MS Comprehensive Exam 2022

STA 207 (100 points)

Read the instructions on Canvas carefully!

Name:

Student ID:

```
library(lme4)
```

In this exam, we investigate the `ChickWeight` dataset in `R`. You can load the data using the following commands. Carefully read the help file of `ChickWeight` before working on the following questions.

```
data(ChickWeight)
```

(a) Briefly summarize all variables in the data set. You need to provide the definition of the variable and quantitative summary.

Solution: (Type your answer here)

```
sapply(ChickWeight,class)
```

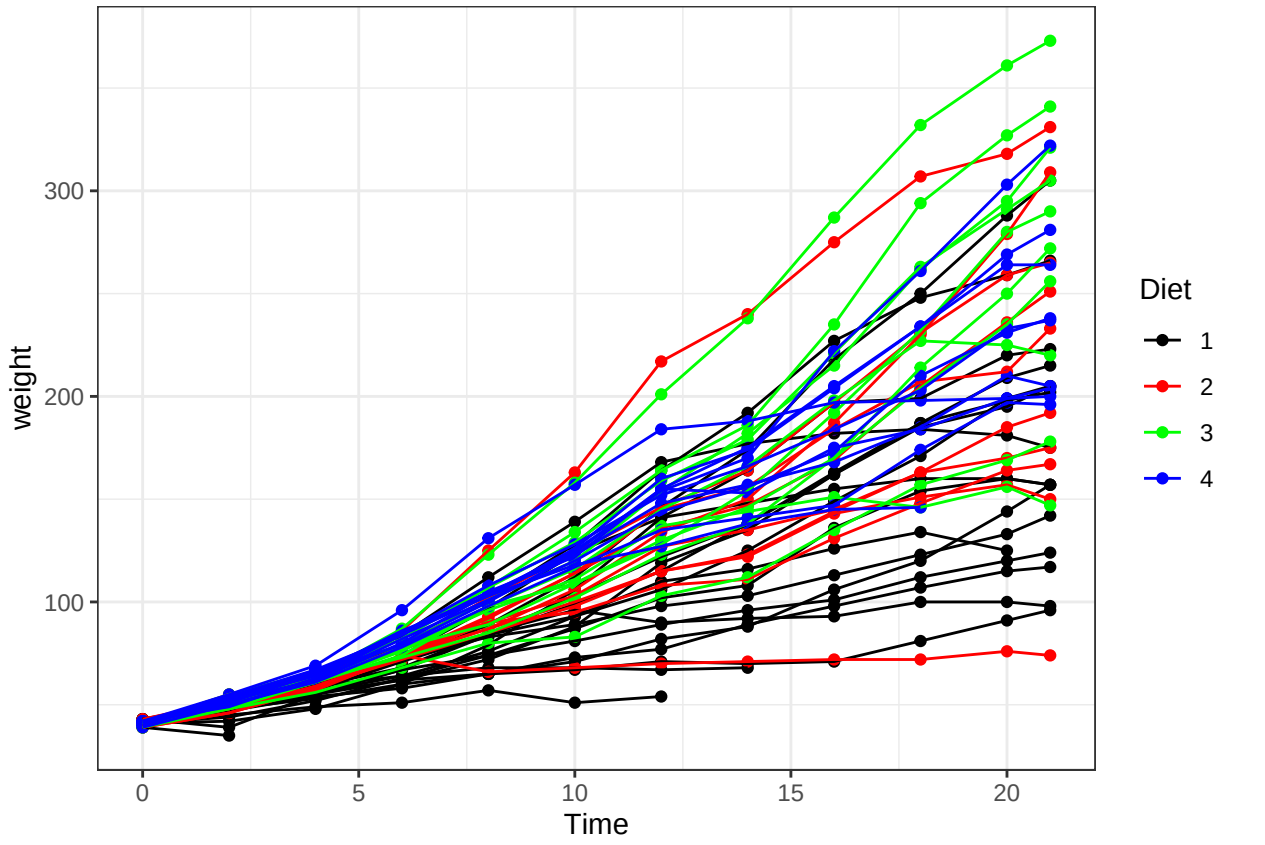
```
## $weight
## [1] "numeric"
##
## $Time
## [1] "numeric"
##
## $Chick
## [1] "ordered" "factor"
##
## $Diet
## [1] "factor"
```

(b) Visualize the weights of each chicks over time in one plot, where (i) each chick is represented by one solid curve, and (ii) the `diet` is color-coded as black (1), red (2), green (3), and blue (4). In addition to the required visualization, you may add any supporting curves, symbols, or any additional plots that you find informative.

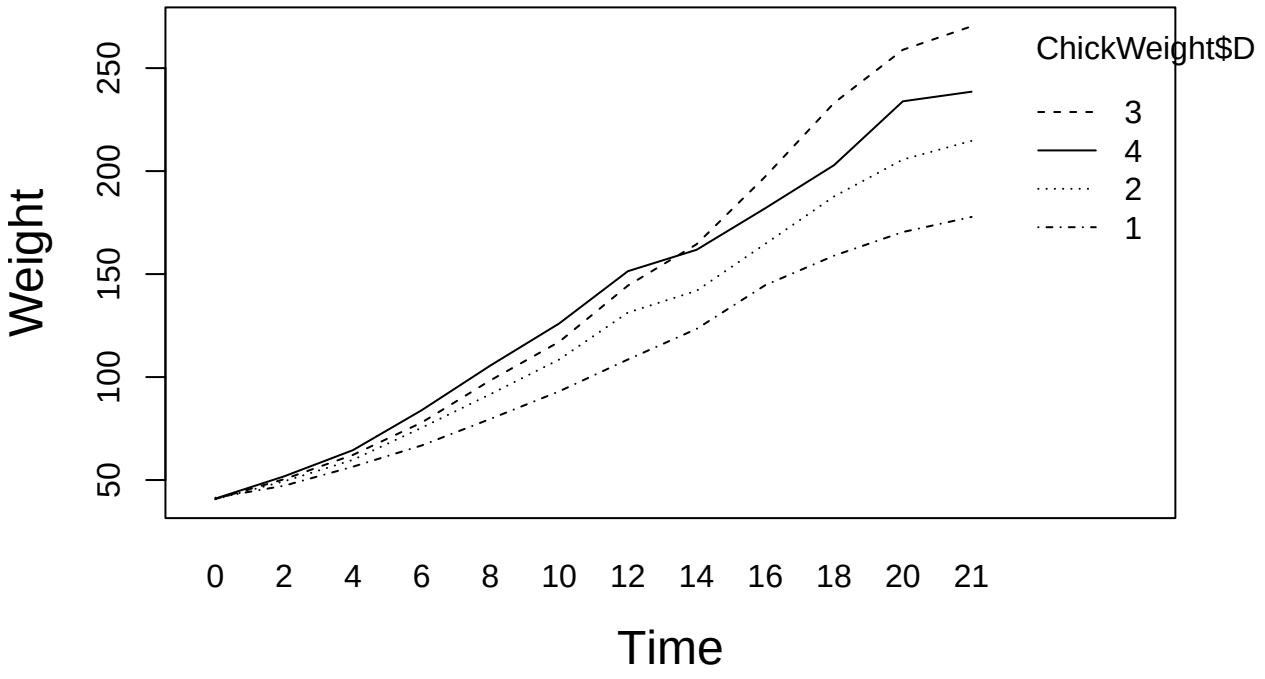
Solution: (Type your answer here)

```
# (b). (Type your code in the space below, if any)
library(ggplot2)
ggplot(ChickWeight, aes(x = Time, y = weight, group = Chick, col = Diet)) +
  geom_point() + stat_summary(fun = mean, geom = "line") + theme_bw() + scale_color_manual(values = c("
                                "red",
```

```
                                "green",
                                "blue"))
```




```
interaction.plot(ChickWeight$Time, ChickWeight$Diet, ChickWeight$weight
                 ,cex.lab=1.5,ylab="Weight",xlab='Time')
```



(c) Write down an appropriate one-way ANOVA model to answer the question whether there is any changes in mean weights at Day 20 across the four diet group. To receive full credits, you need to (i) write down the model, explain your notation, constraint(s) and/or assumptions; (ii) state the null and alternative hypotheses; (iii) state the test result. You can find basic LaTeX commands at the end of this file.

$$Y_{i,j} = \mu + \alpha_i + \epsilon_{i,j},\ j = 1, \ldots, n_i, i = 1, \ldots, 4$$

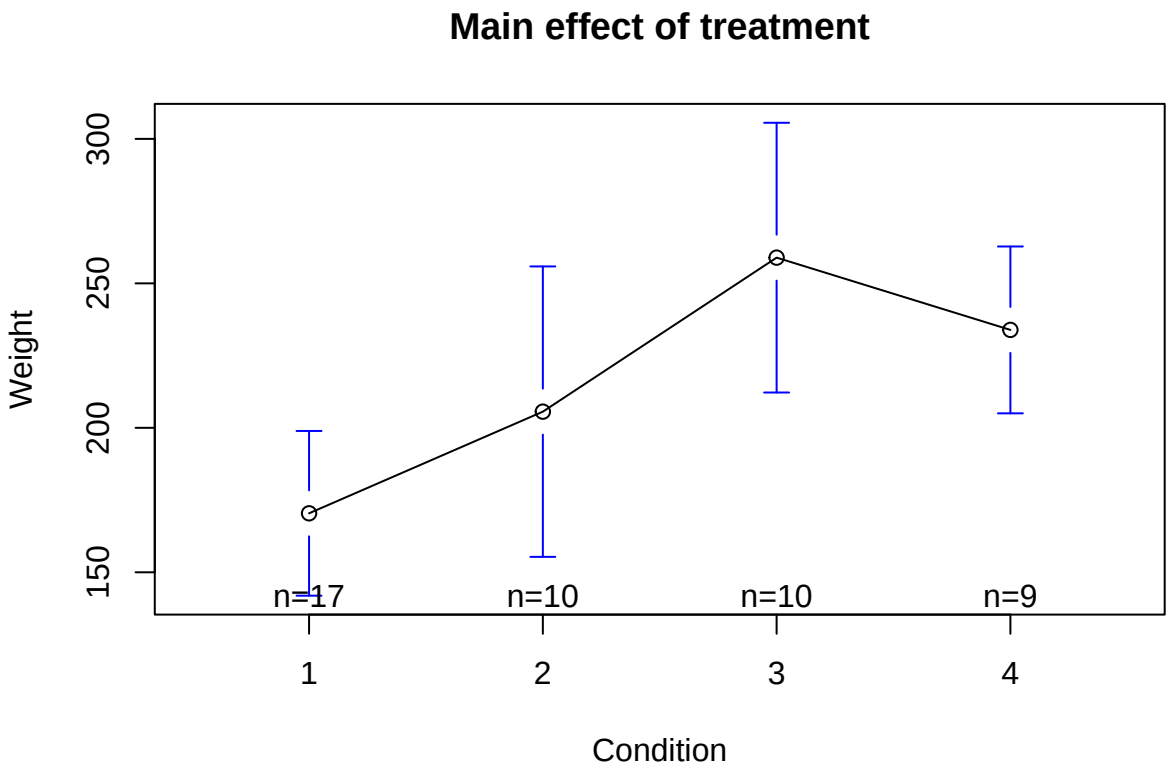
where $\{\alpha_i\}$ satisfies that $\sum_{i=1}^{4} n_i \alpha_i = 0$ and $\{\epsilon_{i,j}\}$ are i.i.d. $N(0, \sigma^2)$.

```
# (c). (Type your code in the space below, if any)

day_20 <- ChickWeight[ChickWeight$Time == 20,]

library(gplots)
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
plotmeans(weight ~ Diet, data = day_20,
          xlab = "Condition", ylab = "Weight",
          main="Main effect of treatment")
```




```
res.aov <- aov(weight ~ as.factor(Diet), data = day_20)
summary(res.aov)
```

```
##                  Df Sum Sq Mean Sq F value  Pr(>F)
## as.factor(Diet)   3  55881   18627   5.464 0.00291 **
## Residuals        42 143190    3409
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

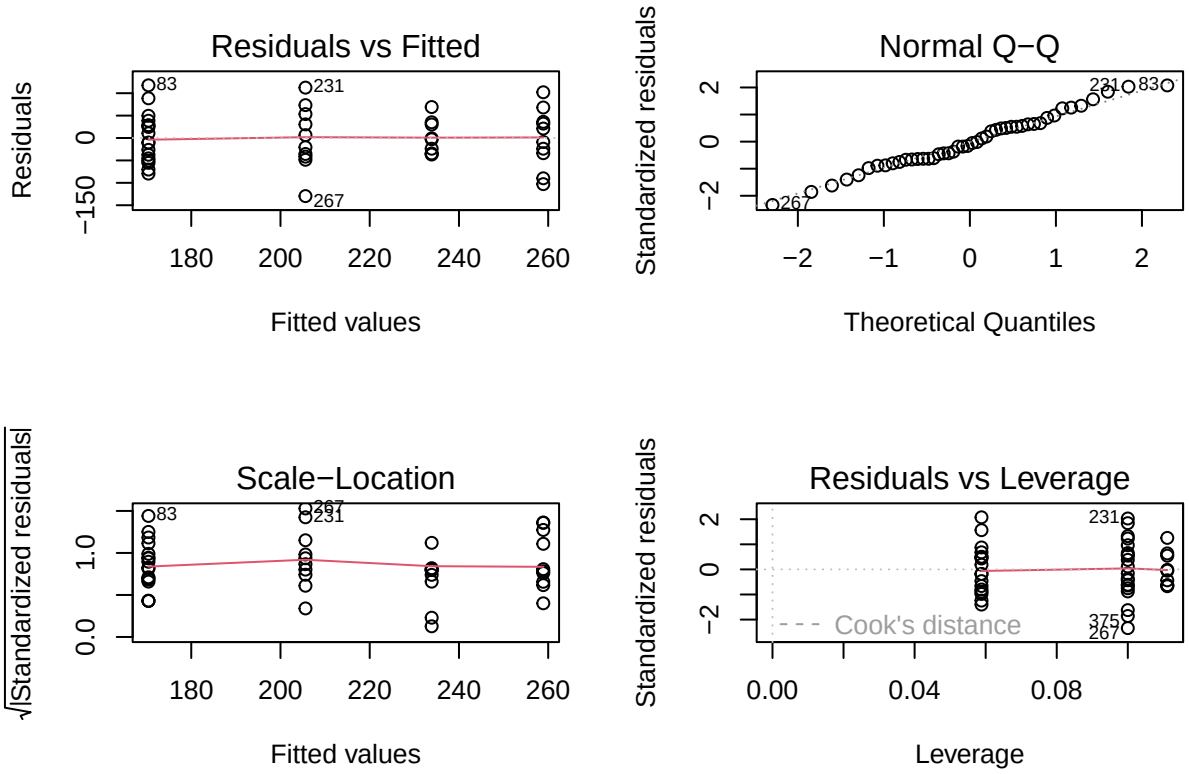
$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0 \quad \text{v.s.} \quad H_A : \text{not all } \alpha_i \text{ are the zero.}$$

We use the F-statistic to test this hypothesis.

(d) For the model fitted in (c), carry out necessary diagnostics to check if the model assumptions are valid. What are your findings?

Solution: (Type your answer here)

```
# (d). (Type your code in the space below, if any)
par(mfrow=c(2,2))
plot(res.aov)
```



Variance Homogeneity

```
bartlett.test(weight ~ as.factor(Diet), data = day_20)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  weight by as.factor(Diet)
## Bartlett's K-squared = 3.2498, df = 3, p-value = 0.3547
```

```
library(car)
```

```
## Loading required package: carData
```

```r
leveneTest(weight ~ as.factor(Diet), data = day_20)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  3  1.1111 0.3553
##       42
```

Normality

```r
# extract the residuals
aov_residuals <- residuals(object = res.aov)
shapiro.test(x = aov_residuals )
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aov_residuals
## W = 0.98573, p-value = 0.8378
```

```r
kruskal.test(weight ~ as.factor(Diet), data = day_20)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  weight by as.factor(Diet)
## Kruskal-Wallis chi-squared = 12.852, df = 3, p-value = 0.004969
```

(e) Write down an appropriate two-way ANOVA model with fixed effect to answer the question whether there is any differences in growth rates across the four diet groups. Here the growth rate can be roughly seen as the effects of `Time` on `weight`. To receive full credits, you need to (i) write down the model, explain your notation, constraint(s) and/or assumptions; (ii) state the null and alternative hypotheses; (iii) state the test result. Hint: You may want to recycle the answer in (c) to save time.

The two-way ANOVA model is as follows:

$$Y_{i,j,k} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{i,j} + \epsilon_{i,j,k}$$

over the cells enumerated as:

$$1 \leq i \leq a, \quad 1 \leq j \leq b, \quad 1 \leq i \leq n_{i,j}$$

subject to the constraints:

$$n_{\cdot,j} = \sum_{i=1}^{a} n_{i,j} \quad \forall j$$

$$n_{i,\cdot} = \sum_{j=1}^{b} n_{i,j} \quad \forall i$$

$$\sum_{i=1}^{a} n_{i,\cdot}\alpha_i = 0$$

$$\sum_{j=1}^{b} n_{\cdot,j}\beta_j = 0$$

$$\sum_{i=1}^{a} n_{i,j}(\alpha\beta)_{i,j} = 0 \text{ for all } j$$

$$\sum_{j=1}^{b} n_{i,j}(\alpha\beta)_{i,j} = 0 \text{ for all } i$$

```
# (e). (Type your code in the space below, if any)
model1 = aov(weight ~ as.factor(Diet) + as.factor(Time), data=ChickWeight)
summary(model1)
```

```
##                  Df  Sum Sq Mean Sq F value Pr(>F)
## as.factor(Diet)   3  155863   51954   40.75 <2e-16 ***
## as.factor(Time)  11 2040908  185537  145.53 <2e-16 ***
## Residuals       563  717785    1275
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model2 <- aov(weight ~ as.factor(Diet) * as.factor(Time), data=ChickWeight)
summary(model2)
```

```
##                                 Df  Sum Sq Mean Sq F value   Pr(>F)
## as.factor(Diet)                  3  155863   51954  43.631  < 2e-16 ***
## as.factor(Time)                 11 2040908  185537 155.812  < 2e-16 ***
## as.factor(Diet):as.factor(Time) 33   86676    2627   2.206 0.000172 ***
## Residuals                      530  631110    1191
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(f) We want to take the chick-specific effect into account. The new mixed effect model is based on the model in (e), where `Time` is treated as a continuous covariate instead of a categorical factor, and a random intercept and a random slope (of `Time`) are added into the model. Report the fitted coefficients of the fixed effects, and summarize your findings from this model. Hint: You do not need to write down the new model, but you may find it helpful.

Solution: (Type your answer here)

```
fit.weights <- lmer(weight ~ Diet + (1 | Time) +
                      (1 | Diet:Time), data = ChickWeight)
summary(fit.weights)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: weight ~ Diet + (1 | Time) + (1 | Diet:Time)
##    Data: ChickWeight
##
## REML criterion at convergence: 5794
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.1302 -0.3252 -0.0182  0.2625  3.4449
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  Diet:Time (Intercept)  118.2   10.87
##  Time      (Intercept) 4052.9   63.66
```

```
##  Residual              1187.9   34.47
## Number of obs: 578, groups:  Diet:Time, 48; Time, 12
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  106.005     18.789   5.642
## Diet2         16.611      5.918   2.807
## Diet3         36.945      5.918   6.243
## Diet4         30.869      5.932   5.204
##
## Correlation of Fixed Effects:
##       (Intr) Diet2  Diet3
## Diet2 -0.137
## Diet3 -0.137  0.436
## Diet4 -0.137  0.435  0.435
```

```
anova(fit.weights)
```

```
## Analysis of Variance Table
##      npar Sum Sq Mean Sq F value
## Diet    3  56090   18697   15.74
```

(g) Assume that the chicks in each diet are randomly selected from the same population, i.e., the enrollment of chicks is independent from any other factors. State the Stable Unit Treatment Value Assumption, write down the potential outcomes (weight at Day 20), and verify whether the randomization assumption holds. (This question will be replaced by another, since causal inference will not be covered this quarter.)

Causal Assumptions:

1. SUTVA: Stable Unit Treatment Value Assumption: This is actually two assumptions.

- No interference: the treatment assigned to one unit does not affect the outcome of another unit. We typically assume no spillover.
- One version of treatment: Outcomes need to be linked to observable data. No missing features.

2. Consistency: The potential outcome under the treatment a is equal to the observed outcome if the actual treatment recieved is a.

$$Y = Y^a \text{ if } A = a, \text{ for all } a$$

3. Ignorability: Treatment assignment is independent from the potential outcomes:

$$Y^0, Y_1 \perp\!\!\!\perp A|X$$

Treatment should be randomly assigned amongst X.

4. Positivity: For every set of values for X, the treatment assignment was not deterministic:

$$P(A = a|X = x) > 0 \text{ for all } a \text{ and } x$$

LaTeX commands

$$Y_{i,j,k}, \mu, \alpha, \beta, \epsilon, i, j, H_0, H_a, \neq, =, ...$$