BIOS 6612 Final Exam - Spring 2021

You may refer to your course notes, homework assignments, labs, textbooks, and published papers only. You are not allowed to use your internet to search for potential solutions/approaches or to work with others. You may use R to check any work or answers, but you **must show all work.**

Answers without work will not receive partial credit. You will enter your responses via Canvas as a quiz that will be in the same order as the questions in the final. **To allow for time to transfer your answers, you will have until 4:30 pm to submit your answers.** Then you will have until 7:30 pm to upload your work (e.g., via images, Rmd files, etc.) to Canvas to be able to receive partial credit for grading.

Read the UCD Honor Code statement below. You must indicate your agreement with this statement when you submit your answers.

I understand that my participation in this examination and in all academic and professional activities as a UCD student is bound by the provisions of the UCD Honor Code. I understand that work on this exam is to be done independently and according to the guidelines provided.

On my honor, I have neither given nor received aid on this examination.

Problem 1 (20 points). Answer the following questions. Circle TRUE or FALSE: (4 points each).

1a. TRUE FALSE When outcome data is missing at random (MAR), GEEs will provide unbiased inference.

FALSE. GLMMS will

1b. TRUE FALSE Wald tests provide appropriate inference for GEE models.

TRUE

1c. TRUE FALSE GLMMs can be used to make inferences about individuals in a population.

TRUE

1d. TRUE FALSE In a generalized linear mixed effects model with a random slope and a random intercept, we assume that the random slope and the random intercept are independent.

FALSE. We assume the random effects and the error term are independent, but each random effect can be correlated with the others.

1e. TRUE FALSE The exponentiated coefficients in Poisson regression can be interpreted as rate ratios.

TRUE

Using the following data to answer Questions 2, 3 and 4.

A study was conducted to determine the effect of alcohol on blood glucose. For the study, blood glucose levels were measured at 10 time points for 7 volunteers given alcohol at time 0. The same experiment was repeated on another occasion with the same subjects but with a dietary additive administered to all subjects. There was no dropout or missing data for this study. The variables in the data set are:

- `Subject`: a factor with levels 1 to 7 identifying the subject whose glucose level is measured.
- `trt`: a factor with levels `control` and `additive` indicating whether the subject received treatment or control at that time point.
- `Time`: : a numeric vector giving the time since alcohol ingestion (in minutes/10).
- `glucose`: a numeric vector giving the blood glucose level (in mg/dl).

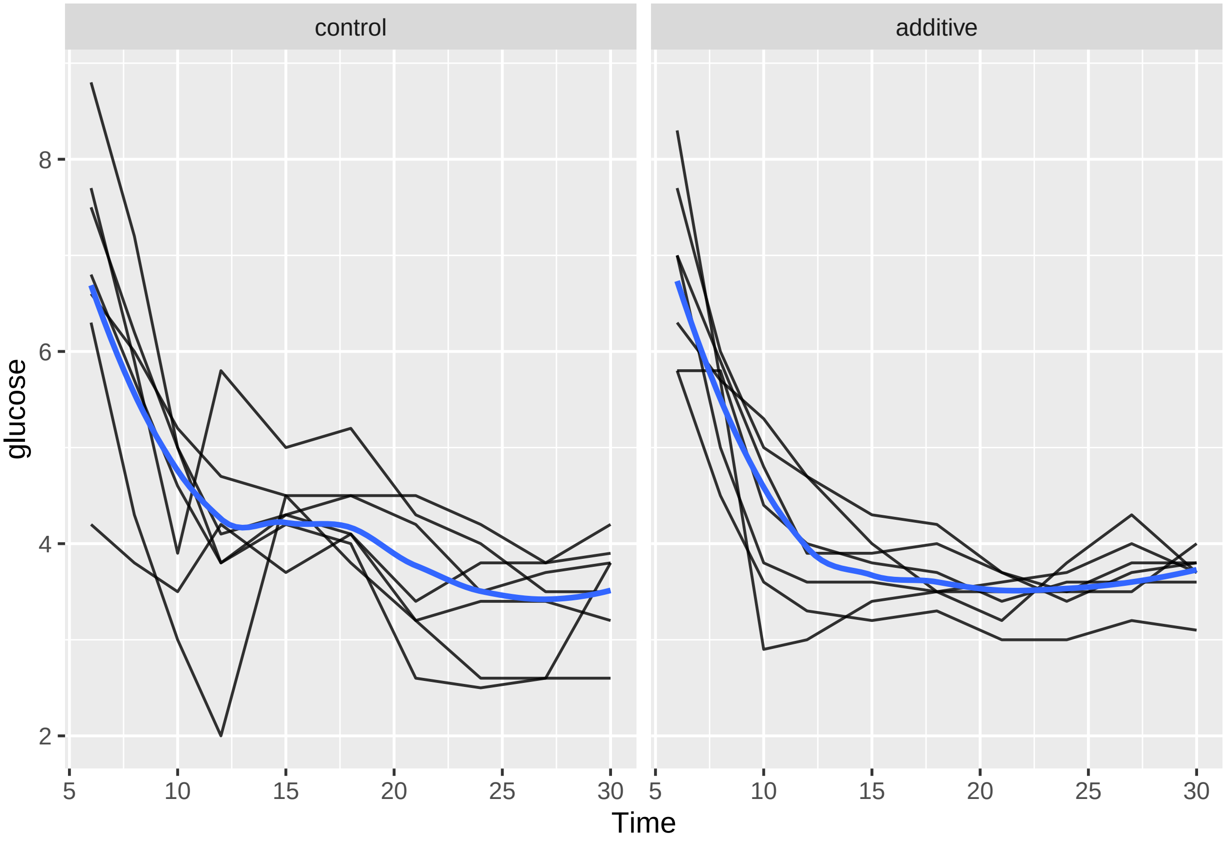
The data for Subject=1 is copied below:

```
##        trt  6   8  10  12  15  18  21  24  27  30
## 1  control 6.3 4.3 3.0 2.0 4.5 3.8 3.2 2.6 2.6 2.6
## 2 additive 5.8 4.5 3.6 3.3 3.2 3.3 3.0 3.0 3.2 3.1
```

Columns indicate time of measurement (in units of minutes/10); the two rows correspond to the two treatments. A spaghetti plot of the data, faceted by treatment, is given below.

```
glucose %>%
  ggplot(aes(Time, glucose)) +
  geom_line(alpha = 0.8, aes(group = Subject)) +
  geom_smooth(method = "loess", se = FALSE) +
  facet_wrap(~trt)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Problem 2 (30 points).

A linear mixed effects model was fitted to this data with a random intercept for `Subject` and fixed effects for `Time`, `trt`, and the interaction between `Time` and `trt`. This is Model 1; output appears in attached document.

```
model1 = lmer(glucose ~ (1 | Subject) + Time*trt,
              data = glucose)

summary(model1)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: glucose ~ (1 | Subject) + Time * trt
##    Data: glucose
##
## REML criterion at convergence: 391.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.6140 -0.6184 -0.0345  0.5053  3.5801
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Subject  (Intercept) 0.1049   0.3239
##  Residual             0.8167   0.9037
## Number of obs: 140, groups:  Subject, 7
##
## Fixed effects:
##                    Estimate Std. Error        df t value Pr(>|t|)
## (Intercept)         6.19087    0.28660  66.68610  21.601  < 2e-16 ***
## Time               -0.10799    0.01377 130.00000  -7.839 1.43e-12 ***
## trtadditive        -0.32766    0.36647 130.00000  -0.894    0.373
## Time:trtadditive    0.01223    0.01948 130.00000   0.628    0.531
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Correlation of Fixed Effects:
##             (Intr) Time   trtddt
## Time        -0.822
## trtadditive -0.639  0.643
## Tim:trtddtv  0.581 -0.707 -0.909
```

```
anova(model1)
```

```
## Type III Analysis of Variance Table with Satterthwaite’s method
##          Sum Sq Mean Sq NumDF DenDF  F value Pr(>F)
## Time     89.342  89.342     1   130 109.3917 <2e-16 ***
## trt       0.653   0.653     1   130   0.7994 0.3729
## Time:trt  0.322   0.322     1   130   0.3940 0.5313
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

2a. Calculate the estimated ICC for this data.

```
as_tibble(summary(model1)$varcor) %>%
  rename("variance" = vcov,
         "sd" = sdcor)
```

```
## # A tibble: 2 x 5
##   grp      var1        var2  variance    sd
##   <chr>    <chr>       <chr>    <dbl> <dbl>
## 1 Subject  (Intercept) <NA>     0.105 0.324
## 2 Residual <NA>        <NA>     0.817 0.904
```

```
round(0.1049/(0.1049 + 0.8167), 3)
```

```
## [1] 0.114
```

2b. Interpret your ICC estimate.

> ICC measures the proportion of total variability due to between-subject variance, so small value means this portion is relatively small; this also means that there is relatively little correlation within subjects

2c. Interpret the interaction between the `Time` and `trt` variables. Is this interaction significant? Provide a test statistic and p-value to support your conclusion.

> Interaction represents the difference in slopes of the mean of glucose with respect to time between treatment and control conditions. Interaction is not significant, F stat of 0.3940 (p=0.5313).

2d Based on trends observed in the spaghetti plot, what other term(s) might we consider adding to our model?

> The effect of glucose on time appears nonlinear. It might be valuable to consider adding a quadratic term for time. Alternatively, we can consider time as a categorical variable as we do in Model 2 below.

Problem 3 (30 points)

Model 2 is fitted to the data, exchanging the linear effect for `Time` with a categorical one. Output for this model, as well as a comparison between model 1 and model 2 is provided below.

```
model2= lmer(glucose ~ (1 | Subject) + factor(Time)*trt,
              data = glucose)

summary(model2)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: glucose ~ (1 | Subject) + factor(Time) * trt
##    Data: glucose
##
## REML criterion at convergence: 277.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.9225 -0.4441 -0.0113  0.4569  3.1057
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Subject  (Intercept) 0.1264   0.3555
##  Residual             0.3866   0.6218
## Number of obs: 140, groups:  Subject, 7
##
## Fixed effects:
##                                Estimate Std. Error        df t value Pr(>|t|)
## (Intercept)                   6.843e+00  2.707e-01  5.572e+01  25.276  < 2e-16
## factor(Time)8                -1.257e+00  3.324e-01  1.140e+02  -3.782 0.000249
## factor(Time)10               -2.529e+00  3.324e-01  1.140e+02  -7.608 8.67e-12
## factor(Time)12               -2.786e+00  3.324e-01  1.140e+02  -8.381 1.55e-13
## factor(Time)15               -2.486e+00  3.324e-01  1.140e+02  -7.479 1.68e-11
## factor(Time)18               -2.529e+00  3.324e-01  1.140e+02  -7.608 8.67e-12
## factor(Time)21               -3.214e+00  3.324e-01  1.140e+02  -9.671  < 2e-16
## factor(Time)24               -3.414e+00  3.324e-01  1.140e+02 -10.273  < 2e-16
## factor(Time)27               -3.500e+00  3.324e-01  1.140e+02 -10.531  < 2e-16
## factor(Time)30               -3.271e+00  3.324e-01  1.140e+02  -9.843  < 2e-16
## trtadditive                   7.735e-14  3.324e-01  1.140e+02   0.000 1.000000
## factor(Time)8:trtadditive    -7.143e-02  4.700e-01  1.140e+02  -0.152 0.879484
## factor(Time)10:trtadditive   -5.714e-02  4.700e-01  1.140e+02  -0.122 0.903453
## factor(Time)12:trtadditive -1.714e-01  4.700e-01  1.140e+02  -0.365 0.716002
## factor(Time)15:trtadditive -6.143e-01  4.700e-01  1.140e+02  -1.307 0.193882
## factor(Time)18:trtadditive -6.429e-01  4.700e-01  1.140e+02  -1.368 0.174106
## factor(Time)21:trtadditive -1.857e-01  4.700e-01  1.140e+02  -0.395 0.693504
## factor(Time)24:trtadditive  7.143e-02  4.700e-01  1.140e+02   0.152 0.879484
## factor(Time)27:trtadditive  3.857e-01  4.700e-01  1.140e+02   0.821 0.413586
## factor(Time)30:trtadditive  1.000e-01  4.700e-01  1.140e+02   0.213 0.831903
##
## (Intercept)                ***
## factor(Time)8              ***
## factor(Time)10             ***
## factor(Time)12             ***
## factor(Time)15             ***
## factor(Time)18             ***
## factor(Time)21             ***
## factor(Time)24             ***
## factor(Time)27             ***
## factor(Time)30             ***
## trtadditive
## factor(Time)8:trtadditive
## factor(Time)10:trtadditive
## factor(Time)12:trtadditive
## factor(Time)15:trtadditive
## factor(Time)18:trtadditive
## factor(Time)21:trtadditive
## factor(Time)24:trtadditive
## factor(Time)27:trtadditive
## factor(Time)30:trtadditive
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
##
## Correlation matrix not shown by default, as p = 20 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
```

```
AIC(model1, model2)
```

```
##        df      AIC
## model1  6 403.2820
## model2 22 321.5519
```

3a Should `Time` be treated as categorical (model 2) or continuous (model 1)? Justify your answer.

> The model with Time as a categorical effect has a lower AIC, so that is the preferred model.

3b. Using Model 2, what is the average glucose level for subjects at `trt= "additive"` and `Time=15`?

```
# solution:
# intercept
6.843 +
  # time effect at 15
  -2.486 +
  # treatment effect
   7.735e-14  +
  # interaction
  -.6143
```

```
## [1] 3.7427
```

3c What additional information would you need to be able to construct a 95% confidence interval for the estimate in 3b?

> Would also need covariance matrix for beta hats and t critical value

Model 3 is fitted to the data, removing the interaction between `Time` and `trt`, but otherwise the same as Model 2. Output from an anova comparing models 2 and 3 is provided below.

```
model3= lmer(glucose ~ (1 | Subject) + factor(Time) + trt,
              data = glucose)

anova(model2, model3)
```

```
## refitting model(s) with ML (instead of REML)
```

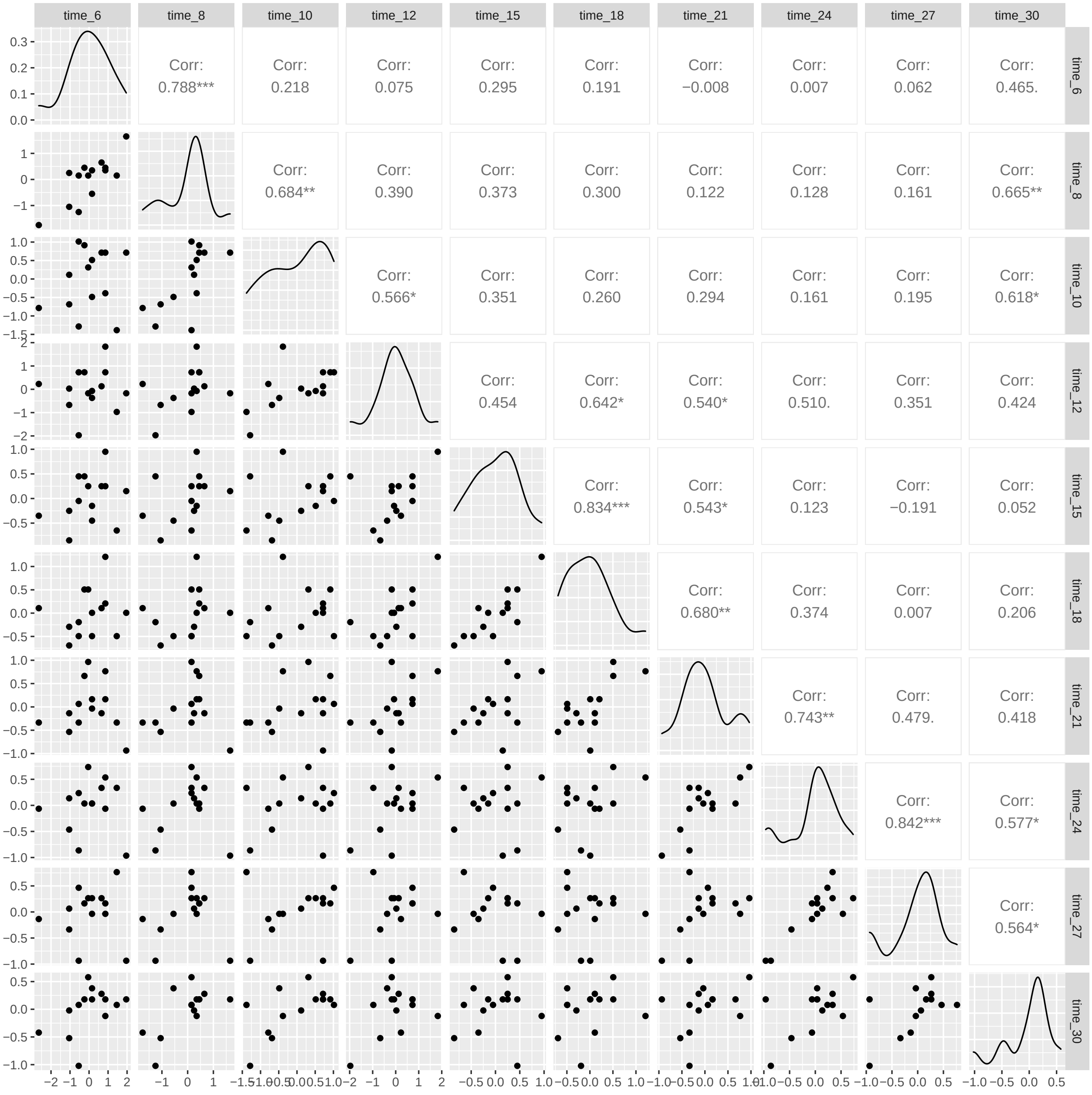
```
## Data: glucose
## Models:
## model3: glucose ~ (1 | Subject) + factor(Time) + trt
## model2: glucose ~ (1 | Subject) + factor(Time) * trt
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model3   13 291.87 330.11 -132.93   265.87
## model2   22 300.82 365.54 -128.41   256.82 9.0441  9     0.4332
```

3d Would you choose model 2 or 3? Is anova appropriate here? Justify your answer.

> Model2 without the interaction is better based on both AIC and anova. Anova is approprate but models should be refit using ML instead of REML.

Problem 4 (20 points)

You also decide to fit a GEE model to this data. Additional exploratory data analysis reveals the following patterns in your correlation.



4a From the plot above, what do you infer about the correlation structure of your data?

Time points closer together are more correlated, and that drops off quickly. After that there is not much consistent structure in the correlation.

A GEE is fitted to the data (model 4), using an exchangeable working correlation structure and the same mean structure as Model 3 (`Time` as categorical, and `trt`, but no interaction term).

```r
model4 = geeglm(glucose ~ factor(Time) + trt,
                id = Subject,
                data = glucose,
                family = gaussian,
                corstr = "exchangeable")

summary(model4)
```

```
##
## Call:
## geeglm(formula = glucose ~ factor(Time) + trt, family = gaussian,
##     data = glucose, id = Subject, corstr = "exchangeable")
##
##  Coefficients:
##                 Estimate Std.err    Wald Pr(>|W|)
## (Intercept)       6.9021  0.3527 382.872  < 2e-16 ***
## factor(Time)8    -1.2929  0.1853  48.684 3.01e-12 ***
## factor(Time)10   -2.5571  0.3281  60.758 6.44e-15 ***
## factor(Time)12   -2.8714  0.3653  61.802 3.77e-15 ***
## factor(Time)15   -2.7929  0.2895  93.054  < 2e-16 ***
## factor(Time)18   -2.8500  0.3042  87.787  < 2e-16 ***
## factor(Time)21   -3.3071  0.3303 100.223  < 2e-16 ***
## factor(Time)24   -3.3786  0.3245 108.377  < 2e-16 ***
## factor(Time)27   -3.3071  0.3176 108.435  < 2e-16 ***
## factor(Time)30   -3.2214  0.2682 144.219  < 2e-16 ***
## trtadditive      -0.1186  0.2228   0.283    0.595
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##              Estimate Std.err
## (Intercept)   0.4619  0.1018
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha   0.3066 0.09962
## Number of clusters:   14  Maximum cluster size: 10
```

```r
anova(model4)
```

```
## Analysis of ’Wald statistic’ Table
## Model: gaussian, link: identity
## Response: glucose
## Terms added sequentially (first to last)
##
##              Df    X2 P(>|Chi|)
## factor(Time)  9 273.8    <2e-16 ***
## trt           1   0.3      0.59
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4b What is the estimated correlation coefficient for model 4?

```
summary(model4)$corr$Estimate
```

```
## [1] 0.3066
```

4c Are the ideal conditions for the sandwich estimator satisfied? Why or why not?

> Somewhat but not fully. Data is balanced and without missingness, but there are relatively few subjects compared to the number of data points per subject. We might need to be careful about choosing our working correlation structure and choose a few more.