MS Comprehensive Exam 2022

STA 207 (100 points)

```
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)

weight: Numeric variable giving the body weight of the chick.

Time: Numeric variable giving the number of days since birth when measure was made.

Chick: Factor with 50 levels giving a unique identifier for the chick.

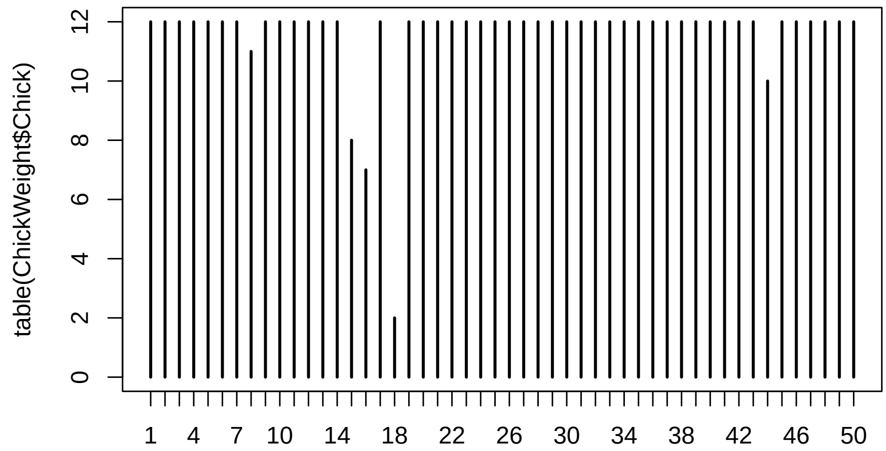
Diet: Factor with 4 levels indicating which experimental diet the chick received.

```
# str(ChickWeight)
# nlevels(ChickWeight$Chick)
# levels(ChickWeight$Chick)
```

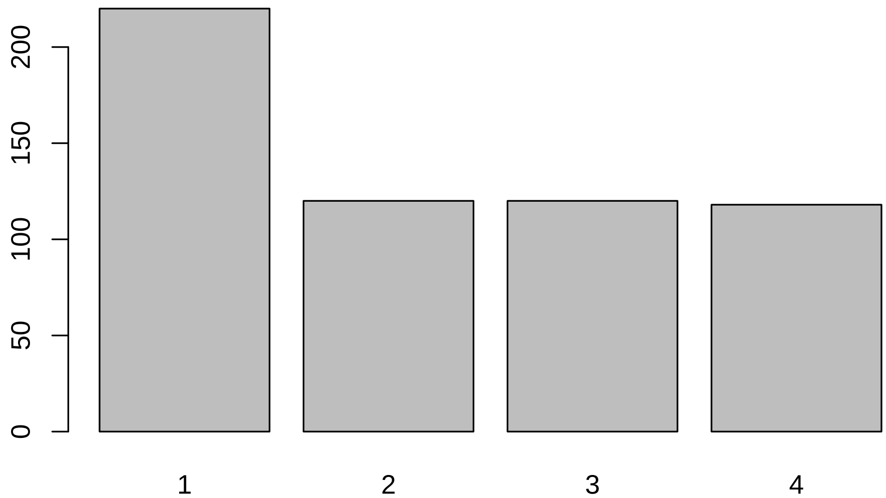
```
summary_table = summary(ChickWeight)
summary_table
```

```
##      weight           Time           Chick     Diet
##  Min.   : 35.0   Min.   : 0.00   13     : 12   1:220
##  1st Qu.: 63.0   1st Qu.: 4.00   9      : 12   2:120
##  Median :103.0   Median :10.00   20     : 12   3:120
##  Mean   :121.8   Mean   :10.72   10     : 12   4:118
##  3rd Qu.:163.8   3rd Qu.:16.00   17     : 12
##  Max.   :373.0   Max.   :21.00   19     : 12
##                                  (Other):506
```

```
plot(table(ChickWeight$Chick))
```




```
barplot(table(ChickWeight$Diet))
```

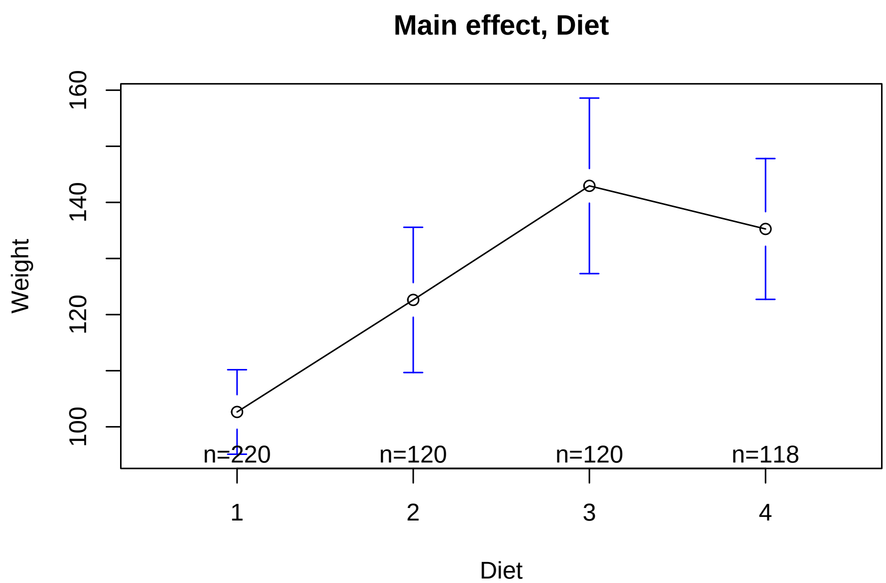



```
# Get summary using Group_by using dplyr
library(dplyr)
summary_table2 = ChickWeight %>% group_by(Diet) %>%
  summarise(
    count=n(),
    min=min(weight),
    Q1=quantile(weight, 0.25),
    mean=mean(weight),
    SD =sd(weight),
    Q3 = quantile(weight, 0.75),
    max=max(weight)
  )
summary_table2
```

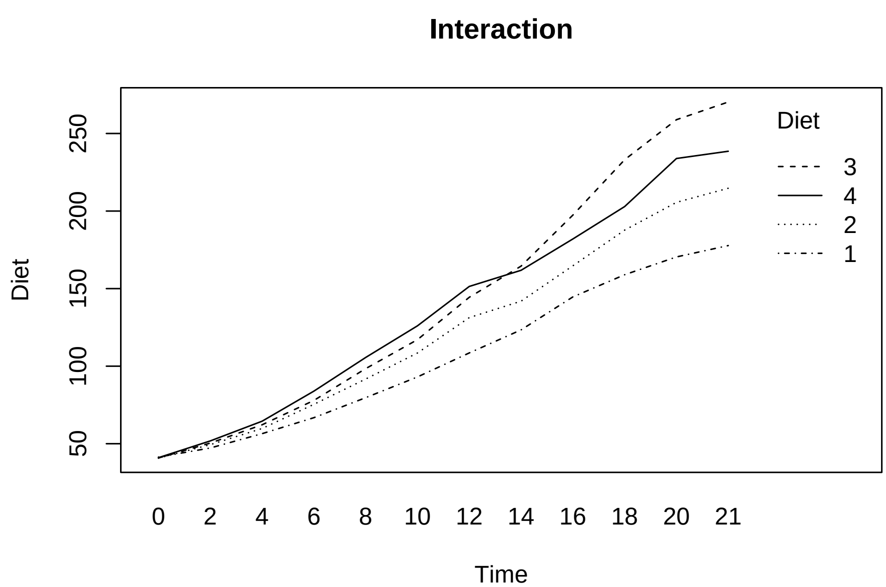
```
## # A tibble: 4 x 8
##   Diet  count   min    Q1  mean    SD    Q3   max
##   <fct> <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 1       220    35  57.8  103.  56.7  136.   305
## 2 2       120    39  65.5  123.  71.6  163    331
```

```
## 3 3        120     39   67.5   143.   86.5   199.    373
## 4 4        118     39   71.2   135.   68.8   185.    322
```

```r
library(gplots)
plotmeans(weight~Diet, data=ChickWeight, xlab="Diet", ylab="Weight",
          main="Main effect, Diet")
```




```r
with(ChickWeight, interaction.plot(x.factor=Time,
                                   trace.factor=Diet,
                                   response=weight,
                            xlab="Time",ylab="Diet", main="Interaction"))
```

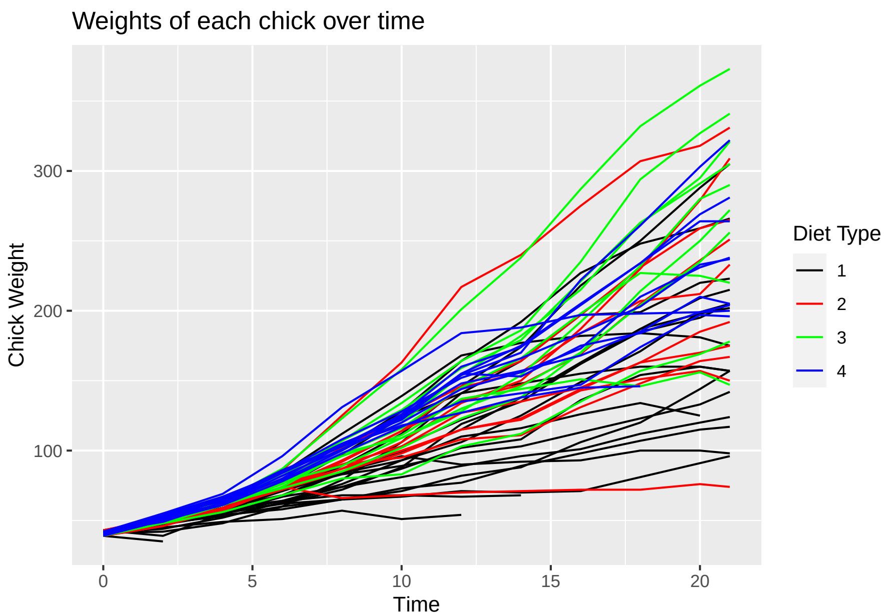


(b) Visualize the weights of each chicks over time in one plot, where (i) each chick is represented by one solid curve, and (ii) the `diet` is color-coded as black (1), red (2), green (3), and blue

(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)

```
# Plot and add title, x and y labels, and legend title. (MAKE SURE TO CHANGE THOSE!)
library(ggplot2)
ggplot(ChickWeight, aes(x=Time, y=weight, group=Chick, color=Diet)) +
  geom_line() +
  scale_color_manual(values = c("black", "red", "green", "blue")) +
  ggtitle("Weights of each chick over time") +
  xlab("Time") + ylab("Chick Weight") + labs(colour="Diet Type")
```



(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.

Solution: (Type your answer here)

One Way Anova Factor Effects Model:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} i = 1, 2, 3, 4 j = 1, \ldots, n_i$$

With Assumptions:

$$\epsilon_{ij} \overset{i.i.d}{\sim} N(0, \sigma^2)$$

With Constraints:

$$\sum_{i=1}^{4} n_i \alpha_i = 0$$

Y_{ij}: represents the **weight** (response variable) at Day 20 of the j-th **chick** (observation) in the i-th **Diet** (factor effect).

μ: represents the population mean **weight** (response variable).

α_i: represents the factor effects of **Diet** (factor effect).

ϵ_{ij}: capture any unexplained effects on **weight** (response variable).

n_i: sample size for the i-th **Diet** (factor effect).

Hypothesis Testing: Testing at a significance level of 0.05

$$H_0 : \alpha_i = 0 \text{ for all i} H_a : \text{at least one } \alpha_i \neq 0$$

```r
# Create a dataframe that keeps only Day 20 values
Day20 = ChickWeight %>% filter(Time == 20)

# Fit One Way Anova model
model1 = lm(weight~as.factor(Diet), data= Day20)
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: weight
##                 Df Sum Sq Mean Sq F value   Pr(>F)
## as.factor(Diet)  3  55881 18627.0  5.4636 0.002909 **
## Residuals       42 143190  3409.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
## Alt. way to fit and get p-values
anova.fit = aov(weight~as.factor(Diet), data=Day20)
summary(anova.fit)
```

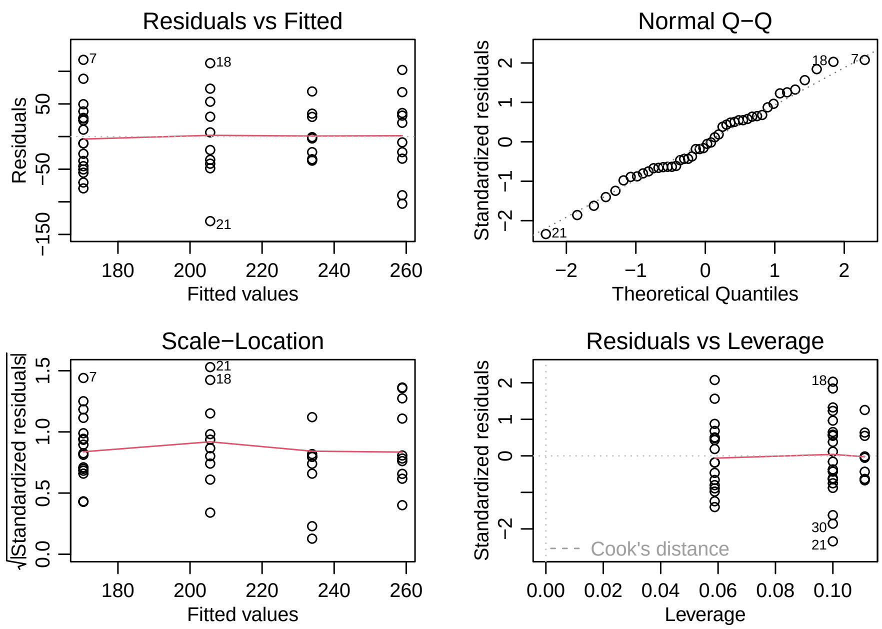
```
##                 Df Sum Sq Mean Sq F value  Pr(>F)
## as.factor(Diet)  3  55881   18627   5.464 0.00291 **
## Residuals       42 143190    3409
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the anova table, since p-value is less than 0.05, we reject the null hypothesis and conclude that there is evidence in the data to suggest that there is significant effect of at least one diet at Day 20.

(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)

```r
par(mfrow=c(2,2), mar=c(3,3,2,2), mgp=c(1.7,.7,0))
plot(anova.fit)
```



In the Residual vs. Fitted value plot, we see no sign of unequal variance or violation of the zero-mean assumption.
The Normal Q-Q plot appears to be normal and satisfy the normality assumption.
All assumptions are met and our model is valid.

(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.

Solution: (Type your answer here)

Two Way Anova Factor Effects Model (Fixed Effects Model):

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} i = 1, \dots, 4 j = 1, \dots, 12 k = 1, \dots, n_{ij}$$

With Assumptions:

$$\epsilon_{ijk} \overset{i.i.d}{\sim} N(0, \sigma^2)$$

With Constraints:

$$\sum_{i=1}^{4} \alpha_i = \sum_{j=1}^{12} \beta_j = 0 \sum_{i=1}^{4} (\alpha\beta)_{ij} = \sum_{j=1}^{12} (\alpha\beta)_{ij} = 0 \ \forall i, j$$

Y_{ijk}: represents the **weight** (response variable) of the k-th **chick** (observation) at j-th **Time** (factor 2 effect) in the i-th **Diet** (factor 1 effect).

$\mu_{..}$: represents the population mean **weight** (response variable).

α_i: represents the factor effects of **Diet** (factor 1 effect).

β_j: represents the factor effects of **Time** (factor 2 effect).

$(\alpha\beta)_{ij}$: represents the interaction term of factor effects of **Time** and **Diet**.

ϵ_{ijk}: capture any unexplained effects on **weight** (response variable).

n_{ij}: sample size for the i-th **Diet** and the j-th **Time** .

$$\mu_{..} = \sum_{i=1}^{4}\sum_{j=1}^{12}\frac{\mu_{ij}}{(4*12)},\ \mu_{i.} = \sum_{j=1}^{12}\frac{\mu_{ij}}{12},\ \mu_{.j} = \sum_{i=1}^{4}\frac{\mu_{ij}}{4}$$

Hypothesis Testing: Testing interaction term at a significance level of 0.05

$$H_0 : \text{All } (\alpha\beta)_{ij} = 0 H_a : \text{At least one}(\alpha\beta)_{ij} \neq 0$$

```r
# Fit Two way Anova model (Fixed Effects)
model2 = lm(weight ~ as.factor(Diet) * as.factor(Time), data=ChickWeight)
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: weight
##                                   Df  Sum Sq Mean Sq  F value    Pr(>F)
## as.factor(Diet)                    3  155863   51954  43.6307 < 2.2e-16 ***
## as.factor(Time)                   11 2040908  185537 155.8123 < 2.2e-16 ***
## as.factor(Diet):as.factor(Time)   33   86676    2627   2.2057  0.000172 ***
## Residuals                        530  631110    1191
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
## Alt. way to fit and get p-values
two_anova_fit = aov(weight ~ as.factor(Diet) * as.factor(Time), data=ChickWeight)
summary(two_anova_fit)
```

```
##                                   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
```

```r
## Alt. way to fit and get p-values
# Full model
full = aov(weight ~ as.factor(Diet) * as.factor(Time), data=ChickWeight)
# Reduced
reduced = aov(weight ~ as.factor(Diet) + as.factor(Time), data=ChickWeight)
# Full/reduced test
anova(reduced, full)
```

```
## Analysis of Variance Table
##
## Model 1: weight ~ as.factor(Diet) + as.factor(Time)
## Model 2: weight ~ as.factor(Diet) * as.factor(Time)
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    563 717785
## 2    530 631110 33     86676 2.2057 0.000172 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the anova table the p-value of the interaction term is less than $\alpha = 0.05$, thus we reject the null hypothesis. We conclude that the interaction term is significant enough to be included in our model. There is evidence in the data to suggest that growth rates differ between different diets.

(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)

The coefficients for the mixed model are reported below. The intercept is the reference class, and represents Diet1 in this model.

```
# Fit Two way Anova model (Mixed Effects)
## (1|Chick) is the random effect of Chick
## (0 + Time|Chick) is the random slope of time given a chick
mixed_model = lmer(weight ~ Time * as.factor(Diet) + (1|Chick) + (0 + Time|Chick),
                   data=ChickWeight)
# Report the coefficients (of the fixed effects)
summary(mixed_model)$coefficients
```

```
##                          Estimate Std. Error    t value
## (Intercept)             33.415500  2.6855268 12.4428103
## Time                     6.280985  0.7091122  8.8575339
## as.factor(Diet)2        -4.781905  4.6081812 -1.0376989
## as.factor(Diet)3       -15.165175  4.6081812 -3.2909242
## as.factor(Diet)4        -1.549349  4.6159467 -0.3356514
## Time:as.factor(Diet)2    2.328151  1.2065259  1.9296319
## Time:as.factor(Diet)3    5.141886  1.2065259  4.2617283
## Time:as.factor(Diet)4    3.258262  1.2073132  2.6987709
```

(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.)

Solution: (Type your answer here)