BIOS 6612 HW 6 Solutions: Generalized Estimating Equations

The data for this assignment comes from a clinical trial of 59 epileptics. For each patient, the number of epileptic seizures was recorded during a baseline period of 8 weeks. Number of seizures was then recorded in 4 consecutive 2-week intervals.

The data is referenced in Lecture 12.3 and given as part of `geepack` package.

- Data set is in wide format (one row per subject)

Each row of the data set contains the following variables:

- `y1`: the outcome (number of seizures) at the first time point
- `y2`: number of seizures at the second time point
- `y3`: number of seizures at the third time point
- `y4`: number of seizures at the fourth time point
- `base`: number of seizures at baseline
- `trt`: the treatment (0 = placebo, 1 = progabide)
- `age`: in years, at baseline

You can load the data into `R` by loading the `geepack` library and calling `data(seizure)`.

Exercise 1: Exploratory Data Analysis (40 points)

1a. (10 pts) Load the data and convert it from wide to long format. Is this data well-balanced? Why or why not?

First I load the data, and convert it from wide format to long format.

- The wide form of the data does not have a unique subject identifier, so we need to create one
- I also factor the treatment variable so placebo is the reference category
- Convert the time variable to numeric

```
data(seizure)

seizure_long = seizure %>%
  mutate(id = row_number(), # data is in wide format and we don't have an id variable
         trt = factor(trt, levels = 0:1, labels = c("placebo", "progabide"))) %>%
  pivot_longer(
    cols = y1:y4,
    names_to = "time",
    values_to = "num_seizures"
  ) %>%
  mutate(time = as.numeric(str_remove(time, "y")))

length(unique(seizure_long$id)) # number of subjects
```

```
## [1] 59
```

```
table(seizure_long$time) # number of subjects per visit
```

```
##
##  1  2  3  4
## 59 59 59 59
```

The data is well balanced, with 59 subjects and 4 visits per subject. There are no missing values.

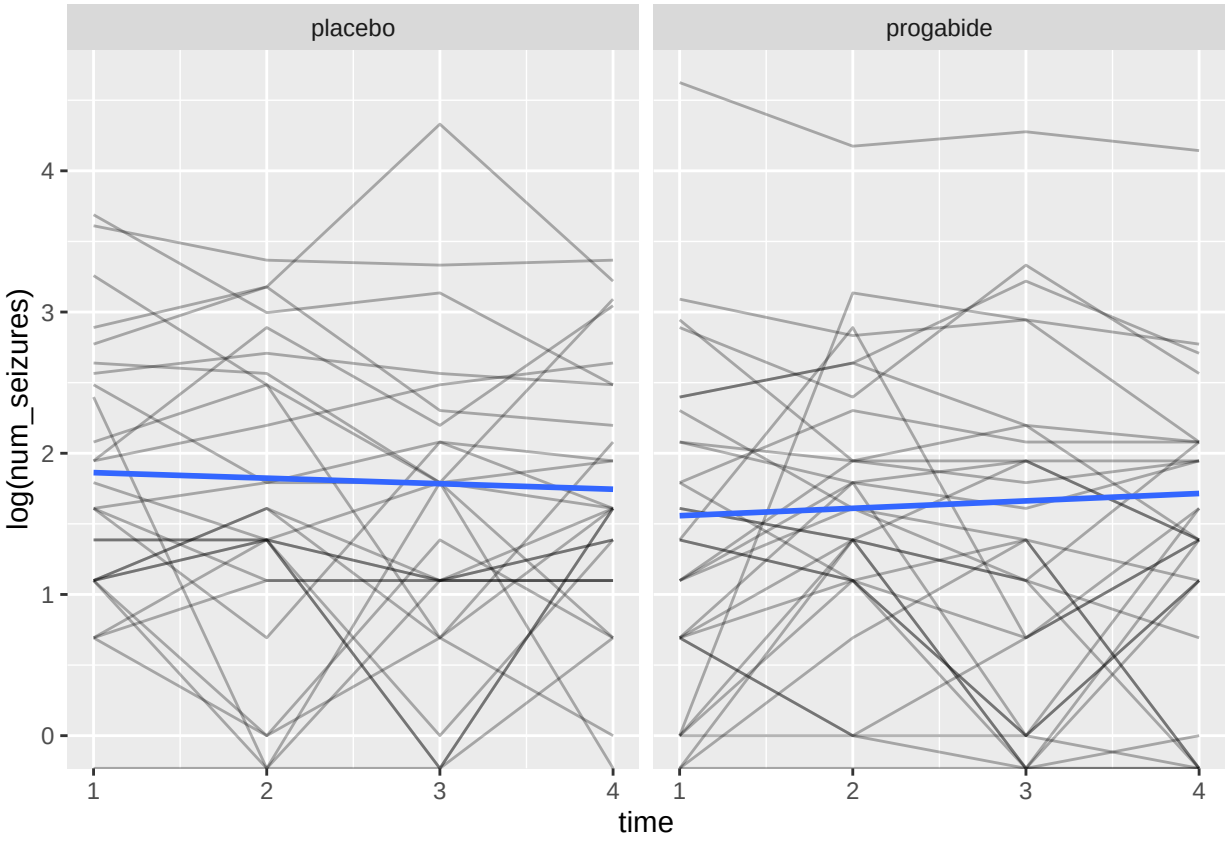
1b. (15 pts) Construct plot(s) to explore the mean structure of your outcome over time, and relationships between the outcome and covariates. Interpret trends that you see.

Below I create plots to explore the mean structure of the data. First are spaghetti plots of the log number of seizures over time, faceted by treatment.

```
seizure_long %>%
  ggplot(aes(time, log(num_seizures))) +
  geom_line(alpha = 0.3,  aes(group = id) )+
  facet_wrap(~ trt) +
  geom_smooth(method = "lm", se = FALSE)
```

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

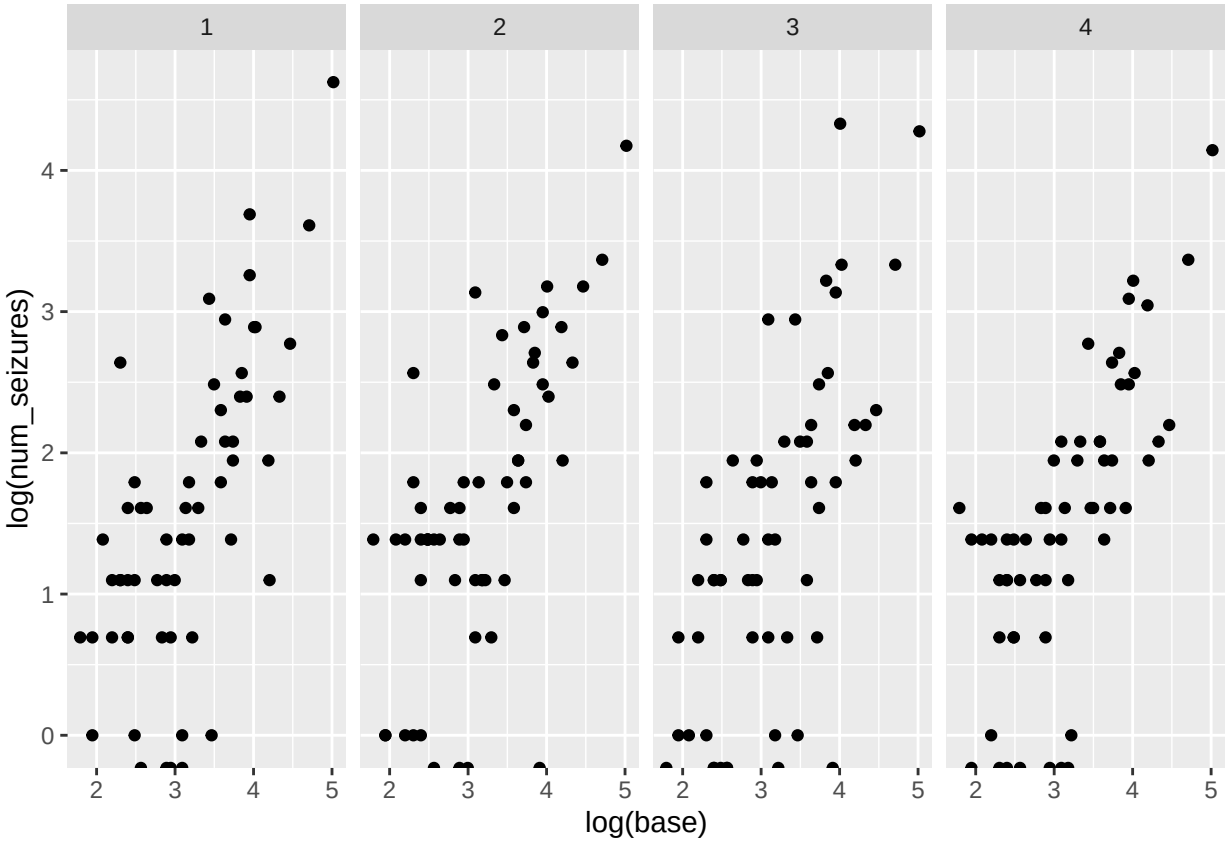
```
## Warning: Removed 23 rows containing non-finite values (stat_smooth).
```



There don't appear to be large differences in number of seizures across treatment groups over time.

I also created a scatterplot of baseline log number of seizures and log number of seizures at each time point. At each time point there appear to be a relationship between number of seizures at baseline and at each time point. This indicates that we should probably control for baseline number of seizures in our model(s).

```
seizure_long %>%
  ggplot(aes(log(base), log(num_seizures))) +
  geom_point() +
  facet_wrap(~time, nrow = 1)
```

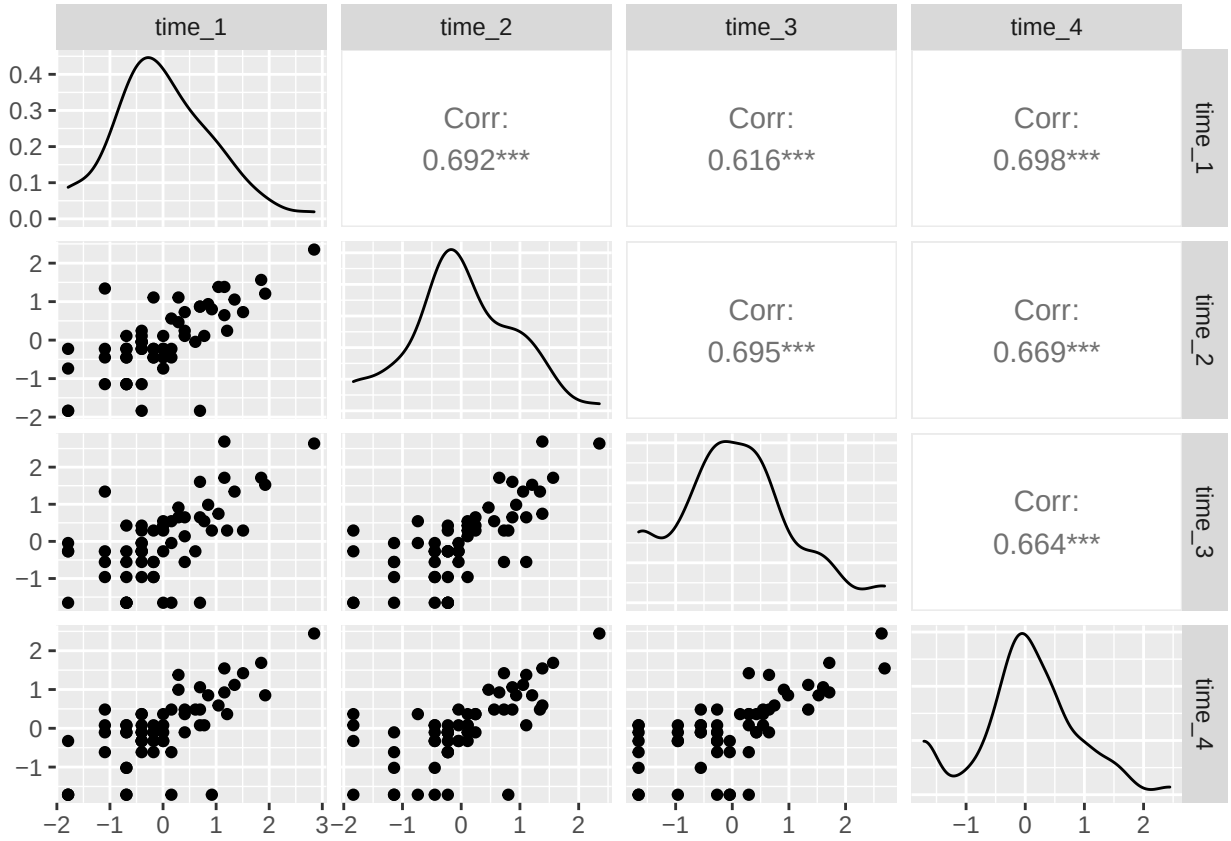


1c. (15 pts) Construct plot(s) to explore the correlation structure of your data. Interpret trends that you see. What might be a good choice of working correlation structure?

- Because we have count data, to calculate residuals I calculate the mean of the log transformed data

```
seizure_residuals = seizure_long %>%
  group_by(time) %>%
  mutate(mean_over_time = mean(log(num_seizures + 1))) %>%
  ungroup() %>%
  mutate(residuals = log(num_seizures + 1) - mean_over_time) %>%
  select(id, time, residuals)

seizure_residuals_wide = seizure_residuals %>%
  pivot_wider(names_from = time,
              names_glue = paste0("time_", "{time}"),
              values_from = residuals) %>%
```

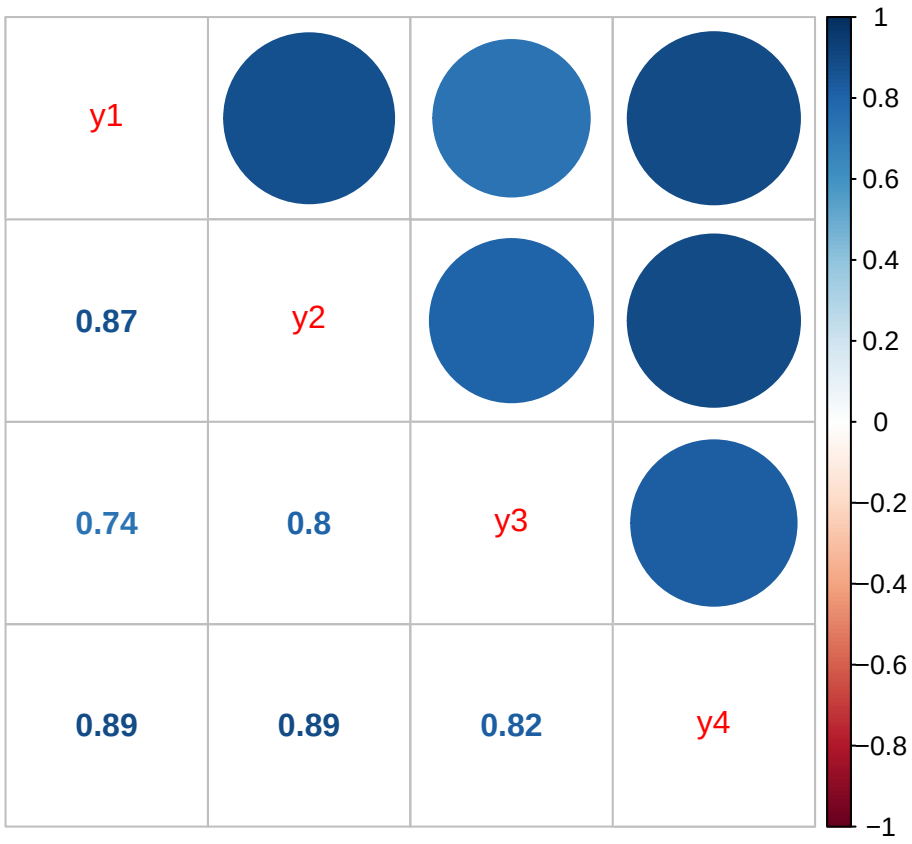
```
  select(-id)

ggpairs(seizure_residuals_wide)
```



There is a moderately high correlation within a subject at all time points, that is similar at each time point. This indicates that an exchangeable working correlation structure is probably appropriate.

Another plot I got have made (that looks at correlation on the raw data instead of correlation among residuals). The correlation levels are different, but conclusions about correlation between time points are similar.

```
seizure %>%
  select(starts_with("y")) %>%
  cor() %>%
  corrplot.mixed()
```



Exercise 2: GEE Model for Count Data (60 points)

2a. (15 pts) Your clinical collaborators are interested in a Poisson GEE model with treatment, age, and time as predictors. They also want to control for number of seizures at baseline. Write out the form of the GEE model. Assume an exchangeable correlation structure.

$$\log\left[E(Y_{ij})\right] = \beta_0 + \beta_1 trt_{ij} + \beta_2 age_{ij} + \beta_3 time_{ij} + \beta_4 base_{ij}$$

- link function is $\log(\cdot)$
- Variance is Poisson variance with a scale parameter ϕ for overdispersion, $Var(Y_{ij}) = \phi E(Y_{ij}) = \mu_{ij}$
  - By default, `geepack` fits a scale parameter to the data.
- Working correlation structure is exchangeable correlation

2b. (15 pts) Fit this model in `R` using the `geepack` package. Provide a table of estimates, standard errors, and P-values. What is the estimated correlation coefficient for your model?

```
seizure_ex = geeglm(num_seizures ~ base + trt + age + time,
                    id = id,
                    data = seizure_long,
                    family = poisson,
                    #scale.fix = TRUE, # specify this for a Poisson model with no overdispersion
                    corstr = "exchangeable")

summary(seizure_ex)
```

```
##
## Call:
## geeglm(formula = num_seizures ~ base + trt + age + time, family = poisson,
##     data = seizure_long, id = id, corstr = "exchangeable")
##
##  Coefficients:
##               Estimate   Std.err    Wald Pr(>|W|)
## (Intercept)   0.810980  0.310429   6.825  0.00899 **
## base          0.022692  0.001256 326.386  < 2e-16 ***
## trtprogabide -0.181357  0.167964   1.166  0.28026
## age           0.019063  0.010040   3.605  0.05759 .
## time         -0.057430  0.034976   2.696  0.10060
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)    5.011   1.612
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha   0.4045 0.07142
## Number of clusters:   59  Maximum cluster size: 4
```

2c. (5 pts) Is this an appropriate model for your data? Should you consider other correlation structures?

The data is perfectly balanced, with few data points per subject relative to number of subjects. These are ideal conditions for the sandwich estimator, so our inference should be robust to mispecification of the working correlation structures.

Another good way to check this is to fit a model with a different correlation structure (independence correlation structure for example) and see if the estimates/standard error/inference substantially changes. If it stays about the same, the sandwich estimator is providing robust estimates.

2d. (15 pts) Provide exponentiated coefficients for your model. Interpret the coefficients of your model on the exponentiated scale.

```
broom::tidy(seizure_ex, exp = TRUE, conf.int = TRUE) %>%
  select(-std.error, -statistic) %>%
  mutate(p.value = format.pval(p.value)) %>%
  knitr::kable(digits = 2)
```

term	estimate	p.value	conf.low	conf.high
(Intercept)	2.25	0.009	1.22	4.13
base	1.02	<2e-16	1.02	1.03
trtprogabide	0.83	0.280	0.60	1.16
age	1.02	0.058	1.00	1.04
time	0.94	0.101	0.88	1.01

The exponentiated coefficients are rate ratios! Only baseline is significant at the 5% level, but all parameters

are interpreted below.

- e^{β_0}: This would be the rate of seizures when all other covariates have a value of 0. In this case the intercept is not interpretable because there are no subjects in the study near age 0 or with close to 0 seizures at baseline.
- $e^{\beta_{base}} = 1.02$: The expected change in rate ratio of seizures associated with a one-seizure increase in number of seizures at baseline is 1.02, controlling for age, treatment, and time period.
- $e^{\beta_{trt}} = 0.834$: The rate of seizures for individuals who received progabide treatment is 0.834 time the rate of seizures for those who received placebo, controlling for age, sex, and baseline number of seizures.
- $e^{\beta_{age}} = 1.02$: The expected change in rate ratio of seizures associated with a one-year increase in age is 1.02, controlling for number of seizures at baseline, treatment, and time period.
- $e^{\beta_{time}} = 0.94$: The expected change in rate ratio of seizures associated with a one-unit increase in time is 0.94, controlling for number of seizures at baseline, treatment, and age.