P8157_hw3_rw2844

Ryan Wei

2023-10-20

Question 1

(a)

The linear mixed effects models to be fitted are

$$Y_{ki} = \beta_0 + \beta_1 \mathbf{I}(\text{diet}_{ki} = 1) + \beta_2 \mathbf{I}(\text{diet}_{ki} = 2) + \beta_3 \cdot \text{Months}_{ki} + \beta_4 \mathbf{I}(\text{diet}_{ki} = 1) \cdot \text{Months}_{ki} + \beta_4 \mathbf{I}(\text{diet}_{ki} = 2) \cdot \text{Months}_{ki} + \gamma_{0k} + \epsilon_{ki},$$

i.e., the mean model plus a random Intercept (model (i)) and

$$Y_{ki} = \beta_0 + \beta_1 \mathbf{I}(\text{diet}_{ki} = 1) + \beta_2 \mathbf{I}(\text{diet}_{ki} = 2) + \beta_3 \cdot \text{Months}_{ki} + \beta_4 \mathbf{I}(\text{diet}_{ki} = 1) \cdot \text{Months}_{ki} + \beta_4 \mathbf{I}(\text{diet}_{ki} = 2) \cdot \text{Months}_{ki} + \gamma_{0k} + \gamma_{1k} \cdot \text{Months}_{ki} + \epsilon_{ki}.$$

i.e., the mean model plus a random Intercept and random slopes for time variable (model (ii)).

Table 1 shows the coefficients estimates and standard error estimates for the fixed effect and the variance components for the random effects of these two models.

Table 1: Coefficients estimates and standard error estimates from the mixed effect model fits

	Fixed effects				Random effects		
	Estimate	SE	t-statistic	p-value		SD	Correlation
Model (i)							
Intercept	250.945	0.482	521.159	< 0.001	Intercept	2.961	
$\mathbf{I}(\text{diet} = 1)$	-0.645	0.681	-0.948	0.345	Residual	1.099	
$\mathbf{I}(\text{diet} = 2)$	-2.660	0.681	-3.909	< 0.001			
Months	0.040	0.025	1.636	0.102			
$\mathbf{I}(\text{diet} = 1) \times$ Months	-0.124	0.033	-3.738	< 0.001			
$\mathbf{I}(\text{diet} = 2) \times$ Months	-0.501	0.033	-15.267	< 0.001			
Model (ii)							
Intercept	250.991	0.511	491.177	< 0.001	Intercept	3.159	
$\mathbf{I}(\text{diet} = 1)$	-0.696	0.722	-0.964	0.337	Months	0.158	-0.407
$\mathbf{I}(\text{diet} = 2)$	-2.688	0.722	-3.722	< 0.001	Residual	1.001	
Months	0.019	0.036	0.527	0.598			
$\mathbf{I}(\text{diet} = 1) \times$ Months	-0.102	0.050	-2.043	0.041			
$\mathbf{I}(\text{diet} = 2) \times$ Months	-0.491	0.050	-9.864	< 0.001			

Interpretation:

Based on the results from the model (ii):

- For the fixed effects:

  - β_0: The average weight for the patients with dietary counseling at baseline (diet=0) at the beginning of the study is 250.991.
  - β_1: Compared to the patients with dietary counseling at baseline (diet=0) at the beginning of the study, on average, patients with dietary counseling at all study visits (diet=1) is 0.696 lighter.
  - β_2: Compared to the patients with dietary counseling at baseline (diet=0) at the beginning of the study, on average, patients with dietary counseling at all visits plus free access to an exercise facility (diet=2) is 2.688 lighter.
  - β_3: For the patients with dietary counseling at baseline (diet=0) and weigh 250.991 at the beginning, on average, with each month after the beginning of the study, there is 0.019 increase in their weight.
  - β_4: Compared to the patients with dietary counseling at baseline (diet=0) at the beginning of the study, on average, patients with dietary counseling at all study visits (diet=1) lose an extra 0.102 weight per month after the beginning of the study.
  - β_5: Compared to the patients with dietary counseling at baseline (diet=0) at the beginning of the study, on average, patients with dietary counseling at all visits plus free access to an exercise facility (diet=2) lose an extra 0.491 weight per month after the beginning of the study.

- For the random effects:

  - The random intercepts, γ_{0k}, characterizes the heterogeneity in the average weight for the patients with dietary counseling at baseline (diet=0) at the beginning of the study. The mean of this effect is 0 and the variance of this effect is $\text{Var}(\gamma_{0k}) = \mathbf{\Sigma}_{\gamma,00} = 9.978$.
  - The random slopes, γ_{1k}, characterizes heterogeneity in the change of the weight after the beginning of the study for the patients with dietary counseling at baseline (diet=0) and weigh 250.991 at the beginning. The mean of this effect is 0 and the variance of this effect is $\text{Var}(\gamma_{1k}) = \mathbf{\Sigma}_{\gamma,11} = 0.025$.
  - The covariance between γ_{0k} and γ_{1k} is $\text{Cov}(\gamma_{0k}, \gamma_{1k}) = \mathbf{\Sigma}_{\gamma,01} = \mathbf{\Sigma}_{\gamma,10} = -0.407$.

- For the residual:

  - The residual standard deviation, 1.0006, represents the variability within the groups, not accounted for by the random effects. It is essentially the standard deviation of the unexplained variability within each group (patient).

(b)

The null hypothesis testing whether the random intercepts/slopes model provides a significantly better fit to the data than the random intercept model is

$$H_0 : \mathbf{G}(\boldsymbol{\alpha}) = \begin{bmatrix} \mathbf{\Sigma}_{\gamma,00} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

and the alternative hypothesis is

$$H_1 : \mathbf{G}(\boldsymbol{\alpha}) = \begin{bmatrix} \mathbf{\Sigma}_{\gamma,00} & \mathbf{\Sigma}_{\gamma,01} \\ \mathbf{\Sigma}_{\gamma,10} & \mathbf{\Sigma}_{\gamma,11} \end{bmatrix},$$

where $\mathbf{\Sigma}_{\gamma,00}, \mathbf{\Sigma}_{\gamma,11}, \mathbf{\Sigma}_{\gamma,01}, \mathbf{\Sigma}_{\gamma,10}$ are defined the same as in part (a).

Figure 1, shows the distribution of the likelihood ratio test (LRT) statistics as well as the distributions of χ^2_1 and χ^2_2. We can see that the distribution of our test statistics under null is in between the distribution of χ^2_1 and χ^2_2, which reflect the fact that the distribution of our test statistics under null is a 50:50 mixture of a χ^2_1 distribution and a χ^2_2 distribution.

To get the p-value of this test, we can compare the observed LRT statistic to a 50:50 mixture of χ^2_1 and χ^2_2 distributions.

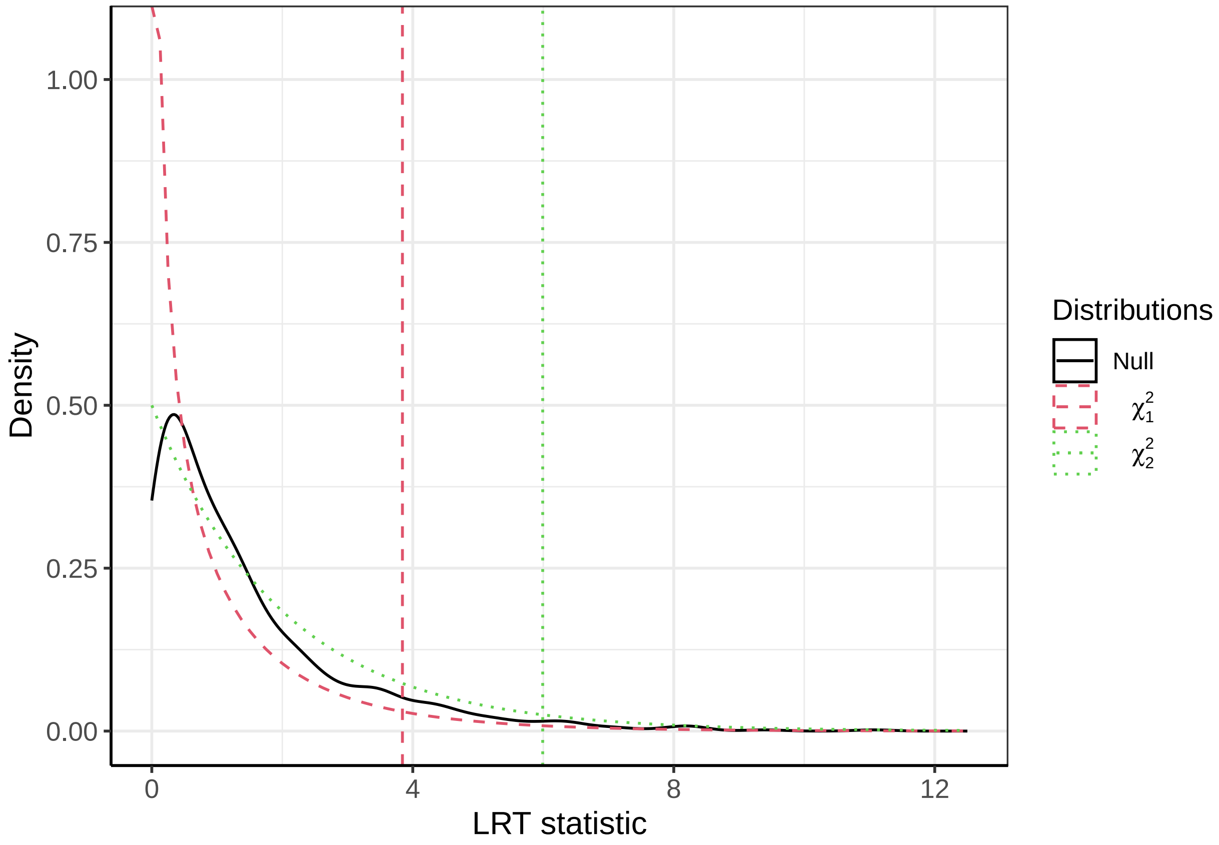

Figure 1: Distribution of the LRT statistics under the null with the distribution of χ^2_1 and χ^2_2. The vertical lines shows the 95^{th} percentiles.

Conclusion:

Since the observed LRT statistic is 70.381, the corresponding p-value is < 0.001. Therefore, we reject the null hypothesis at 0.05 significance level and conclude that the random intercept/slopes model gives us a better fit of the data compared to the random intercept model.

(c)

For the residual analysis, we focused on the normalized stage-one (cluster-level) residuals (as well as the predicted random intercepts), that is

$$\boldsymbol{\epsilon}_k = \mathbf{Y}_k - \mathbf{X}_k\boldsymbol{\beta} - \mathbf{Z}_k\boldsymbol{\gamma}_k,$$

since the marginal (population-level) residuals $\boldsymbol{e}_k = \mathbf{Y}_k - \mathbf{X}_k\boldsymbol{\beta}$ is just the unnormalized stage-one residual plus the predicted random effects.

Figure 2 shows the boxplots of $\hat{\epsilon}_{ki}$ versus diet and months. From the plots we can see that there is inconclusive evidence regarding heteroskedasticity by diet and month, which means that the assumed specification of the mean model is adequate.

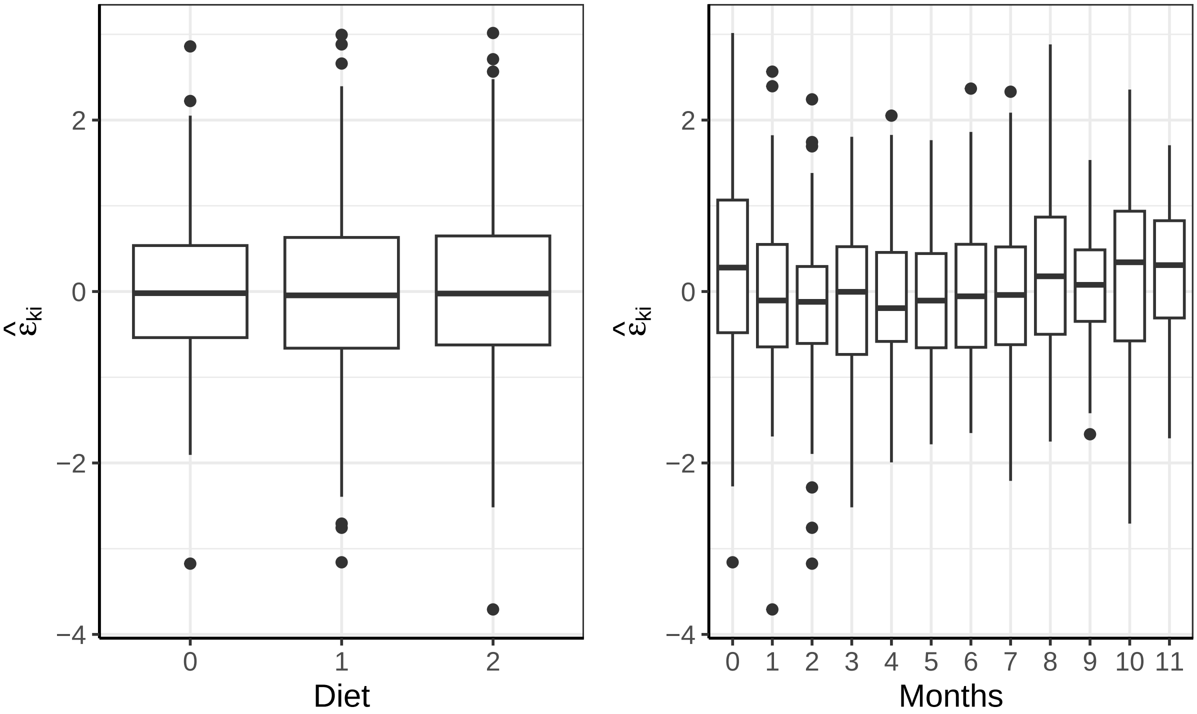

Figure 2: Distribution of the stage-one residuals, stratified by diet categories and months.

Figure 3 shows the $\hat{\epsilon}_{ki}$ versus lag($\hat{\epsilon}_{ki}$) (lag 1). From the plot we can see that there is not indication of unaccounted for local correlation or any residual serial dependence. The plot of stage-one residuals versus fitted values also shows that there is no residual mean-variance relationship (figure not shown).

Figure 4 shows the quantile-quantile plots of stage-one residuals $\hat{\epsilon}_{ki}$ and random intercepts $\hat{\gamma}_{ki}$. From the plots we can see linearity between theoretical and sample quantiles in both cases in the center. However, there are some evidence of heavier than Normal tails in the stage-one residuals while the random interceprs seems fine.

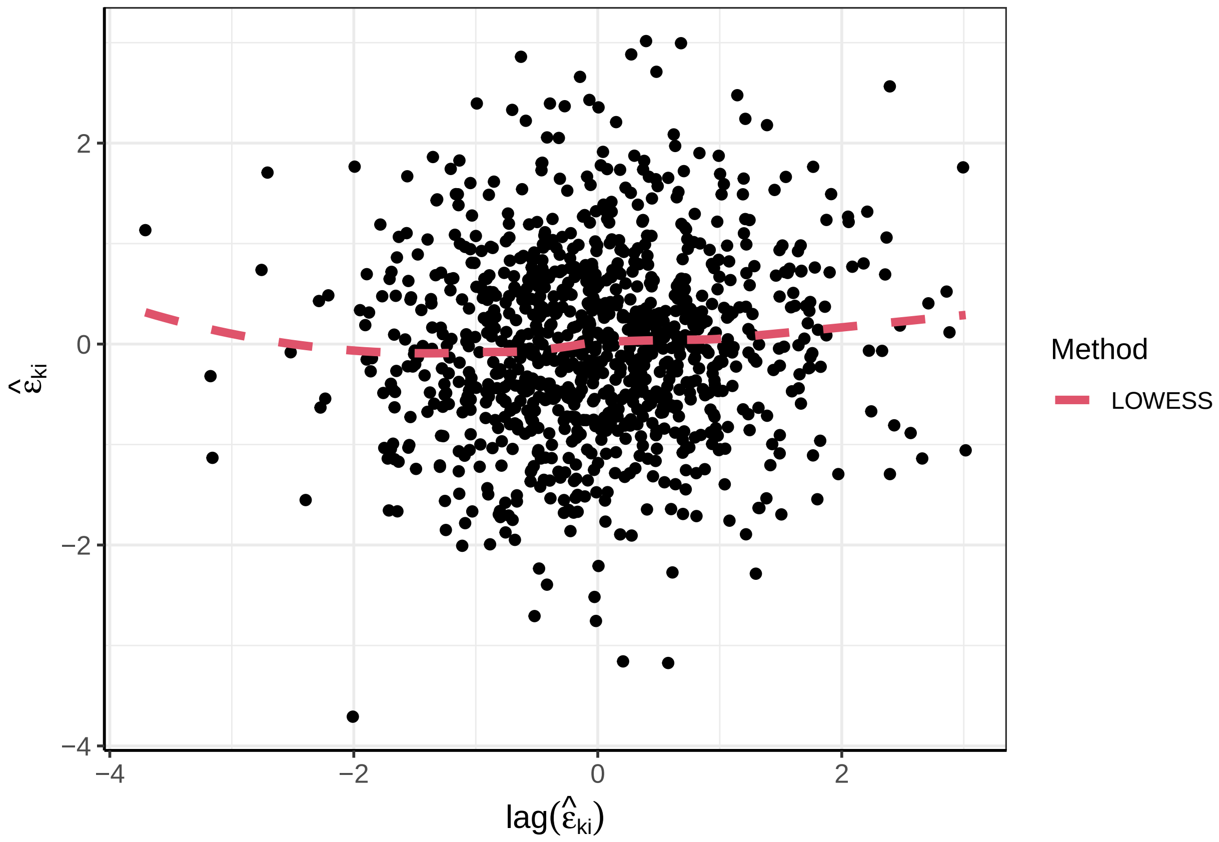

Figure 3: Residuals versus lagged residuals, with loess smoothed curve.

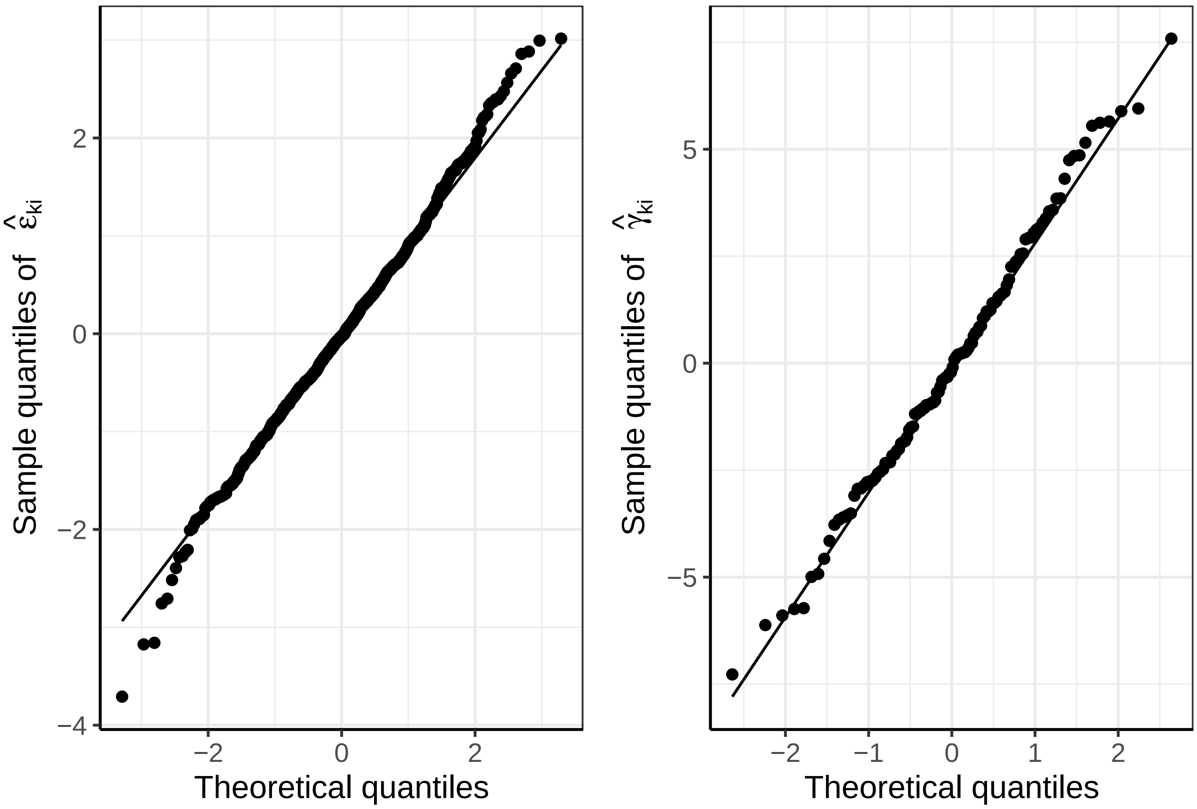

Figure 4: Q-Q plots of stage-one residuals (left panel) and random intercepts (right panel).

(d)

Table 2 shows the coefficients estimates and standard error estimates for the marginal model fits using GEE with different dependence structures.

Table 2: Coefficients estimates from the marginal model fits, with different working correlation models.

	Estimate	$\text{SE}_{\text{robust}}$	Wald Statistics	p-value
Working independence (GEE-I)				
Intercept	251.017	0.581	186570.277	< 0.001
$\mathbf{I}(\text{diet} = 1)$	-0.861	0.755	1.300	0.254
$\mathbf{I}(\text{diet} = 2)$	-2.809	0.785	12.811	< 0.001
Months	-0.016	0.082	0.038	0.846
$\mathbf{I}(\text{diet} = 1) \times \text{Months}$	0.008	0.099	0.007	0.934
$\mathbf{I}(\text{diet} = 2) \times \text{Months}$	-0.400	0.105	14.628	< 0.001
Working exchangeable (GEE-E)				
Intercept	250.945	0.559	201672.333	< 0.001
$\mathbf{I}(\text{diet} = 1)$	-0.646	0.736	0.771	0.38
$\mathbf{I}(\text{diet} = 2)$	-2.661	0.748	12.637	< 0.001
Months	0.040	0.035	1.309	0.253
$\mathbf{I}(\text{diet} = 1) \times \text{Months}$	-0.124	0.055	5.077	0.024
$\mathbf{I}(\text{diet} = 2) \times \text{Months}$	-0.500	0.050	99.206	< 0.001
ρ	0.863	0.120		
Working AR-1 (GEE-AR1)				
Intercept	251.130	0.583	185237.650	< 0.001
$\mathbf{I}(\text{diet} = 1)$	-0.703	0.785	0.802	0.37
$\mathbf{I}(\text{diet} = 2)$	-2.246	0.800	7.874	0.005
Months	0.010	0.042	0.057	0.811
$\mathbf{I}(\text{diet} = 1) \times \text{Months}$	-0.119	0.060	3.907	0.048
$\mathbf{I}(\text{diet} = 2) \times \text{Months}$	-0.522	0.056	87.443	< 0.001
ρ	0.965	0.032		

Interpretation:

Based on the results from the model GEE-E:

- β_0: The average weight for the patients with dietary counseling at baseline (diet=0) at the beginning of the study is 250.945.
- β_1: Compared to the patients with dietary counseling at baseline (diet=0) at the beginning of the study, on average, patients with dietary counseling at all study visits (diet=1) is 0.646 lighter.
- β_2: Compared to the patients with dietary counseling at baseline (diet=0) at the beginning of the study, on average, patients with dietary counseling at all visits plus free access to an exercise facility (diet=2) is 2.661 lighter.
- β_3: For the patients with dietary counseling at baseline (diet=0) and weigh 250.945 at the beginning, on average, with each month after the beginning of the study, there is 0.040 increase in their weight.
- β_4: Compared to the patients with dietary counseling at baseline (diet=0) at the beginning of the study, on average, patients with dietary counseling at all study visits (diet=1) lose an extra 0.124 weight per month after the beginning of the study.
- β_5: Compared to the patients with dietary counseling at baseline (diet=0) at the beginning of the study, on average, patients with dietary counseling at all visits plus free access to an exercise facility (diet=2) lose an extra 0.500 weight per month after the beginning of the study.
- ρ: The dependence among observations for any given patients appeared to be large, since $\rho = 0.863$, meaning that any two observation for a given patient is positively correlated.

(e)

The null hypothesis testing whether the rate of weight loss differs for the treatment groups is

$$H_0 : \beta_4 = \beta_5 = 0, \text{or}, \ \mathbf{Q}\boldsymbol{\beta} = \mathbf{0},$$

and the alternative hypothesis for this test is

$$H_1 : \beta_4 \neq 0, \beta_5 \neq 0, \text{or}, \ \mathbf{Q}\boldsymbol{\beta} \neq \mathbf{0},$$

Using the GEE-E estimator $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5)^T$, we can construct a Wald-type test statistic

$$(\mathbf{Q}\hat{\boldsymbol{\beta}})^T(\mathbf{Q}\,\widehat{\text{Cov}}[\hat{\boldsymbol{\beta}}]\,\mathbf{Q}^T)^{-1}(\mathbf{Q}\hat{\boldsymbol{\beta}}) \sim \chi^2_2,$$

where $\mathbf{Q} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$.

Base on the GEE-E estimates, our Wald statistic is 104.834, and the corresponding p-value is < 0.001. Therefore, we reject the null hypothesis at 0.05 significance level and conclude that the rate of weight loss differs for the treatment groups.

(f)

When the mean model is correctly specified,

- **Point Estimates**: Both linear mixed effects model (LMMs) (in part (a)) and marginal model using GEE (in part (d)) provide consistent point estimates. Since the marginal mean model $E[\mathbf{Y}_k \mid \mathbf{X}_k] = \mathbf{X}_k\boldsymbol{\beta}$ is the same in two models. However, the interpretation is slightly different since we need conditional on the random effect to interpret the mean response in linear mixed effects model.
- **Standard Error Estimates**: GEEs have an edge in robustness, particularly when the correlation structure is uncertain. LMMs require a correctly specified model, including the random effects structure, for valid standard error estimates.

In summary, while both models can provide consistent point estimates under correct mean model specification, GEEs offer more robust standard errors in the face of uncertainty regarding the correlation structure. This robustness can be particularly advantageous in complex datasets where the correlation structure is not well understood or is difficult to model accurately.

Appendix: Code for this report

```
knitr::opts_chunk$set(echo = FALSE, message = F, warning = F)
options(knitr.kable.NA = '')
library(tidyverse)
library(caret)
library(latex2exp)
library(gstat)
library(sp)
library(nlme)
library(kableExtra)
library(geepack)
write_matex <- function(x) {
  begin <- "$$\\begin{bmatrix}"
  end <- "\\end{bmatrix}$$"
  X <-
    apply(x, 1, function(x) {
      paste(
        paste(x, collapse = "&"),
        "\\\\"
      )
    })
  writeLines(c(begin, X, end))
}
theme_set(
  theme_bw()+
  theme(
    plot.title = element_text(size = 16, hjust = 0.5),
    axis.title.x = element_text(size = 12),
    axis.title.y = element_text(size = 12),
    axis.text = element_text(size = 10),
    axis.line = element_line(color = "black", size = 0.5),
  )
)
load("../../datasets/WeightLoss/WtLoss.Rdata")
fit1.ML = lme(fixed = weight ~ as.factor(diet)*time,
              random = reStruct(~ 1 | id),
              data = wtloss,
              method = "ML")
fit1.sum = summary(fit1.ML)
fit1.ranef = as.data.frame(VarCorr(fit1.ML)[,"StdDev"])
fit1.fixed = fit1.sum$tTable

fit2.ML = lme(fixed = weight ~ as.factor(diet)*time,
              random = reStruct(~ time | id),
              data = wtloss,
              method = "ML")
fit2.sum = summary(fit2.ML)
fit2.ranef = as.data.frame(VarCorr(fit2.ML)[,2:3])
fit2.fixed = fit2.sum$tTable

lme.sum.tab = rbind(fit1.fixed, fit2.fixed) %>% as.data.frame(row.names = c(1:12)) %>% mutate(terms = r
  mutate(`p-value` = ifelse(`p-value` < 0.001, "< 0.001", round(`p-value`,3)))
```

```r
lme.sum.tab %>% select(terms, everything()) %>% knitr::kable(digits = 3, booktab = T, row.names = F, col
  pack_rows("Model (ii)", 7, 12) %>% kable_styling(latex_options = "hold_position" )
fit1.sim = simulate.lme(fit1.ML,nsim = 1000, seed = 1504, method = "ML",m2 =fit2.ML)
#fit2.sim = simulate.lme(fit2.ML,nsim = 1000, seed = 1504, method = "ML")
lrt.stat.null = -2 *(fit1.sim$null$ML[,2] - fit1.sim$alt$ML[,2])

ggplot() +
  geom_density(aes(x = lrt.stat.null, linetype = "Null", color = "Null")) +
  stat_function(fun = dchisq, args = list(df = 1), aes(color = "df = 1", linetype = "df = 1")) +
  stat_function(fun = dchisq, args = list(df = 2), aes(color = "df = 2", linetype = "df = 2")) +
  xlim(0,12.5)+
  geom_vline(color =2, lty = 2, xintercept = qchisq(0.95, 1))+
  geom_vline(color = 3, lty = 3, xintercept = qchisq(0.95, 2))+
  scale_linetype_manual(name = "Distributions", values = c(1,2,3),breaks = c("Null", "df = 1", "df = 2"
  scale_color_manual(name = "Distributions", values = c(1,2,3),breaks = c("Null", "df = 1", "df = 2"),
  xlab("LRT statistic") +
  ylab("Density")
lrt.stat.obs = -2*(logLik(fit1.ML) - logLik(fit2.ML))
mix.chiSq = c(rchisq(1e6, df = 1), rchisq(1e6, df = 2))
lrt.pval = mean(mix.chiSq > lrt.stat.obs)
epsHat <- resid(fit1.ML, type="normalized")
gammaHat <- ranef(fit1.ML)[,1]
#Investigate heteroskedasticity in the standardized stage 1 Residuals:
resid_df = cbind(epsHat, wtloss)

box1 = ggplot() +
  geom_boxplot(aes(x = as.factor(wtloss$diet), y = epsHat)) +
  xlab("Diet") +
  ylab(TeX("$\\hat{\\epsilon}_{ki}$"))
box2 = ggplot() +
  geom_boxplot(aes(x = as.factor(wtloss$time), y = epsHat)) +
  xlab("Months") +
  ylab(TeX("$\\hat{\\epsilon}_{ki}$"))
library(patchwork)
box1 + box2
#Compare Residuals and lagged Residuals to investigate potential serial dependence:
ggplot() +
  geom_point(aes(x = lag(epsHat), y = epsHat)) +
  geom_smooth(method = "loess", se = FALSE, aes(x = lag(epsHat), y = epsHat,linetype = "LOWESS", color =
  scale_linetype_manual(name = "Method", values = c(2),breaks = c("LOWESS"))+
  scale_color_manual(name = "Method", values = c(2),breaks = c("LOWESS")) +
  xlab(TeX("lag$(\\hat{\\epsilon}_{ki})$")) +
  ylab(TeX("$\\hat{\\epsilon}_{ki}$"))
#Q-Q plots:
qq.eps =
  ggplot() + geom_qq(aes(sample = epsHat)) + geom_qq_line(aes(sample = epsHat), fullrange = T) +
  xlab(TeX("Theoretical quantiles$")) +
  ylab(TeX("Sample quantiles of $\\ \\hat{\\epsilon}_{ki}$"))

qq.gamma =
  ggplot() + geom_qq(aes(sample = gammaHat)) + geom_qq_line(aes(sample = gammaHat), fullrange = T) +
  xlab(TeX("Theoretical quantiles$")) +
  ylab(TeX("Sample quantiles of $\\ \\hat{\\gamma}_{ki}$"))
```

```
qq.eps + qq.gamma

fit.GEE.I = geeglm(weight ~ as.factor(diet)*time, id = id, data = wtloss, family = gaussian, scale.fix
fit.GEE.E = geeglm(weight ~ as.factor(diet)*time, id = id, data = wtloss, family = gaussian, scale.fix
fit.GEE.AR1 = geeglm(weight ~ as.factor(diet)*time, id = id, data = wtloss, family = gaussian, scale.fi
#summary(fit.GEE.I)
#summary(fit.GEE.E)
#summary(fit.GEE.AR1)

# formatting results
GEE.I.sum = summary(fit.GEE.I)
GEE.E.sum = summary(fit.GEE.E)
GEE.AR1.sum = summary(fit.GEE.AR1)

GEE.I.coef = GEE.I.sum$coefficients
GEE.E.coef = GEE.E.sum$coefficients
GEE.AR1.coef = GEE.AR1.sum$coefficients

GEE.I.coef$terms = c("Intercept", "$\\mathbf{I}(\\text{diet} = 1)$", "$\\mathbf{I}(\\text{diet} = 2)$",
GEE.E.coef$terms = c("Intercept", "$\\mathbf{I}(\\text{diet} = 1)$", "$\\mathbf{I}(\\text{diet} = 2)$",
GEE.AR1.coef$terms = c("Intercept", "$\\mathbf{I}(\\text{diet} = 1)$", "$\\mathbf{I}(\\text{diet} = 2)$

GEE.E.corr = GEE.E.sum$corr
GEE.E.corr["Wald"] = NA
GEE.E.corr["Pr(>|W|)"] = NA
GEE.E.corr["terms"] = "$\\rho$"
GEE.AR1.corr = GEE.AR1.sum$corr
GEE.AR1.corr["Wald"] = NA
GEE.AR1.corr["Pr(>|W|)"] = NA
GEE.AR1.corr["terms"] = "$\\rho$"

GEE.E.coef = rbind(GEE.E.coef, GEE.E.corr)
GEE.AR1.coef = rbind(GEE.AR1.coef, GEE.AR1.corr)

GEE.I.coef$corstr = GEE.I.sum$corstr
GEE.E.coef$corstr = GEE.E.sum$corstr
GEE.AR1.coef$corstr = GEE.AR1.sum$corstr

# summary table
gee.sum.tab <- rbind(GEE.I.coef, GEE.E.coef, GEE.AR1.coef)

gee.sum.tab %>%
  mutate(`Pr(>|W|)` = ifelse(`Pr(>|W|)` < 0.001, "< 0.001", round(`Pr(>|W|)`,3)))%>% select(terms,every
  pack_rows("Working exchangeable (GEE-E)", 7, 13) %>% pack_rows("Working AR-1 (GEE-AR1)", 14, 20)
wald.stat = t(coef(fit.GEE.E)[5:6]) %*% solve(vcov(fit.GEE.E)[5:6,5:6]) %*% (coef(fit.GEE.E)[5:6])
wald.p.value = pchisq(wald.stat,2,lower.tail = F)
```