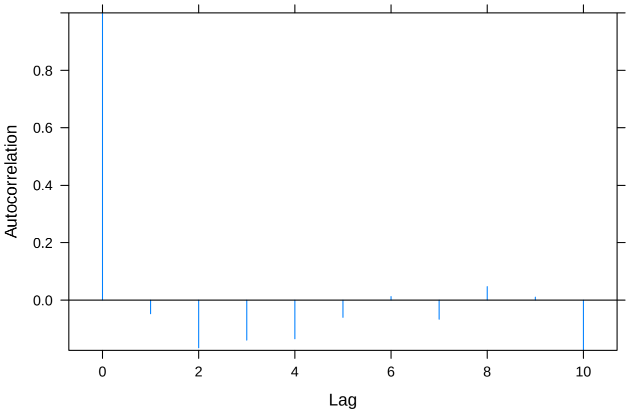
In order to correctly test the significance of each of these random effects, we would need to use the parametric bootstrap as we do not actually know the distribution of the LR statistic under the null (it being a complex mix of chi-squares). However, it is interesting to note that if we do a LRT between the above two models, the test soundly rejects the more complex model. Even though we cannot trust the p-value here (0.8695), it is unlikely to change such that an opposite finding would result.

Taking the simpler model as the best candidate for our final model we see that the introduction of a random effect on the intercept term allows the model to be simplified still further as the Drug use variable is no longer significant (p-val 0.0527). Removing this and re-fitting yields a final model with an AIC of 14226.69.

A residual plot, a variogram and an ACF for the residuals from this final model are shown below:



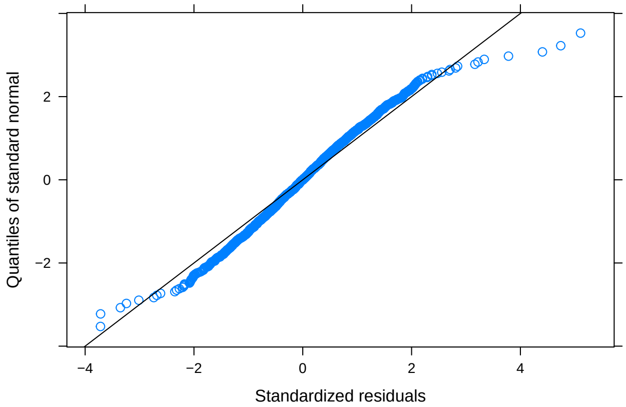


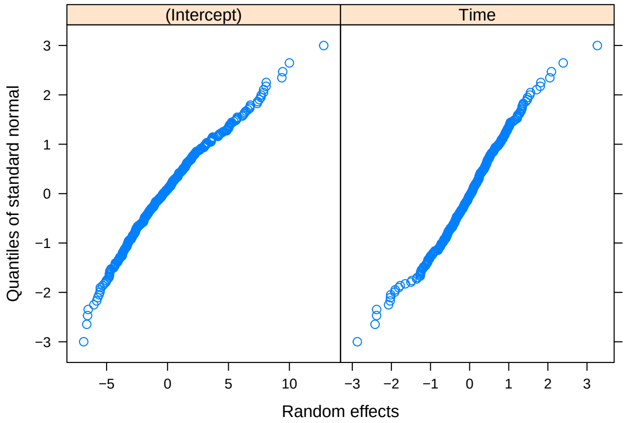


These plots indicate that this model has captured much of the between-subject and within-subject sources of variance as well as noise due to measurement error. In the ACF there is still some unrepresented correlation in the residuals but at lag 10 this is unlikely to be of concern as this is where data begins to become more sparse as the study continues and subjects are lost to follow up. The components of variability that this model does represent can be broken down as follows (derivation of these components is outlined in the appendix document):

1. The variance due to the intercept random effect is $\sigma_b^2 \approx 14.978$
2. The variance due to the correlated process is $\sigma_U^2 \approx 7.169$
3. The variance due to random noise from measurement error is $\sigma_e^2 \approx 9.657$

Plots assessing the normality of the residuals from both fixed effects and random effects are shown below.





Again, this shows that the assumptions made by the model are fundamentally sound although we do note deviations from normality in the tails which may warrant further investigation (e.g. trying a heavier tailed distribution such as a t-distribution). Regardless, for the purposes of this work we appear to have a satisfactory model.

The model fixed effects estimates are shown below:

```
##                 Value Std.Error   DF t-value p-value
## (Intercept) 29.0521    0.3327 1999 87.3190  0.0000
## Time        -0.2911    0.2183 1999 -1.3340  0.1824
## smoker       1.8491    0.3567 1999  5.1843  0.0000
## Time0       -5.2730    0.5605 1999 -9.4077  0.0000
## Time1        2.8271    0.3901 1999  7.2466  0.0000
## Time3        0.5543    0.1233 1999  4.4972  0.0000
## Cesd        -0.0412    0.0135 1999 -3.0622  0.0022
## Sex          0.0993    0.0363 1999  2.7381  0.0062
## Time:smoker -0.5762    0.1622 1999 -3.5532  0.0004
```

Overall, the explanatory variables regarding behaviours that might be considered “risky”, such as use of recreational drugs, multiple sexual partners and smoking all have positive estimated coefficients and therefore seem to be associated with increased levels of CD4 cell counts.

Higher depression scores are associated with a decrease in CD4 cell counts as of course does the progression of time. Note that for smokers, the interaction effect is negative indicating that the time affect on smokers is more pronounced than for non-smokers. Again, there is absolutely no basis here to infer any causality.

The final mixed effects fitted model is shown below. Note that smoking status can (and does) vary throughout the study hence the use of the j suffix.

$$\begin{aligned}E(Y_{ij}) &= \beta_1 + \beta_2 t_{ij} + \beta_3 smoker_{ij} + \beta_4(t_{ij} - 0)_+ + \beta_5(t_{ij} - 1)_+ + \beta_6(t_{ij} - 3)_+ \\ &+ \beta_7 Cesd_{ij} + \beta_8 Sex_{ij} + \beta_9 t_{ij} \times smoker_{ij} + b_{1i} + b_{2i} t_{ij} \\ &= 29.052133 - 0.291148 t_{ij} + 1.849093 smoker_{ij} - 5.272980(t_{ij} - 0)_+ + 2.827089(t_{ij} - 1)_+ \\ &+ 0.554318(t_{ij} - 3)_+ - 0.041219 Cesd_{ij} + 0.099269 Sex_{ij} - 0.576220 t_{ij} \times smoker_{ij} + b_{1i} + b_{2i} t_{ij}\end{aligned}$$

The combination of random effects in the model, specifically on the intercept term (which induces a compound symmetry structure) and the exponential model for covariance structure results in a hybrid model for representing overall variability of the response.

This is particularly useful as it combines the advantages of the parametric exponential model given the highly unbalanced data at hand, with the real world empirical observations that measurement error is always present and therefore within-subject correlation cannot be zero. In addition, it can be shown (see Week 5 lecture 1) that the within-subject correlation is always less than one, which again aligns with real world experience.

Application to Individual Trajectories

Subject ID 30119 has 12 measurements spanning a six year period of the study which are fairly well balanced both before and after sero-conversion (represented here by the 5th and 6th measurements). Table 5 shows the variances extracted from the diagonal of the estimated variance-covariance matrix for this subject. It can be seen that the variance decreases prior to sero-conversion and subsequently increases throughout the post sero-conversion period.

Table 5: Estimated Variances for Subject 30119

1	2	3	4	5	6	7	8	9	10	11	12
35.66	33.55	32.24	31.67	31.65	32.09	33.21	36.17	38.57	42.71	47.2	52.68

Table 6: Estimated Correlations for Subject 30119

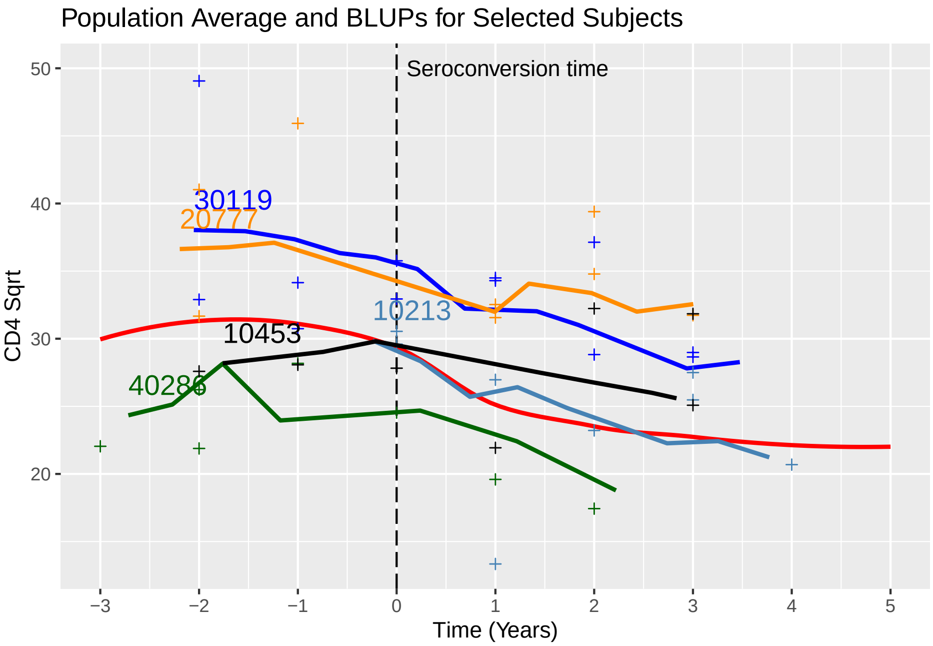
1	2	3	4	5	6	7	8	9	10	11	12
1	0.61	0.53	0.48	0.45	0.41	0.36	0.29	0.26	0.21	0.17	0.13

Table 6 shows the first (representative) row of the correlation matrix for this subject where we observe persistent positive correlation over the whole study period. This correlation decays fairly slowly with time overall but appears to decay slightly faster during the post sero-conversion period. Note that correlations do not decay to zero even when measurements are taken years apart - in this case over a six year period. The full estimated covariance and correlation matrices for this subject are included in the appendix document.

In this section we have selected five subjects from the dataset such that they span the sero-conversion point and represent a range of high, medium and low responders as specified. For each subject, trajectories have been estimated using empirical BLUPs and plotted below. In addition, the estimated population mean response curve is shown in red for reference and the transformed observed CD4 cell counts for each subject are plotted as coloured crosses.

Note that each BLUP declines over time with several exhibiting varying slopes for each subject at the three knot points in the original model. Some BLUPs track observations closely (e.g. subjects 30119 and 10213) while others display far more variability over time (e.g. subjects 40286 and 20777).

The BLUP for subject 10213 closely tracks the population mean response whereas the BLUP for subject 20777 apears to diverge after sero-conversion. Both are examples of how individual subject variability is weighted against population mean response when BLUPs are calculated.



Discussion of Modelling

The main issues encountered in this work were to do with convergence and model specification syntax. The large number of observations precluded any feasible use of an unstructured covariance structure as a base to compare candidate covariance models using Likelihood Ratio Tests on REML fits.

Moreover, the various combinations of the “form” and “weights” syntax in gls, lme and variograms made it difficult to be certain that all parts lined up correctly. This is obviously a matter of a lack of experience with longitudinal data analysis (that currently being somewhat less than nine weeks in total) and one that will be overcome with practice.

In terms of the behaviour of the data itself, it was not clear whether to proceed down a more complex path and fit cubic splines and possibly other group effects regarding use of recreational drugs. Furthermore, this work focussed on modelling the CD4 cell count response post sero-conversion given the wide range of observed counts leading up to that time point. This allowed a more simple modelling approach given the observed 3-stage decline in response in this period and favoured a more parsimonious final model.

Lastly, the assumption of normality of residuals, while it appears to be reasonable, would be re-assessed in further work. This is fundamentally count data and it would make sense to attempt a model using a GLMM with a canonical Poisson link.

References

[1] Fitzmaurice GM, Laird NM, and Ware JH. *Applied Longitudinal Analysis, 2nd. Ed.* John Wiley and Sons, 2011.

[2] Diggle PJ, Heagerty P, Liang K-Y, and Zeger SL. *Analysis of Longitudinal Data, 2nd. Ed.* Oxford University Press, 2002.

[3] Diggle PJ and Zeger SL. *Semiparametric Models for Longitudinal Data with Application to CD4 Cell Numbers in HIV Seroconverters* . Biometrics, vol. 50, no. 3, pp. 689-699, (1994).