MATH 5900 – ADVANCED DATA ANALYTICS

ASSIGNMENT 4

TAIWO JEGEDE – E00755963

APPLIED DATA SCIENCE

SOFTWARE USED: R

QUESTION 1

I

First of all, Marginal and conditional longitudinal models are two approaches that we use when analyzing longitudinal data.

Whenever we want are interested in the population-level effect, and the focus is on estimating average effects across all subject we would use marginal longitudinal models. And Conditional longitudinal models are more suitable in statements regarding individual-specific effects or trends and we want to model the correlation structure within subjects explicitly.

The conclusions we draw from the models can sometimes be different because it depends on the research question and the level of interest. Marginal models provide insights into average effects across the population, ignoring individual variability, while conditional models offer a more nuanced understanding by considering subject-specific effects and accounting for within-subject correlation.

II

Three descriptive statistics commonly used to explore longitudinal data sets are:

Mean: The mean provides a measure of central tendency, indicating the average value of a variable across all time points or subjects so when we calculate the mean for each time, the point can show the overall trend or pattern in the data over time and it helps us in understanding the general level of the variable and how it changes longitudinally.

Standard Deviation: The standard deviation measures the dispersion or variability of data points around the mean. In a longitudinal context, it can indicate how much individual data points

deviate from the average at each time point. A higher standard deviation suggests greater variability in the data, while a lower standard deviation indicates more consistency over time.

Correlation Coefficient: The correlation coefficient assesses the strength and direction of a linear relationship between two variables. In longitudinal data analysis, calculating correlations between variables at different time points can reveal patterns of association over time. Understanding how variables are related longitudinally can provide insights into dependencies and interactions within the data set.

These descriptive statistics offer valuable insights into longitudinal data by summarizing key characteristics such as central tendency, variability, and relationships between variables across multiple time points or subjects

III

The (semi-) variogram is the same as the variogram, and it is a plot we use to analyze spatial data. Here, the variogram can be used to explore the spatial or temporal dependence structure within the data. The plot shows how the variance between pairs of observations changes as a function of distance or time lag.

IV

Random Intercept: Groups have different baselines (varying intercepts) but the effect of predictors (slopes) is the same for everyone.

Random Slopes: Both baselines and how predictors affect them (slopes) can vary across groups, capturing individual differences.

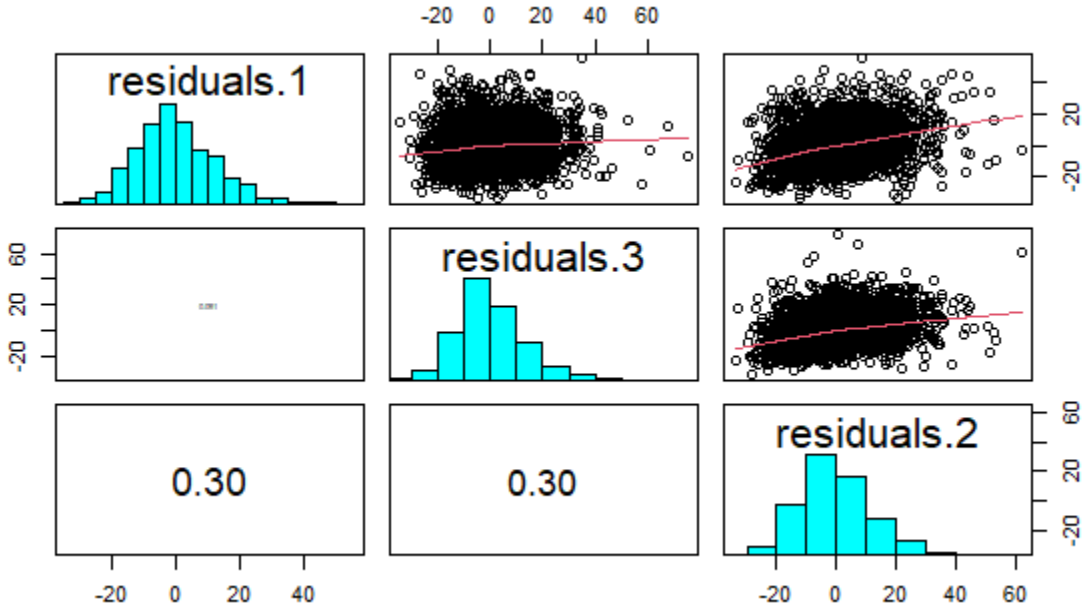
The random slopes model estimates multiple variance parameters: σ^2, σ^2_{u1}, and σ^2_{u2}, in addition to the residual variance σ^2_e.

QUESTION 2

We are going to be using data from the Framingham Heart Study to see if there are differences in heart rate between men and women over time, while considering how age, BMI, and cholesterol levels might also play a role.

I

Scatter plot matrix



From what i observed from the scatterplots, the residuals tend to move together, especially between Residuals.1 and Residuals.3. And again, the scatterplot between Residuals.2 and Residuals.3 shows a more scattered pattern, implying greater variability.

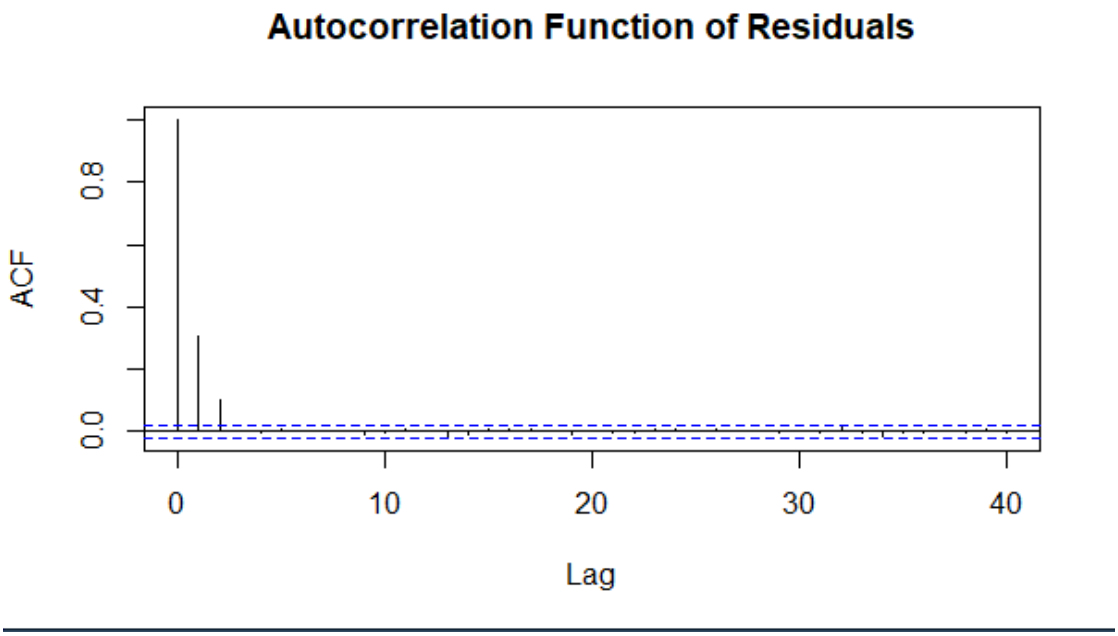
Pearson Correlation matrix

```
> cor(rDataWide[,-1],use="complete")
            residuals.1 residuals.3 residuals.2
residuals.1  1.00000000  0.09132489   0.3030304
residuals.3  0.09132489  1.00000000   0.2983122
residuals.2  0.30303039  0.29831218   1.0000000
```

From the output I got above, the correlation values between residuals measured at different times are generally positive, ranging from approximately 0.091 to 0.303 and they suggest that there is some degree of association between the residuals at different time points. But then the correlations are not extremely high, indicating that while there is some relationship between the residuals at different times, it is not exceptionally strong.

II

Autocorrelation function of Residuals



Interpretation

The ACF plot suggests that the model's residuals are white noise (random fluctuations) and do not exhibit any systematic patterns which is desirable because it indicates that the model adequately explains the data's temporal structure.

III

Presenting an appropriate model:

A marginal longitudinal model was used because it can actually handle longitudinal data and time-varying covariate, we can analyze repeated measures overtime, while we consider correlations between observations from the same individuals and incorporating the effects of Age, bmi and the total cholesterol**.**

Picking an appropriate mean model.

```
# Fit a linear mixed-effects model
model <- lmer(HEARTRTE ~ SEX + PERIOD + AGE + BMI + TOTCHOL + (1 | RANDID), data = df)
summary(model)
```

IV

After fitting in an appropriate model, we then try to find the Generalized estimating equations, we have four different correlation structure.

We have compound symmetry, auto regressive, unstructured and independent. After fitting in the models, I then went ahead to find the QIC, which is the Quasi-Likelihood under Independence criterion.

The way we interpret the QIC is similar to AIC and BIC. A smaller QIC means a better fit. So from my output below, the Independent structure has the smallest QIC value, so we say it is the best fit.

```
> QIC(GEEFHSModel_UN)
       QIC       QICu  Quasi Lik        CIC     params       QICC
1768506.93 1768501.41 -884243.70       9.76       7.00 1768506.98
> QIC(GEEFHSModel_AR1)
       QIC       QICu  Quasi Lik        CIC     params       QICC
 1.769e+06  1.769e+06 -8.844e+05  9.822e+00  7.000e+00  1.769e+06
> QIC(GEEFHSModel_EXCH)
       QIC       QICu  Quasi Lik        CIC     params       QICC
 1.768e+06  1.768e+06 -8.842e+05  9.805e+00  7.000e+00  1.768e+06
> QIC(GEEFHSModel_ind)
       QIC       QICu  Quasi Lik        CIC     params       QICC
1767171.13 1767164.36 -883575.18      10.38       7.00 1767171.15
>
```

Interpretations of Predictors:

PERIOD: The estimated coefficients for PERIOD indicate the change in heart rate across different measurement periods. Compared to the reference period, individuals have, on average:

1.294 units higher heart rate during period 2.

1.647 units higher heart rate during period 3.

SEX: The coefficient for SEX suggests that females have, on average, 2.470 units higher heart rate compared to males.

TOTCHOL: Each one-unit increase in total cholesterol levels is associated with a 0.017-unit increase in heart rate, on average.

AGE: The coefficient for AGE indicates a non-significant association between age and heart rate ($p = 0.28$). There is no significant change in heart rate associated with one-year increase in age.

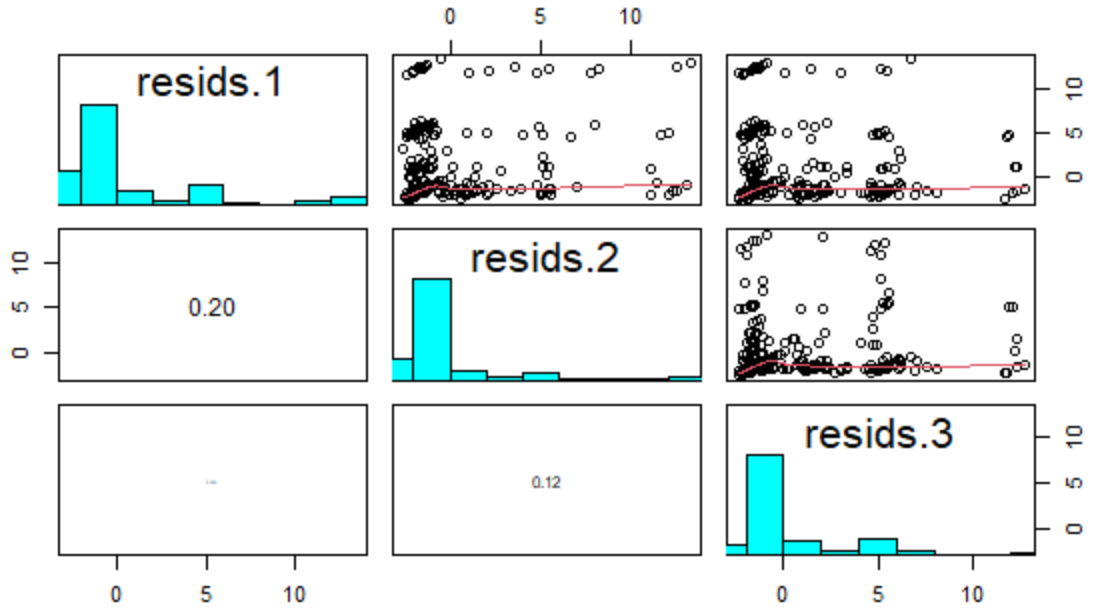
BMI: Each one-unit increase in BMI is associated with a 0.180-unit increase in heart rate, on average).

Answering the Question about Heart Rate:

- Heart rate is significantly influenced by several factors:
- It tends to increase across different measurement periods.
- Females have higher heart rates compared to males.
- Higher total cholesterol levels and BMI are associated with higher heart rates.
- However, there is no significant association between age and heart rate in this model.

QUESTION 3

I



Resids.1 vs. Resids.2: - The points exhibit a slight upward trend, indicating that the model may have underestimated "days sick" for some individuals in round 2.

Resids.1 vs. Resids.3: The points show a more pronounced upward trend, suggesting a consistent underestimation of "days sick" in round 3 compared to round 1.

Resids.2 vs. Resids.3: The points exhibit a similar upward trend, suggesting consistent underestimation in round 3.

Pearson Correlation

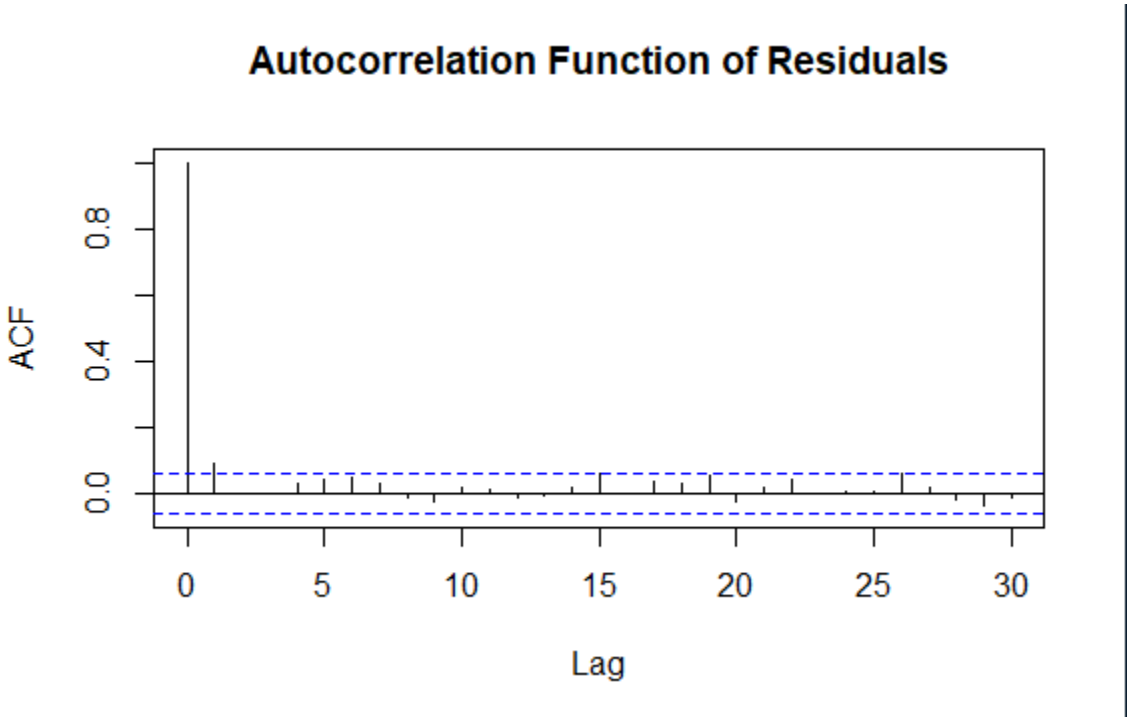
```
> cor(rDataWide[,-1],use="complete")
           resids.1  resids.2   resids.3
resids.1 1.00000000 0.2006418 0.02517439
resids.2 0.20064179 1.0000000 0.12277519
resids.3 0.02517439 0.1227752 1.00000000
```

Interpretation

The correlations are quite low, which implies that the residuals from different survey rounds are not strongly related.

II

Autocorrelation function



Interpretations

The spikes in the ACF plot represent the autocorrelations. Notably, all the spikes fall within the blue dashed lines (confidence intervals) and this suggests that the residuals are not significantly autocorrelated.

III

An appropriate model

Random Intercept Model:

```
> RIModel = lmer(days.sick~as.factor(survey.round)+gender+bmi+age+(1|childid),
+                data=child_data,REML=TRUE)
> summary(RIModel)
Linear mixed model fit by REML ['lmerMod']
Formula: days.sick ~ as.factor(survey.round) + gender + bmi + age + (1 |
    childid)
   Data: child_data

REML criterion at convergence: 5800.2

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.2659 -0.5015 -0.3557  0.0184  3.8820

Random effects:
 Groups   Name        Variance Std.Dev.
 childid  (Intercept) 1.261    1.123
 Residual             9.647    3.106
Number of obs: 1110, groups:  childid, 370

Fixed effects:
                            Estimate Std. Error t value
(Intercept)                 4.231569   1.315411   3.217
as.factor(survey.round)2 -0.576204   0.230135  -2.504
as.factor(survey.round)3  0.006968   0.233990   0.030
gender                      0.329977   0.221690   1.488
bmi                        -0.099689   0.078240  -1.274
age                        -0.023329   0.006611  -3.529

Correlation of Fixed Effects:
            (Intr) a.(.)2 a.(.)3 gender bmi
as.fctr(.)2 -0.066
as.fctr(.)3 -0.013  0.511
gender      -0.103  0.014  0.023
bmi         -0.971  0.003 -0.033  0.044
age         -0.415 -0.120 -0.217 -0.099  0.244
```

Fitted model:

Wit = 4.23 – 0.58survey.round2 + 0.0069survey.round3 +0.329gender – 0.099bmi – 0.023age

```
> RIModel
Linear mixed model fit by REML ['lmerMod']
Formula: days.sick ~ as.factor(survey.round) + gender + bmi + age + (1 |
    childid)
   Data: child_data
REML criterion at convergence: 5800.211
Random effects:
 Groups   Name        Std.Dev.
 childid  (Intercept) 1.123
 Residual             3.106
Number of obs: 1110, groups:  childid, 370
Fixed Effects:
             (Intercept)  as.factor(survey.round)2  as.factor(survey.round)3
                4.231569                 -0.576204                  0.006968
                  gender                       bmi                       age
                0.329977                 -0.099689                 -0.023329
>
```

Interpretations:

Intercept: The estimated intercept is 4.23, representing the expected number of sick days when all other predictors are zero.

as.factor(survey.round):

For survey round 2, there is a decrease in the expected number of sick days by 0.58 compared to round 1.

For survey round 3, there is a slight increase in the expected number of sick days by 0.007 compared to round 1.

gender: Females, on average, have a higher expected number of sick days (coefficient = 0.33) compared to males.

bmi: Each unit increase in BMI is associated with a slight decrease in the expected number of sick days (coefficient = -0.10).

age: Each year increase in age is associated with a slight decrease in the expected number of sick days (coefficient = -0.023).

Random Time Slope Modelling

```
> # RANDOM TIME (SLOPE) MODEL #
> RTModel = lmer(days.sick~as.factor(survey.round)+gender+bmi+age+(1+survey.round|childi
d)
+                ,data=child_data,REML=TRUE)
> summary(RTModel)
Linear mixed model fit by REML ['lmerMod']
Formula: days.sick ~ as.factor(survey.round) + gender + bmi + age + (1 +
    survey.round | childid)
   Data: child_data

REML criterion at convergence: 5773.7

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.4320 -0.4603 -0.3089  0.0287  4.1921

Random effects:
 Groups   Name         Variance Std.Dev. Corr
 childid  (Intercept)  13.113   3.621
          survey.round  1.891   1.375    -0.94
 Residual               7.755   2.785
Number of obs: 1110, groups:  childid, 370

Fixed effects:
                          Estimate Std. Error t value
(Intercept)               4.296865   1.294557   3.319
as.factor(survey.round)2 -0.571097   0.218683  -2.612
as.factor(survey.round)3  0.016047   0.254735   0.063
gender                    0.306745   0.218010   1.407
bmi                      -0.100223   0.076843  -1.304
age                      -0.024487   0.006504  -3.765

Correlation of Fixed Effects:
            (Intr) a.(.)2 a.(.)3 gender bmi
as.fctr(.)2 -0.080
```

Fitted effects:

Wit = 4.29 – 0.57*survey.round2 + 0.016*survey.round3 + 0.306*gender – 0.100*bmi – 0.024*age

```
> RTModel
Linear mixed model fit by REML ['lmerMod']
Formula: days.sick ~ as.factor(survey.round) + gender + bmi + age + (1 +
    survey.round | childid)
   Data: child_data
REML criterion at convergence: 5773.659
Random effects:
 Groups   Name         Std.Dev. Corr
 childid  (Intercept)  3.621
          survey.round 1.375    -0.94
 Residual              2.785
Number of obs: 1110, groups:  childid, 370
Fixed Effects:
            (Intercept)  as.factor(survey.round)2  as.factor(survey.round)3
                4.29686                   -0.57110                   0.01605
                 gender                        bmi                       age
                0.30675                   -0.10022                  -0.02449
>
```

Interpretations:

survey.round:

For survey round 2, there is a decrease in the expected number of sick days by 0.57 compared to round 1.

For survey round 3, there is a slight increase in the expected number of sick days by 0.016 compared to round 1.

gender: Females, on average, have a higher expected number of sick days (coefficient = 0.31) compared to males.

bmi: Each unit increase in BMI is associated with a slight decrease in the expected number of sick days (coefficient = -0.10).

age: Each year increase in age is associated with a slight decrease in the expected number of sick days (coefficient = -0.024).

Explaining why a conditional model is appropriate

We used a conditional longitudinal model because each child were measured multiple times, we can call this repeated measures and because of this we had a nested data structure. A conditional longitudinal model allows us to account for the dependency among observations within the same child and to examine how predictors such as survey round, gender, BMI, and age influence the outcome variable (number of sick days) over time while controlling for individual-specific differences.

IV

Evaluating the fit of the model using the AIC and BIC

```
> AIC(RIModel,RTModel)
        df      AIC
RIModel  8 5816.211
RTModel 10 5793.659
> BIC(RIModel,RTModel)
        df      BIC
RIModel  8 5856.308
RTModel 10 5843.780
```

Interpretation:

AIC and BIC are both measures for comparing models, with lower values indicating a better fit.

Both AIC and BIC favor the Random Time Slope Model (RTModel) over the Random Intercept Model (RIModel) for better fit to the data.

V

Appropriate interpretations of the predictors

Survey Round (Time):

The predictor as.factor(survey.round) represents the time at which each child was measured, with three levels corresponding to three survey rounds.

In both RIModel and RTModel, survey round 2 is associated with a decrease in the expected number of sick days compared to the reference level (survey round 1), while survey round 3 shows a slight increase. However, the differences are not statistically significant in some cases (e.g., RTModel: survey round 3).

BMI (Body Mass Index):

The predictor bmi represents the body mass index of each child.

In both models, there is a slight decrease in the expected number of sick days for each unit increase in BMI, holding other predictors constant. However, the effect is not statistically significant in some cases.

Gender:

The predictor gender represents the gender of each child. In both models, females tend to have a higher expected number of sick days compared to males, holding other predictors constant.

Age:

The predictor age represents the age of each child. There is a slight decrease in the expected number of sick days for each year increase in age, holding other predictors constant.

Research Question

A study on child illness looked at how factors like survey round (time), gender, BMI, and age affected sick days. Results showed a decrease in sick days from round 1 to 2, then a slight rise in round 3. These changes weren't always significant. Girls tended to have more sick days than boys, and higher BMI/age were linked to slightly fewer sick days, though significance

varied. Overall, the study suggests these factors influence sick days, with individual variations possible.