HW 4

Andrew Jowe

1

a. The errors seem to be normally distributed because the dots on the QQ plot are close to the QQ line.
b. There seems to be outliers because there are dots that are more than 5 errors away from the fitted mean on the errors vs means plot. ??
c. It appears that the groups don't have equal variances because the vertical spread of the dots are not the same for all fitted means on the errors vs fitted means plot.
d. I would not remove any observations because the sample size for each group is already pretty small. More data means that our estimates are more accurate.

2

a. Since the p-value is typically greater than α (at most 0.1), we fail to reject the null hypothesis. Therefore, the data follows a normal distribution.
b. Since the p-value is typically greater than α (at most 0.1), we fail to reject the null hypothesis. Therefore, the variances of each group are approximately equal.
c. I would not suggest transforming the data because our data already fits the ANOVA assumptions.
d. Our confidence intervals and p-values are likely to be reliable because we know the ANOVA assumptions hold.

3

a. It appears that the ANOVA assumptions are not met, because the graphs suggest that the data may not have equal variances between groups, and we cannot say for sure if the data follows a normal distribution. The Errors vs Group Means graph has different spreads for all group means, suggesting non-equal variances. The dots on the normal QQ plot do not follow the QQ line for the points beyond the ± 1 theoretical quantiles, suggesting that the data may not be normally distributed. However, the histogram of errors show that the data is approximately normal. In the best case scenario, we need to remove outliers so that the data follows a normal distribution.
b. The null hypothesis for the Shapiro-Wilks test is that the data follows a normal distribution. The alternative hypothesis is that the data does not follow a normal distribution.
c. The null hypothesis for the Brown-Forstyhe test is that the groups all have equal variances. The alternative hypothesis is that the groups do not all have equal variances (at least one of the variances are not equal).
d. No, because we are not sure if the ANOVA assumptions hold.

4

a. We have a 0.0734 chance of a type I error, where we reject the null hypothesis when the null hypothesis is true. In other words, this is the chance of stating that the data is not normal when it actually is.
b. We have a 0.0000304 chance of a type I error, where we reject the null hypothesis when the null hypothesis is true. In other words, this is the chance of stating that at least one of the variances are not equal when all groups actually have equal variances.
c. We would suggest transforming the variables because we conclude the variances are not equal between all groups as we will probably reject the null hypothesis of the Brown Forsythe test using typical α values.
d. One downside is that it can make the interpretation of the data more complex. When we interpret, now we have to account for the function we used to transform the data.

5

a. Yes, because the dots on the QQ plot are a lot closer to the line compared to before.
b. No, although the p-value for the Shapiro Wilks test is a lot higher than before, making it more likely that we would fail to reject the null hypothesis or accept that the data is normally distributed, the Brown Forsythe test did not improve that much as it's p-value is still lower than 0.01. We will most likely still reject the null hypothesis for that test given typical α values, concluding that the groups do not all have equal variances, violating our ANOVA assumptions.
c. Yes, it appears that the groups have constant variances as the spreads for the different group means are around the same on the Errors vs Group Means chart.
d. No, although the p-value for the Brown Forsythe test is a lot higher than before, making it more likely that we would fail to reject the null hypothesis or accept that the data is normally distributed, the Shapiro Wilks test did not improve that much as it's p-value is still lower than 0.01. We will most likely still reject the null hypothesis for that test given typical α values, concluding that the data is not normally distributed, violating our ANOVA assumptions.
e. None of the datasets are suitable because all of them violate ANOVA assumptions.

6

a. $Y_{ijk} = \mu_{..} + \gamma_i + \delta_j + \epsilon_{ijk}$ The assumptions and constraints are:
    i. The ANOVA assumptions hold, that is:
        - All Y_{ijk} were randomly sampled
        - The i groups are independent of each other
        - The j groups are independent of each other
        - $\epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$
    ii. Observations should be independent between groups as a constraint since there is no factor effect between groups in our model.
b. $Y_{ijk} = \mu_{..} + \gamma_i + \delta_j + (\gamma\delta)_{ij} + \epsilon_{ijk}$ The assumptions are:
    i. The ANOVA assumptions hold, that is:
        - All Y_{ijk} were randomly sampled
        - The i groups are independent of each other
        - The j groups are independent of each other
        - $\epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$
    ii. Observations can be dependent between groups since there is a factor effect between groups in our model.

c. There seems to be a significant effect of smoking cigarettes on sleep, because there is a change in sleep when we change this factor. The average sleep in hours for those who smoke is 5.90 compared to 7.04 for those who don't smoke.
d. There seems to be a significant effect of smoking marijuana on sleep, because there is a change in sleep when we change this factor. The average sleep in hours for those who smoke is 6.71 compared to 7.073 for those who don't.
e. There seems to be an interaction effect between smoking cigarettes and smoking marijuana on sleep. For those who don't smoke cigarettes, smoking marijuana decreases the average sleep by 0.18 hours. However, for those who do smoke cigarettes, smoking marijuana increases the average sleep by 0.97.

7

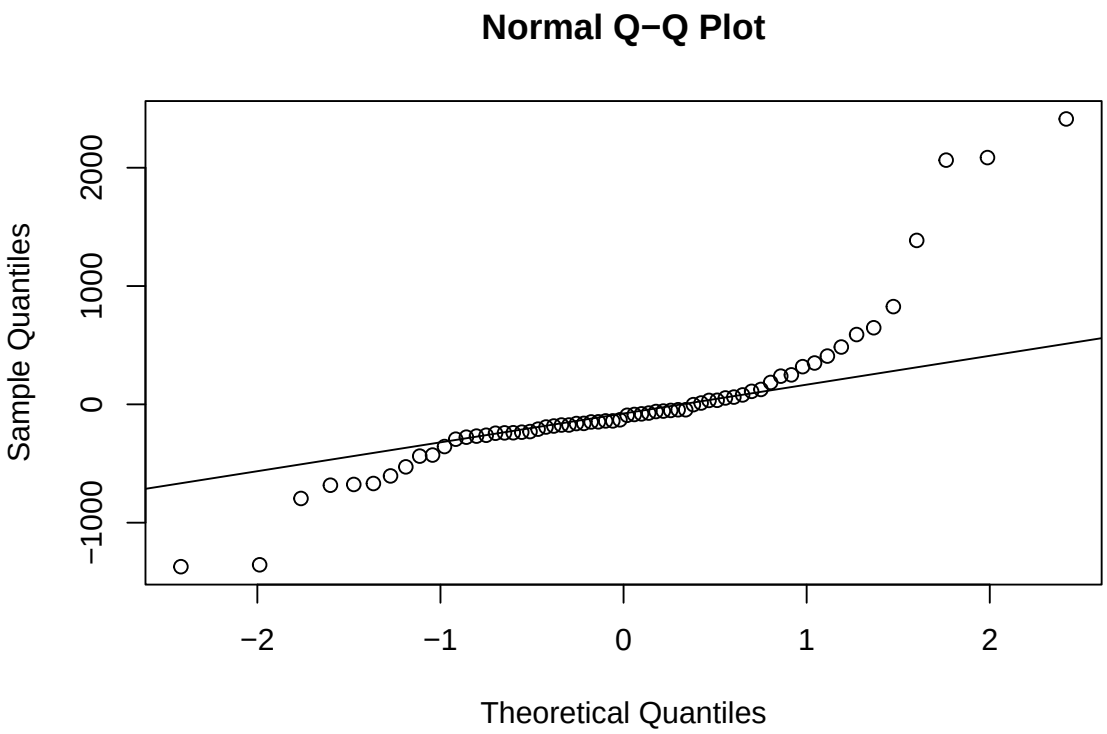
a. We need two values of γ_i to estimate since there are two possibilities for smoking cigarettes.
b. We need two values of δ_j to estimate since there are two possibilities for smoking marijuana.
c. We need four values of $(\gamma\delta)_{ij}$ to estimate since there are two possibilities for smoking cigarettes multiplied by two possibilities for smoking marijuana.
d. In total, we have $(a-1)+(b-1)+1$ parameters to estimate. In our model $Y_{ijk} = \mu_{..} + \gamma_i + \delta_j + \epsilon_{ijk}$, we have 1 parameter to estimate for $\mu_{..}$, $a-1$ parameters to estimate for γ_i, and $b-1$ parameters to estimate for δ_j.
e. In total, we have $(a-1)+(b-1)+(a-1)(b-1)+1$ parameters to estimate. In our model $Y_{ijk} = \mu_{..} + \gamma_i + \delta_j + (\gamma\delta)_{ij} + \epsilon_{ijk}$, we have 1 parameter to estimate for $\mu_{..}$, $a-1$ parameters to estimate for γ_i, $(a-1)(b-1)$ parameters to estimate for $(\gamma\delta)_{ij}$, and $b-1$ parameters to estimate for δ_j. If we simplify our result: $(a-1)+(b-1)+(a-1)(b-1)+1 \rightarrow a-1+b-1+a*b-a-b+1+1 \rightarrow a+b+a*b-a-b \rightarrow a*b$ $\therefore$ We have $a*b$ total parameters.

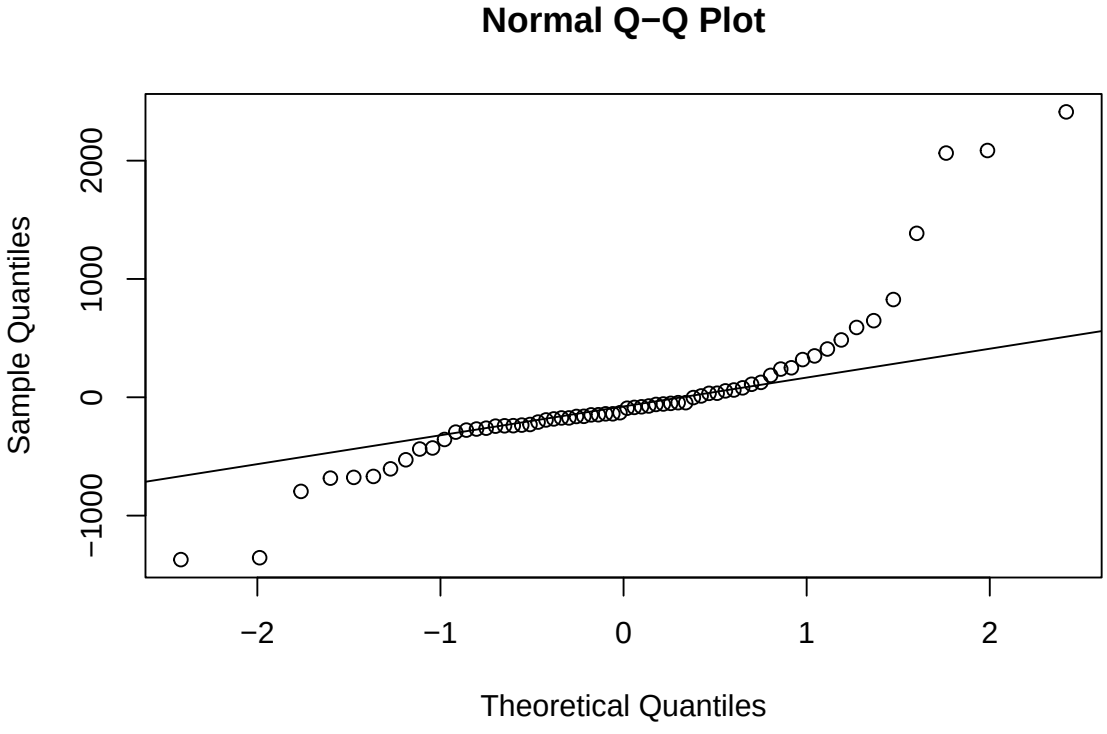
8

a. False. Our null hypothesis for this test is that the data is normal. A smaller p-value makes it less likely for us to accept the null hypothesis, making it more likely that the data is not normally distributed.
b. False. It is possible for the normality assumption to hold, but i groups have difference variances.
c. True. This matches our answer in 7d.
d. False. This does not match our answer in 7e.

I

a. In the Normal Q-Q plot, the dots beyond ± 1 theoretical quantiles do not conform to the normal line. The graph suggests that the data is not normally distributed, violating our ANOVA assumptions.


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−1000
0
1000
2000
−2
−1
0
1
2


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−1000
0
1000
2000
−2
−1
0
1
2

b.

c. $p = 0$

d. $p = 0.0033$

e. 0% of the data was removed.

II

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.

## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.

## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

a.

Appendix

I

```
library(car)
cancer <- read.csv("Cancer.csv")
cancer_model <- lm(Survival ~ Organ, data = cancer)
```

I-a

```
qqnorm(cancer_model$residuals)
qqline(cancer_model$residuals)
```

I-b

```
alpha <- 0.01

cancer$ei <- cancer_model$residuals
nt <- nrow(cancer)
a <- length(unique(cancer_model$Organ))
r_ij <- rstandard(cancer_model)

t_cutoff <- qt(1 - alpha / (2 * nt), nt - a)
cancer$is_outlier <- abs(r_ij) > t_cutoff
num_outliers_removed <- sum(cancer$is_outlier == TRUE)

cancer_new <- cancer[!cancer$is_outlier, ]
cancer_model_new <- lm(Survival ~ Organ, data = cancer)
```

```r
qqnorm(cancer_model_new$residuals)
qqline(cancer_model_new$residuals)
```

I-c

```r
ei <- cancer_model_new$residuals
sw_test <- shapiro.test(ei)
```

I-d

```r
the_bf_test <- leveneTest(ei ~ Organ, data = cancer_new, center = median)
the_p_value <- the_bf_test[[3]][1]
```

I-e

```r
portion_of_data_removed <- num_outliers_removed / nt
```

II

```r
library("EnvStats")

do_diagnostic <- function(y, x, data) {
  model <- lm(y ~ x, data = data)
  ei <- model$residuals
  sw_test <- shapiro.test(ei)
  sw_test_p <- sw_test$p.value
  the_bf_test <- leveneTest(ei ~ x, data = data, center = median)
  the_p_value <- the_bf_test[[3]][1]
  return(c(sw_test_p, the_p_value))
}
```

II-a

```r
lambda_1 <- boxcox(
  cancer_model,
  objective.name = "PPCC",
  optimize = TRUE
)$lambda
lambda_2 <- boxcox(
  cancer_model,
  objective.name = "Shapiro-Wilk",
  optimize = TRUE
```

```
)$lambda
lambda_3 <- boxcox(
  cancer$Survival,
  objective.name = "Log-Likelihood",
  optimize = TRUE
)$lambda

yt_l1 <- (cancer$Survival ** lambda_1 - 1) / lambda_1
t_cancer_l1 <- data.frame(Survival = yt_l1, Organ = cancer$Organ)
t_cancer_l1_res <- do_diagnostic(
  t_cancer_l1$Survival,
  t_cancer_l1$Organ,
  t_cancer_l1
)
sw_test_l1 <- t_cancer_l1_res[1]
bf_test_l1 <- t_cancer_l1_res[2]

yt_l2 <- (cancer$Survival ** lambda_1 - 1) / lambda_1
t_cancer_l2 <- data.frame(Survival = yt_l2, Organ = cancer$Organ)
t_cancer_l2_res <- do_diagnostic(
  t_cancer_l2$Survival,
  t_cancer_l2$Organ,
  t_cancer_l2
)
sw_test_l2 <- t_cancer_l2_res[1]
bf_test_l2 <- t_cancer_l2_res[2]

yt_l3 <- (cancer$Survival ** lambda_1 - 1) / lambda_1
t_cancer_l3 <- data.frame(Survival = yt_l3, Organ = cancer$Organ)
t_cancer_l3_res <- do_diagnostic(
  t_cancer_l3$Survival,
  t_cancer_l3$Organ,
  t_cancer_l3
)
sw_test_l3 <- t_cancer_l3_res[1]
bf_test_l3 <- t_cancer_l3_res[2]
```