Homework 3

Denis Ostroushko

Question 1

There were 4 rows of data where MS levels were not measured. We remove these observations from the data.

1 - A

Our linear predictor which will allow you to assess whether the effect of treatment changes over time is given below. We consider an interaction between treatment assignment variable and time to find if changes of MS level over time are different for treatment and control groups. Figure 1 shows that the average trend over time might be different for the two groups.

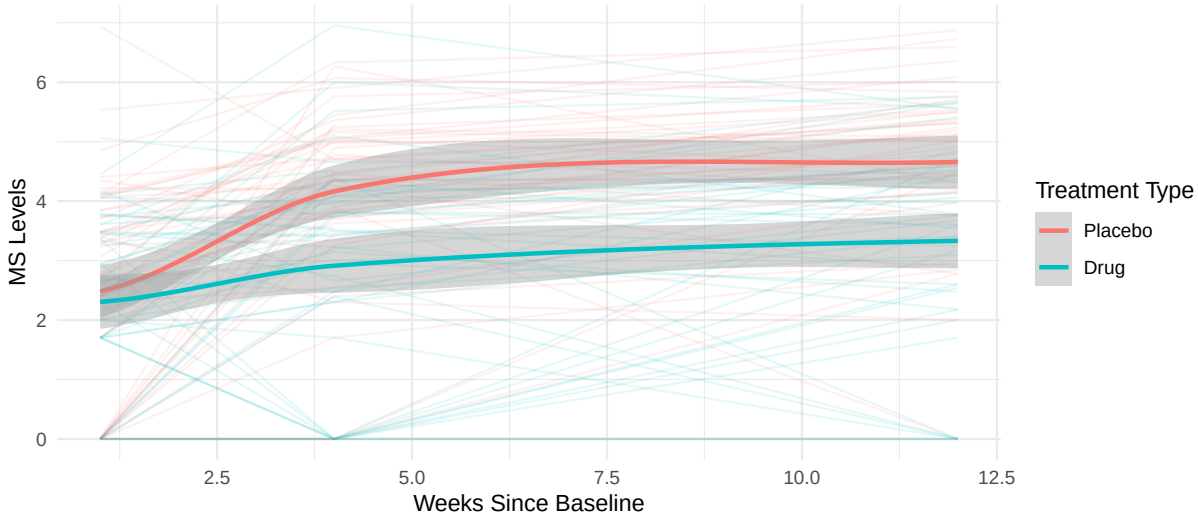

Figure 1: Average MS levels over time

Linear predictor:

$$E[MS_{ij}] = \beta_0 + \beta_1\ Treatment_{ij} + \beta_2\ Time\ in\ Weeks_{ij} + \beta_3 Treatment_{ij} \times Weeks_{ij}$$

1 - B

(i)

Table 1 displays coefficients for the effect of time and treatment on MS levels.

Table 1: 1B - Gaussian GLM Model Estimates

Predictor Term	Estiamte	CI	P-value
(Intercept)	2.81	(2.39, 3.22)	0.00
time_weeks	0.17	(0.11, 0.23)	0.00
trtgroupDrug	-0.44	(-1.02, 0.15)	0.14
time_weeks:trtgroupDrug	-0.09	(-0.17, 0)	0.04

- As time since baseline increases by 1 week, the average level of MS in the Placebo group increased by 0.17 units, after adjusting for other predictors. Effect is bounded by the (0.11, 0.23) confidence interval, suggesting a statistically significant effect of time on the average levels of MS in the Placebo group.
  - Figure 1 suggests that the linear effect might not be the most appropriate for these data. Average MS levels rise fast and then stay stable for the Placebo group, so we way want to consider non-linear terms for a more refined statistical model
- Estimated coefficient for interaction between time effect and treatment group is -0.09 with p-value < 0.05. Therefore, there is strong statistical evidence that the effect of time on MS levels is different between the two groups. With each additional week in the treated group, on average, MS levels increased by 0.17-0.09 = 0.08 units, after adjusting for other variables. The effect is bounded by the (0.02, 0.14) 95% confidence interval

(ii)

Obtained coefficient estimates are not trustworthy. Given that we correctly specified a mean function for the linear predictor, our estimates will be consistent and unbiased estimate for the true effects of covariates on predictors. Visual exploration showed that the we might not be specifying all model terms correctly, as there is some visual evidence that over time MS levels for the Placebo group change in a non-linear way. Obtained coefficients β_i are consistent estimates for the linear effect of time on the average MS levels, but the true effect is likely a non-linear function.

(iii)

We have cluster variant and invariant predictors. Consider an individual to be a cluster in these data.

1. Treatment group effect: this is a cluster invariant variable, therefore, we should expect to see a higher variance estimate for this coefficient, when accounting for the correlation structure present in the data.
2. Time effect for placebo and treated subjects is cluster variant comparative, therefore, when accounting for the correlation structure, we should expect variance estimates to be smaller for these two model estimates.

(iv)

using a robust method for variance estimation will not make model output more close to reality because Gaussian GLM still does not account for the correlation in the data.

1 - C

(i)

Table 2 summarizes model estimates using four types of working correlation.

1. In terms of strength and significance of effects, all four model agree on significance levels of all four predictors. All four models seem to agree that the main effect of treatment is strong, but not statistically significant at the 95% significance level, however, all models agree that the change in MS levels over time is statistically different for the two groups.
2. Overall, all four types of correlation structures produce very similar results, I suspect that with a 101 unique clusters in the data, we have a large enough sample size that variance estimator can correct for any differences imposed by the correlation structures.
3. When fitting a Gaussian family model with an identity link, we expect model point estimates for Gaussian GLM and Gaussian GEE with independence or exchangeable correlation structure to match. In our case, we observed this behavior, with a minor caveat that due to data incompleteness caused by missingness, exchangeable and independence correlation structures have small differences in the point estimates and variance estimates.
4. Time measured in weeks is a cluster variant covariate, and each of the four models correctly estimates smaller SE for this predictor when compared with the Gaussian GLM in Table 1

Table 2: Comparison of Correlation Structure Impact

Predictor Term	Estiamte	CI	P-value
Independece Correlation			
(Intercept)	2.81	(2.36, 3.25)	0.00
time_weeks	0.17	(0.13, 0.21)	0.00
trtgroupDrug	-0.44	(-1.03, 0.16)	0.15
time_weeks:trtgroupDrug	-0.09	(-0.14, -0.03)	0.00
Exchangeable Correlation			
(Intercept)	2.81	(2.37, 3.25)	0.00
time_weeks	0.17	(0.13, 0.21)	0.00
trtgroupDrug	-0.45	(-1.05, 0.15)	0.14
time_weeks:trtgroupDrug	-0.08	(-0.14, -0.02)	0.01
AR-1 Correlation			
(Intercept)	2.78	(2.34, 3.23)	0.00
time_weeks	0.15	(0.11, 0.19)	0.00
trtgroupDrug	-0.43	(-1.03, 0.17)	0.16
time_weeks:trtgroupDrug	-0.07	(-0.13, -0.01)	0.02
Unstructured Correlation			
(Intercept)	2.90	(2.47, 3.34)	0.00
time_weeks	0.14	(0.1, 0.18)	0.00
trtgroupDrug	-0.51	(-1.11, 0.09)	0.09
time_weeks:trtgroupDrug	-0.06	(-0.12, 0)	0.04

5. Treatment group is a cluster invariant covariate and each model should estimate a higher SE for this term. However, it looks like the results are pretty similar between Gaussian GLM from (1-b) and the four GEE models. I write this off as a coincidence.

(ii)

Results for exchangeable and independence correlation structures would be identical when the data are balanced and compete. In the case with these data, there are 4 people with only 2 observation and not 3. Therefore, there is a small degree of imbalance and incompleteness in the data, which results in a small discrepancy between the estimates from the independence and exchangeable correlation structures models.

For comparison, I imputed missing/incomplete values of MS values with average MS value for a given time point and group combination. This was done for the purpose of creating a complete data set. Table 3 shows the results of running GEE models on a data set with equal number of observations in each cluster of data, therefore creating a balanced and complete

Table 3: Independece and Exchangeable correlation structures with imputed means of MS to create complete data

Predictor Term	Estiamte	CI	P-value
Independece Correlation			
(Intercept)	2.81	(2.36, 3.25)	0.00
time_weeks	0.17	(0.13, 0.21)	0.00
trtgroupDrug	-0.44	(-1.04, 0.16)	0.15
time_weeks:trtgroupDrug	-0.08	(-0.14, -0.03)	0.00
Exchangeable Correlation			
(Intercept)	2.81	(2.36, 3.25)	0.00
time_weeks	0.17	(0.13, 0.21)	0.00
trtgroupDrug	-0.44	(-1.04, 0.16)	0.15
time_weeks:trtgroupDrug	-0.08	(-0.14, -0.03)	0.00

data set. Results of independence and exchangeable correlation structure-based models are identification, further showing that the reason for the difference in our original data set is due to data structures of clusters.

(iii)

- As time since baseline increases by 1 week, the average level of MS in the Placebo group increased by 0.17 units, after adjusting for other predictors. Effect is bounded by the (0.13, 0.21) 95% confidence interval, suggesting a statistically significant effect of time on the average levels of MS in the Placebo group.
- Estimated coefficient for interaction between time effect and treatment group is -0.08 with p-value < 0.05. Therefore, there is strong statistical evidence that the effect of time on MS levels is different between the two groups. With each additional week in the treated group, on average, MS levels increased by 0.17-0.08 = 0.09 units, after adjusting for other variables. The effect is bounded by the (0.05,0.13) 95% confidence interval.

1 - D

(i)

Average MS levels at week one for a subject in the treatment group is given in terms of β coefficients:

$$E[MS_{ij}] = \beta_0 + \beta_1 + \beta_2 + \beta_3$$

Or alternatively,

$$E[MS_{ij}] = 2.81 + 0.17 - 0.45 - 0.08 = 2.45$$

Variance of this point estimate is a complicated function of numerous covariance terms, therefore, we use `emmeans` to obtain an asymptotic 95% confidence interval

Table 4: Estiamted average MS for treated group one week post-baseline

Point Estimate	95% Confidence Interval
2.45	(2.07, 2.83)

(ii)

Similarly, average MS levels at week one for a subject in the placebo group is given in terms of β coefficients:

$$E[MS_{ij}] = \beta_0 + 4 * \beta_1$$

Or alternatively,

$$E[MS_{ij}] = 2.81 + 4 * 0.17 = 3.49$$

Variance of this point estimate is a complicated function of numerous covariance terms, therefore, we use `emmeans` to obtain an asymptotic 95% confidence interval

Table 5: Estiamted average MS for placebo group four weeks post-baseline

Point Estimate	95% Confidence Interval
3.48	(3.12, 3.85)

1 - E

(i)

$$E[MS_{ij}] = \beta_0 + \beta_1 I(Time_{ij} = 4) + \beta_2 I(Time_{ij} = 12) + \beta_3 I(Treatment_{ij} = "Drug") +$$
$$\beta_4 I(Time_{ij} = 4) \times I(Treatment_{ij} = "Drug") + \beta_5 I(Time_{ij} = 12) \times I(Treatment_{ij} = "Drug")$$

(ii)

A

β_3 expresses the difference between treatment and control groups at one week. Therefore, given a regression equation stated above, we can develop a hypothesis to test the effect of treatment at different follow up time points:

1. Null hypothesis: no difference between treatment assignment groups at every time point, i.e. $\beta_3 = \beta_4 = \beta_5 = 0$
2. Alternative hypothesis: not all β_i for $i = 3, 4, 5$ are equal to zero, i.e. average MS levels differ between the two groups at some time points.

B

Baseline to Four weeks change

For this question, we assume that measurements at week one are the baseline levels: Therefore average MS levels for treated at baseline is given by:

$$E[MS_{i1}|Week = 1, Group = Drug] = \beta_0 + \beta_3$$

Average MS levels at 4 weeks for treated are:

$$E[MS_{i4}|Week = 4, Group = Drug] = \beta_0 + \beta_1 + \beta_3 + \beta_4$$

Therefore, average change for treated is stated as

$$E[MS_{i4}|Week = 4, Group = Drug] - E[MS_{i1}|Week = 1, Group = Drug] = \beta_1 + \beta_4$$

Similarly, we estimate average MS levels for placebo group at baseline as

$$E[MS_{i1}|Week = 1, Group = Placebo] = \beta_0$$

Average MS levels at 4 weeks for placebo are:

$$E[MS_{i4}|Week = 4, Group = Placebo] = \beta_0 + \beta_1$$

Therefore, average change for treated is stated as

$$E[MS_{i4}|Week = 4, Group = Placebo] - E[MS_{i1}|Week = 1, Group = Placebo] = \beta_1$$

Therefore, we can state our hypothesis:

1. Null: there is no difference in average change from baseline to four weeks between treated and placebo groups, i.e. $\beta_4 = 0$
2. Alternative: There is some difference in mean change from baseline to four weeks between the two groups, i.e. $\beta_4 \neq 0$

Four to Twelve Week Change

Average MS values at 12 weeks for treated are given by:

$$E[MS_{i12}|Week = 12, Group = Drug] = \beta_0 + \beta_2 + \beta_3 + \beta_5$$

Average difference between treated at twelve weeks and four weeks is:

$$E[MS_{i12}|Week = 12, Group = Drug] - E[MS_{i4}|Week = 4, Group = Drug] = \\ (\beta_0 + \beta_2 + \beta_3 + \beta_5) - (\beta_0 + \beta_1 + \beta_3 + \beta_4) = \beta_2 + \beta_5 - \beta_1 - \beta_4$$

Average MS value at 12 weeks for placebo group is given by:

$$E[MS_{i12}|Week = 12, Group = Placebo] = \beta_0 + \beta_2$$

Average difference between placebo at twelve weeks and four weeks is:

$$E[MS_{i12}|Week = 12, Group = Placebo] - E[MS_{i4}|Week = 4, Group = Placebo] = \\ (\beta_0 + \beta_2) - (\beta_0 + \beta_1) = \beta_2 - \beta_1$$

The difference in two average changes is: $\beta_2 + \beta_5 - \beta_1 - \beta_4 - \beta_2 + \beta_1 = \beta_5 - \beta_4$

Therefore, we can state our hypothesis:

1. Null: there is no difference in average change from four to twelve weeks between treated and placebo groups, i.e. $\beta_5 - \beta_4 = 0$, or, $\beta_5 = \beta_4$
2. Alternative: There is some difference in mean change from baseline to four weeks between the two groups, i.e. at least one of β_5 or β_4 are not zero

Question 2

2 - A

Estimated parameters are not the same, therefore this structure cannot be exchangeable. After a quick check, it looks like $0.2044^2 \neq 0.1936$, so we do not have an AR-1 structure. Entries off-diagonal are not zero, therefore, we do not have an independence correlation structure. The only remaining choice of the four we commonly use is unstructured.

We have an unstructured working correlation.

2 - B

First, calculate the value of logit of probability post-baseline for control subjects:

$$logit[P(Y_{ij} = 1|Post_{ij} = 1, ICGroup_{ij} = 0)] = 0.1676 - 0.3238 = -0.156$$

Inverting the function, we obtain probability of correct answer among controls in the post-baseline period:

$$P(Y_{ij} = 1|Post_{ij} = 1, ICGroup_{ij} = 0) = \frac{exp(-0.156)}{1 + exp(-0.156)} = 0.461$$

Now, calculate the value of logit of probability post-baseline for treated subjects:

$$logit[P(Y_{ij} = 1|Post_{ij} = 1, ICGroup_{ij} = 1)] = 0.1676 - 0.3238 - 0.1599 + 1.0073 = 0.691$$

Inverting the function, we obtain probability of correct answer among treated subjects in the post-baseline period:

$$P(Y_{ij} = 1|Post_{ij} = 1, ICGroup_{ij} = 1) = \frac{exp(0.691)}{1 + exp(0.691)} = 0.666$$

2 - C

In the presence of an interaction term, $\hat{\beta}_1$ is the effect of post-baseline period on the log-odds of correctly answering the question among controls. On average, after adjusting for other predictors, the log-odds of correctly answering the question in the post-baseline period were 0.3238 less than the log-odds among controls in at the baseline visit.

Alternatively, the odds of a correct answer among controls were approximately 27% lower in the post-baseline period when compared with the baseline visit, after adjusting for other predictors.

2 - D

log-odds ratio for the treated patients in the post-baseline period in terms of $\hat{\beta}_i$ coefficients is:

$$\log[\frac{p}{1-p}] = \beta_0 + \beta_1 + \beta_2 + \beta_3$$

log-odds ratio for the treated people at the baseline is:

$$\log[\frac{p}{1-p}] = \beta_0 + \beta_2$$

Therefore, odds ratio, or log-odds ratio, or other comparison will be a function of β_1 and β_3 representing the effect of post-baseline time period and an additional effect specific to the treated subjects, respectively.

Thus, to calculate a 95% confidence interval we need variance of an estimator that depends on both β_1 and β_3, which are correlated. Covariance of β_1 and β_3 is a missing piece that we need for the calculation of variance of an odds ratio for the comparison that we wish to perform.

2 - E

In the context of our problem, $exp(\beta_3)$ estimates the additional percentage increase in the odds of answering the question correctly for the treated population in the post-baseline period, after adjusting for other predictors.

2 - F

We have two predictors: treatment assignment, which is cluster invariant, and an indicator for the post-baseline time point, which is cluster variant.

Ignoring correlation will results in a smaller standard error estimate for the treatment assignment effect. Therefore, we will conduct anti-conservative inference, and detect effects which might not be truly statistically significant.

Ignoring correlation structure for the post-baseline indicator variable, we will ignore the nature of within-cluster variation of this variable, and will estimate standard errors that are higher than what it should be. This will results in conservative inference and, potentially, inability to detect truly statistically significant effects.