Code Appendix

Zirui Zhang

```
library(tidyverse)
library(ggplot2)
library(patchwork)
library(nlme)
library(lme4)
library("car")
library(geepack)
```

```
load("~/Documents/2023Fall/P8157/P8157/WtLoss.RData")
data = wtloss |>
  mutate(diet = as.factor(diet))
p0 = ggplot(data, aes(x = time, y = weight, group = id, color = as.factor(diet))) +
  geom_line() +
  facet_grid(~diet) +
  theme_classic()
```

```
fit1 = lme(fixed = weight ~ diet*time, random=reStruct(~ 1 | id), data=data, method="ML")
fit2 = lme(fixed = weight ~ diet*time, random=reStruct(~ time | id), data=data, method="ML")
# fixed effect
fixed = data.frame(
  coef.fit1 = c(summary(fit1)$coefficients$fixed),
  sd.fit1 = c(sqrt(diag(summary(fit1)$varFix))),
  coef.fit2 = c(summary(fit2)$coefficients$fixed),
  sd.fit2 = c(sqrt(diag(summary(fit2)$varFix)))
)
rownames(fixed) = c("Intercept, b0",
                    "Main effect for diet1, b1", "Main effect for diet2, b2",
                    "Main effect for time, b3",
                    "Interaction for diet1, b4", "Interaction for diet2, b5")
colnames(fixed) = c("Est.fit1", "SE.fit1","Est.fit2", "SE.fit2")
# random effect
random = data.frame(
  ran.fit1 = c(as.numeric(VarCorr(summary(fit1))[1,2]), NA, summary(fit1)$sigma),
  ran.fit2 = c(as.numeric(VarCorr(summary(fit2))[1,2]), as.numeric(VarCorr(summary(fit2))[2,2]), summary
)
rownames(random) = c("SD of random intercepts", "SD of random slope",
                     "SD of errors")
colnames(random) = c("fit1", "fit2")
```

```
sim = simulate(fit1, nsim = 1000, seed = 1504, fit2, method = "ML")
lrt = data.frame(stat = -2*(-sim$alt$ML[, "logLik"]+sim$null$ML[, "logLik"]))
lrt_95 = quantile(lrt$stat, 0.95)
data1 = data.frame(x = rchisq(1000, df = 1))
```

```r
data2 = data.frame(x = rchisq(1000, df = 2))
lrt_1 = quantile(data1$x, 0.95)
lrt_2 = quantile(data2$x, 0.95)
p1 = ggplot() +
  geom_histogram(data = data1, aes(x = x, y = ..density..),
                 bins = 100, fill = "blue", alpha = 0.5) +
  geom_histogram(data = data2, aes(x = x, y = ..density..),
                 bins = 100, fill = "green", alpha = 0.5) +
  geom_histogram(data = lrt, aes(x = stat, y = ..density..),
                 bins = 100, fill = "black", alpha = 0.8) +
  geom_vline(xintercept = lrt_95, linetype = "dashed", color = "black") +
  geom_vline(xintercept = lrt_1, linetype = "dashed", color = "blue") +
  geom_vline(xintercept = lrt_2, linetype = "dashed", color = "green") +
  labs(x = "LRT statistic",
       y = "Density") +
  annotate("text", x = 2.5, y = 0.6, label = "Chi-square 1", color = "blue") +
  annotate("text", x = 6.5, y = 0.6, label = "Chi-square 2", color = "green") +
  annotate("text", x = 4.7, y = 0.6, label = "LRT", color = "black") +
  theme_minimal()
```

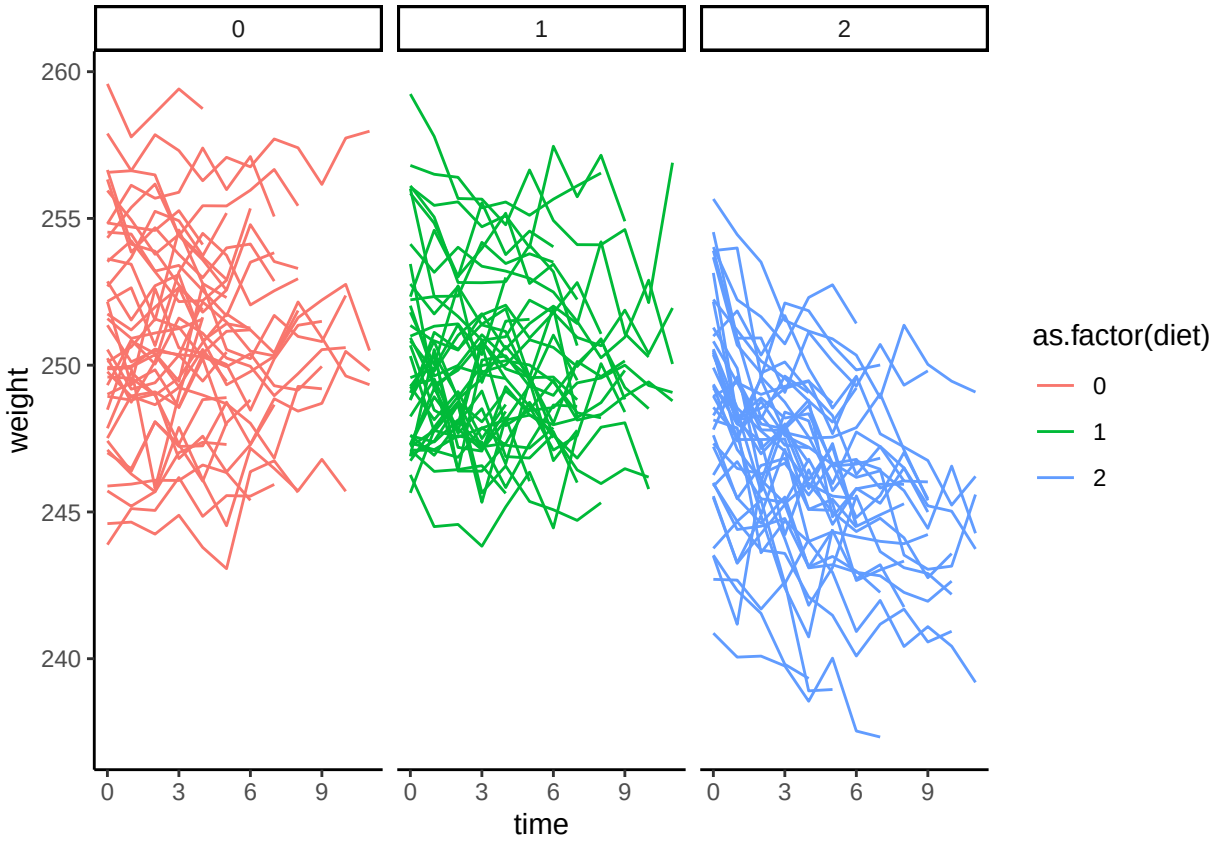
```r
# residuals - stage 1 and random intercept
epsHat = data.frame(eps = resid(fit1, type="normalized"))
gammaHat = data.frame(gam = ranef(fit1)[,1])
epsHat$diet = as.factor(data$diet)
epsHat$time = as.factor(floor(data$time/2))
# box plot mean model - diet - stage 1
p2 = ggplot(epsHat, aes(x = diet, y = eps)) + geom_boxplot()
# box plot mean model - time - stage 1
p3 = ggplot(epsHat, aes(x = time, y = eps)) + geom_boxplot()
# scatterplot for dependence model - stage 1
p4 = epsHat |> mutate(time = as.numeric(time)) |> filter(time > 3) |>
  ggplot(aes(x = lag(eps), y = eps)) + geom_point() +geom_smooth()
```

```r
fit.I = geeglm(weight ~ diet*time, id=id, data, family=gaussian, scale.fix=TRUE, corstr="independence")
fit.E = geeglm(weight ~ diet*time, id=id, data, family=gaussian, scale.fix=TRUE, corstr="exchangeable")
fit.AR = geeglm(weight ~ diet*time, id=id, data, family=gaussian, scale.fix=TRUE, corstr="ar1")
est = data.frame(
  est = c(summary(fit.I)$coefficients[1][[1]][2:6],
          summary(fit.E)$coefficients[1][[1]][2:6],
          summary(fit.E)$geese$correlation[1,1],
          summary(fit.AR)$coefficients[1][[1]][2:6],
          summary(fit.AR)$geese$correlation[1,1]),
  se = c(summary(fit.I)$coefficients[2][[1]][2:6],
         summary(fit.E)$coefficients[2][[1]][2:6],
         summary(fit.E)$geese$correlation[1,2],
         summary(fit.AR)$coefficients[2][[1]][2:6],
         summary(fit.AR)$geese$correlation[1,2])
)
rownames(est) = c("GEE-I: diet1", "GEE-I: diet2", "GEE-I: time",
                  "GEE-I: time*diet1", "GEE-I: time*diet2",
                  "GEE-E: diet1", "GEE-E: diet2", "GEE-E: time",
                  "GEE-E: time*diet1", "GEE-E: time*diet2","GEE-E: rho",
                  "GEE-AR1: diet1", "GEE-AR1: diet2", "GEE-AR1: time",
```

```
                  "GEE-AR1: time*diet1", "GEE-AR1: time*diet2", "GEE-AR1: rho")
colnames(est) = c("Est", "SE")
```

```
fit.E.2 = geeglm(weight ~ diet+time, id=id, data, family=gaussian, scale.fix=TRUE, corstr="exchangeable
```

```
# spaghetti plot for different diets.
p0
```



We can see the intercepts across different clusters have high variability, and different clusters tend to have different slopes at the same time points.

0.1 Question (a)

```
knitr::kable(fixed, format = "markdown")
```

	Est.fit1	SE.fit1	Est.fit2	SE.fit2
Intercept, b0	250.9450770	0.4800621	250.9912464	0.5094672
Main effect for diet1, b1	-0.6454756	0.6786109	-0.6963624	0.7201476
Main effect for diet2, b2	-2.6599846	0.6784852	-2.6879129	0.7200520
Main effect for time, b3	0.0400853	0.0244266	0.0191093	0.0361432
Interaction for diet1, b4	-0.1241404	0.0331105	-0.1021718	0.0498554
Interaction for diet2, b5	-0.5006710	0.0326964	-0.4909484	0.0496217

```
knitr::kable(random, format = "markdown")
```

	fit1	fit2
SD of random intercepts	2.961354	3.1589179
SD of random slope	NA	0.1580983
SD of errors	1.098885	1.0005857

Interpretation for the fixed effects of Model 2:

- β_0: 250.99, the average baseline weight for patients assigned to diet 0 is 250.99 pounds.
- β_1: -0.6964, the average decrease in baseline weight when comparing those on diet 1 to diet 0 is 0.6964.
- β_2: -2.6879, the average decrease in baseline weight when comparing those on diet 2 to diet 0 is 2.6879.
- β_3: 0.0191, the average increase in weight for those on diet 0 is 0.0191 with unit increase in time.
- β_4: -0.1022, the average difference in weight decrese along unit increase in time when comparing those on diet 1 to those on diet 0 is 0.1022.
- β_5: -0.4909, the average difference in weight decrese along unit increase in time when comparing those on diet 2 to those on diet 0 is 0.4909.

Interpretation for the variance components of Model 2:

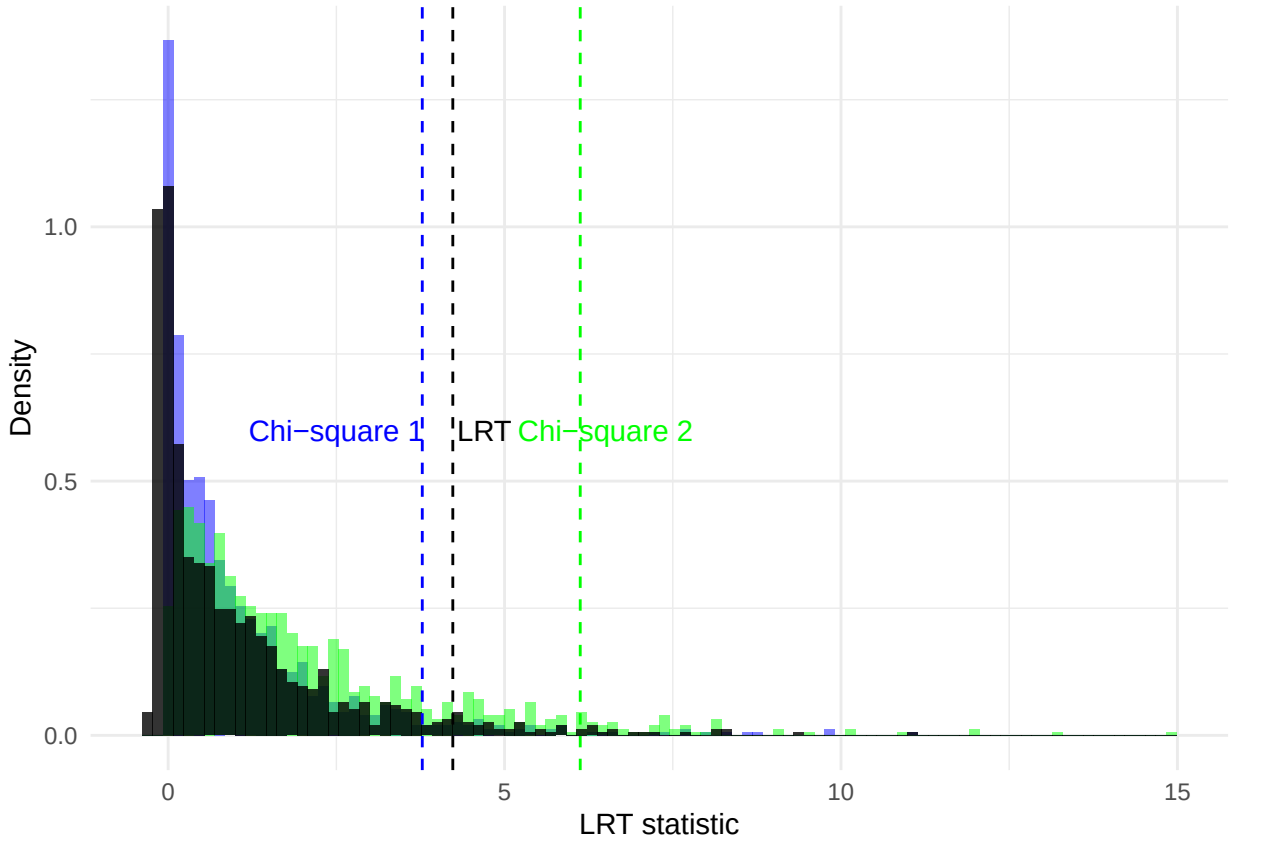
- $\sigma_{\gamma,0}$: 3.1589, standard deviation of the random intercept, represents the amount of variability in the intercepts among different clusters. The variation is pretty large.
- $\sigma_{\gamma,1}$: 0.1581, standard deviation of the random slope, represents the amount of variability in the effect of time among different clusters.

0.2 Question (b)

$$H_0 : G(a) = \begin{bmatrix} \Sigma_{\gamma,00} & 0 \\ 0 & 0 \end{bmatrix}$$

$$H_1 : G(a) = \begin{bmatrix} \Sigma_{\gamma,00} & \Sigma_{\gamma,01} \\ \Sigma_{\gamma,10} & \Sigma_{\gamma,11} \end{bmatrix}$$

```
p1
```

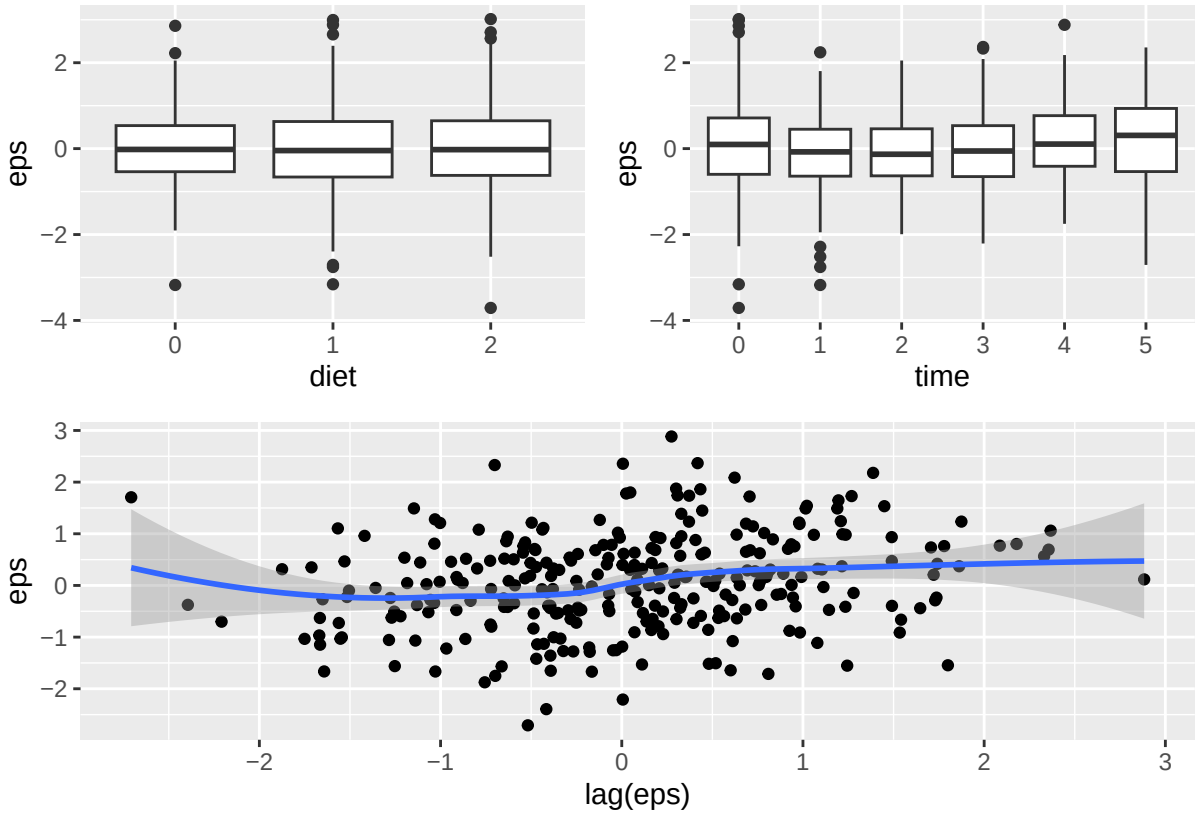


The sampling distribution seems to be a 50:50 mixture of the χ^2_1 and χ^2_2 distribution. The one-side 95% percentile of lrt lies between those of the χ^2_1 and χ^2_2 distribution.

From this plot, we may make the preliminary conclusion that the random slope term is significant, the random slope/intercept model is more adequate than the random intercept model.

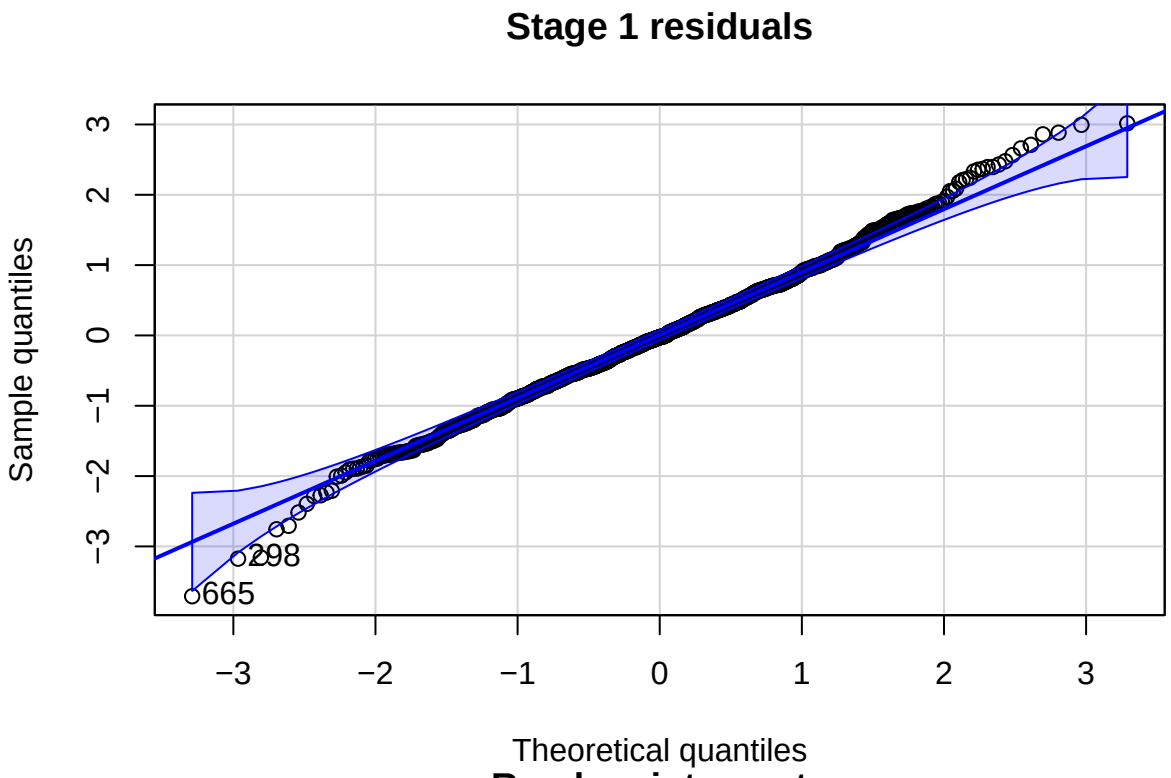
0.3 Question (c)

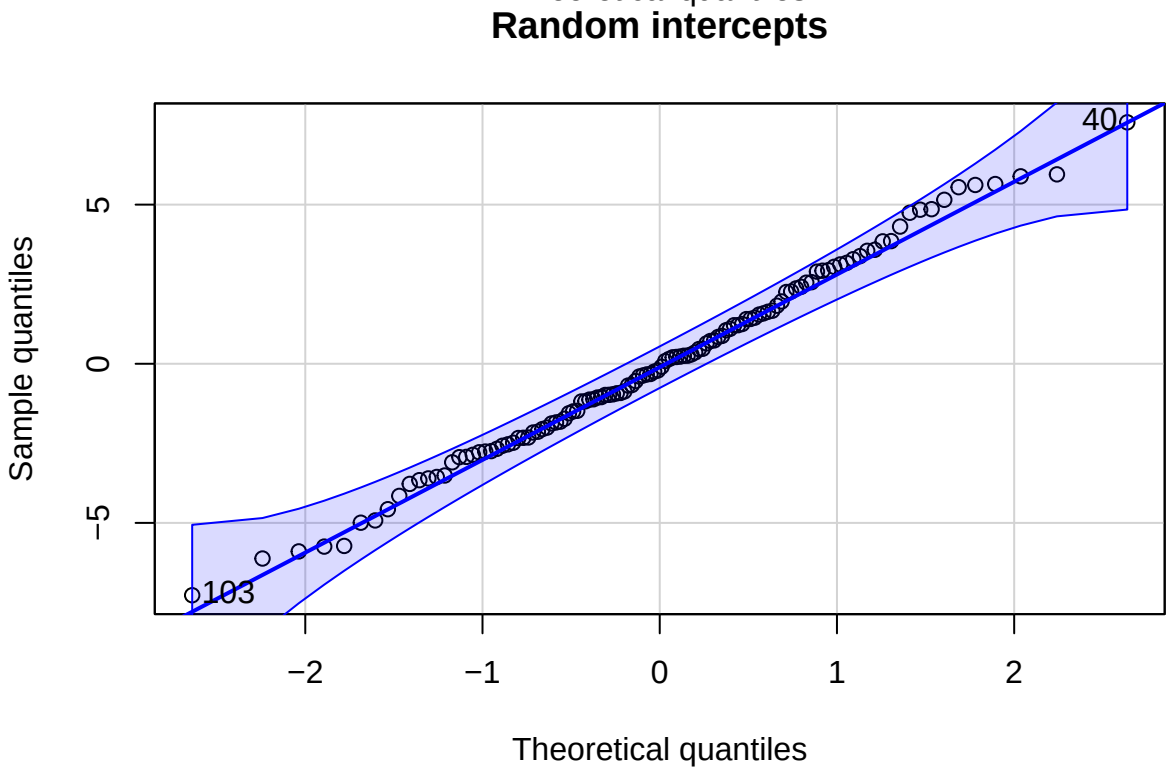
```
# box plot mean model - diet - stage 1
# box plot mean model - time - stage 1
# scatterplot for dependence model - stage 1
(p2+p3)/p4
```



- In the standardized stage 1 residual plot, there's inconclusive evidence regarding heteroskedasticity by diet, but presents strong evidence of heteroskedasticity by time.
- In the residual and lagged residual plot, there presents no local correlation, thus no serial dependence.

```
# qqplot for normality - both
qqPlot(epsHat$eps, xlab = "Theoretical quantiles", ylab = "Sample quantiles", main = "Stage 1 residuals
qqPlot(gammaHat$gam, xlab = "Theoretical quantiles", ylab = "Sample quantiles", main = "Random intercep
```






```
## [1] 705 401
```

- In the Q-Q plots, the stage 1 residuals and random effect estimates seem perfectly normal.

All the results are consistent with what we concluded from part(b).

0.4 Question (d)

```
knitr::kable(est, format = "markdown")
```

	Est	SE
GEE-I: diet1	-0.8612126	0.7552699
GEE-I: diet2	-2.8090850	0.7848149
GEE-I: time	-0.0160138	0.0824101
GEE-I: time*diet1	0.0081454	0.0985074
GEE-I: time*diet2	-0.4004588	0.1047030
GEE-E: diet1	-0.6461907	0.7361619
GEE-E: diet2	-2.6605211	0.7484207
GEE-E: time	0.0398574	0.0348323
GEE-E: time*diet1	-0.1236704	0.0548847
GEE-E: time*diet2	-0.5002941	0.0502293
GEE-E: rho	0.8625142	0.1200589
GEE-AR1: diet1	-0.7030401	0.7849039
GEE-AR1: diet2	-2.2458993	0.8003557
GEE-AR1: time	0.0101187	0.0423468
GEE-AR1: time*diet1	-0.1188261	0.0601179
GEE-AR1: time*diet2	-0.5221817	0.0558418
GEE-AR1: rho	0.9654686	0.0319882

Interpretation for GEE-E model:

- β_1: -0.6462, the average decrease in baseline weight when comparing those on diet 1 to diet 0 is 0.6462.
- β_2: -2.6605, the average decrease in baseline weight when comparing those on diet 2 to diet 0 is 2.6605.
- β_3: 0.0399, the average increase in weight for those on diet 0 is 0.0399 with unit increase in time.
- β_4: -0.1237, the average difference in weight decrese along unit increase in time when comparing those on diet 1 to those on diet 0 is 0.1237.
- β_5: -0.5003, the average difference in weight decrese along unit increase in time when comparing those on diet 2 to those on diet 0 is 0.5003.

0.5 Question (e)

$$H_0 : \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} = \begin{bmatrix} \beta_4 \\ \beta_5 \end{bmatrix} = 0$$

$$H_1 : \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} = \begin{bmatrix} \beta_4 \\ \beta_5 \end{bmatrix} \neq 0$$

```
anova(fit.E.2, fit.E)
```

```
## Analysis of 'Wald statistic' Table
##
## Model 1 weight ~ diet * time
## Model 2 weight ~ diet + time
##   Df     X2 P(>|Chi|)
## 1  2 104.83 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We had better use the model with the interaction term between time and diet.

Description:

In the 3 diet groups, patients on diet1 and diet2 lose more weight than patients on diet0 across the same time period/at the same time points; and patients on diet2 lose more than those on diet1 across the same time period/at the same time points.

0.6 Question (f)

If the mean model is correctly specified, the point estimates from the LMM and GEE model should all be consistent (unbiased). When the dependence structure is correctly specified, the LMM standard errors should be valid. When the dependence structure is not necessarily correct, with the sandwich-based estimator, we could still get valid standard error estimates for GEE.