UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Data Science for Biomedicine

Lab 1

Tim Boleslawsky

Exploratory Analysis

Table of the mean cholesterol by time and diet:

	Diet	Time	mean_cholesterol
1	1	PostChol	193.625
2	1	PreChol	240.000
3	2	PostChol	206.125
4	2	PreChol	230.750
5	3	PostChol	218.500
6	3	PreChol	225.250
7	4	PostChol	224.875
8	4	PreChol	221.750

Figure 1: Mean cholesterol by time and diet

From the table we can see that for diets 1-3 the cholesterol levels seem to have decreased on average after the treatment. For diet 4 we have an increase of cholesterol levels, although a small one, after the treatment. We can also see that diet 1 seemed to have the biggest effect. But we have to note that the pre-treatment cholesterol levels in the diet 1 group were the highest. This may have an effect on efficacy.

Box-plot by time and diet:

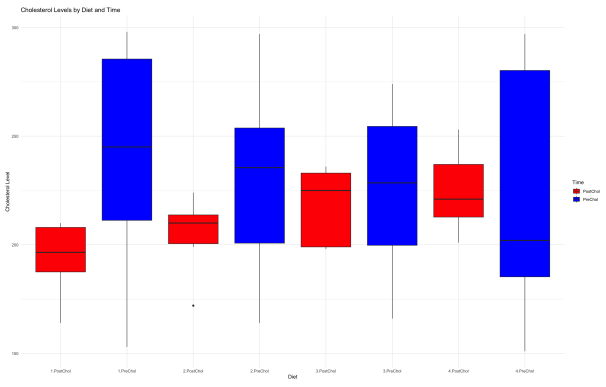

Figure 2: Box-plot by time and diet

The Box-plot confirms what we saw in the table. We additionally see that the spread of the post-treatment cholesterol levels seem to be much smaller. This is evident in all four diets.

When discussing statistical hypotheses to test the observed differences, we can go about it in two ways:

- If we want to compare the cholesterol levels before and after the treatment, we can use a paired t-test. Then the hypothesis would be for example: The mean

cholesterol level after treatment is lower than before. In this case we would not differentiate between diets. The null hypothesis would then be: $H_0 : \mu_p re = \mu_p ost$

- We also could compare the diets. Then we can use an ANOVA test to compare the post-treatment means of all the diets. The hypothesis then would for example be: "At least one diet reduces the cholesterol levels more than all the other diets". A possible null hypothesis would then be: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

Code:

```
# Exploratory Analysis
diet <- read.table(file="diet.txt",header=T,sep=" ")

# Table
diet_long <- diet %>% pivot_longer(cols = c(PreChol, PostChol),
    names_to = "Time", values_to = "Cholesterol")
diet_grouped <- diet_long %>% group_by(Diet, Time)
diet_summarized <- diet_grouped %>%
    summarise(mean_cholesterol = mean(Cholesterol))

# Boxplot
diet_boxplot <- ggplot(diet_grouped,
                       aes(x = interaction(Diet, Time, lex.order = TRUE),
                       y = Cholesterol, fill = Time)) +
                  geom_boxplot() +
                  labs(title = "Cholesterol Levels by Diet and Time",
                     x = "Diet",
                     y = "Cholesterol Level") +
                  theme_minimal() +
                  scale_fill_manual(values = c("PreChol" = "blue",
                      "PostChol" = "red"))
```

ANOVA Analysis

Table of the ANOVA analysis:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Diet	3	4594	1531.3	5.235	0.00538
Residuals	28	8190	292.5		

Figure 3: ANOVA Analysis

From the above table in figure 3, we can see the results of the one-way ANOVA analysis on the post-treatment data. The degree of freedom of 3 shows us that we compare the four different diets. The F-value of 5.235 indicates the ratio of variability between diet groups compared to variability within the groups. A higher F-value suggests that diet has a strong effect on cholesterol levels. The p-value is 0.00538, which is statistically significant at $\alpha = 0.05$. Since $p < 0.05$, we reject the null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$, meaning at least one diet group has a significantly different mean cholesterol level.

Table of the linear regression analysis on the diet data:

```
Call:
lm(formula = PostChol ~ Diet, data = diet)

Residuals:
    Min      1Q  Median      3Q     Max
-34.125 -12.125   2.687  14.406  28.125

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  193.625      6.047  32.022  < 2e-16 ***
Diet2         12.500      8.551   1.462  0.15493
Diet3         24.875      8.551   2.909  0.00703 **
Diet4         31.250      8.551   3.654  0.00105 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.1 on 28 degrees of freedom
Multiple R-squared:  0.3594,    Adjusted R-squared:  0.2907
F-statistic: 5.235 on 3 and 28 DF,  p-value: 0.005384
```

Figure 4: Linear Model Analysis

Instead of simply using the aov() function, we can also first fit a linear regression model and then perform an analysis on the output we get from this. In the table in figure 4, we can see the output of the linear regression model fitted on the diet data. The residuals represent the differences between observed and predicted values in our regression model. We want a median of zero, in this case it is 2.687, which suggests a slight model bias. We can also see that the min and max values are pretty far apart, which could be because we have outliers. In a linear regression model the coefficients tell us the effect of one variable on the target. In the output of this function, we can see the results of a t-test, which tests the statistical difference between diet 1 and the other diets. We can conclude from the p-values that diet 2 gives not significantly different outcomes to diet 1, while diet 3 and 4 do. We can also see that diet 4 is the most different from diet 1. The lm() function also provides us with a F-test to tell us, if the model as a whole is significant. In this case, we can see a p-value of 0.005384 which leads us to believe that at least one

diet significantly explains the outcome variable.

In this output, a few advantages of this approach, compared to the simple ANOVA approach, become apparent. First we can do regression-based interpretations. This means we can not only see, if there is a difference between the diets, we can also see the size of the difference. Second, we can work with covariates (ANCOVA) and are able to better handle categorical and continuous data. Third, and this might not be obvious from the above output, ANOVA (aov()) assumes balanced designs (equal sample sizes across groups), while regression + ANOVA handles unbalanced designs more robustly.

Output of the function emmeans(), when used on the result of the ANOVA analysis:

```
Diet emmean   SE df lower.CL upper.CL t.ratio p.value
1       194 6.05 28      181      206  32.022  <.0001
2       206 6.05 28      194      219  34.090  <.0001
3       218 6.05 28      206      231  36.136  <.0001
4       225 6.05 28      212      237  37.191  <.0001

Confidence level used: 0.95
```

Figure 5: Least Squares Means and Confidence Intervals

In the output of the function emmeans(), visible in figure 5, we can see the least square means (column "emmean") and the confidence intervals for each diet using a confidence level of 0.95. When we look at the output of the linear regression model in figure 4, we can see, that we get the least square means for diet 1, by looking at the estimate of the intercept. The least square means for the other diets can now be constructed by adding the appropriate estimate to the estimate of the intercept (diet 1 least squares mean). For example for diet 2 this would look like this: $193.6 + 12.5 = 206.1$. The confidence intervals are constructed by the following formula: $\text{CI} = \text{lsmean} \pm t_{\text{critical}} \times \text{SE}$. The value for t_{critical} in the case of a 0.95 confidence level and degrees of freedom of 28, is 2.05. If we put in these numbers to the formula, we get the confidence levels visible in figure 5.

Code:

```
# ANOVA analysis
diet$Diet <- as.factor(diet$Diet) # ensure "Diet" is categorical
anova_result <- aov(PostChol ~ Diet, data = diet)
anova_summary <- summary(anova_result)

lm_anova_output <- lm(PostChol ~ Diet, data = diet)
lm_anova_summary <- summary(lm_anova_output)
lm_anova_result <- anova(lm_output)

anova_post_means <- emmeans(lm_output, ~ Diet)
anova_post_means_summary <- summary(anova_post_means)
```

ANCOVA Analysis (Equal Slopes)

Table of the ANCOVA analysis with equal slopes:

```
            Df Sum Sq Mean Sq F value  Pr(>F)
PreChol      1    130   130.4   0.430 0.51752
Diet         3   4465  1488.4   4.908 0.00753 **
Residuals   27   8188   303.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 6: ANCOVA Analysis with Equal Slopes

To evaluate the results of the ANCOVA analysis with equal slopes, shown in figure 6, we mainly want to focus on the p-values. The p-value for the pre-treatment cholesterol is 0.51752. This p-value is greater than 0.05, which means that we fail to reject the null hypothesis. Therefore, we conclude that pre-treatment cholesterol does not have a statistically significant effect on post-treatment cholesterol.
The p-value for the diet is 0.00753. This p-value is less than 0.05, meaning that diet has a statistically significant effect on post-diet cholesterol levels. Therefore, we can reject the null hypothesis that the means for the different diets are equal.

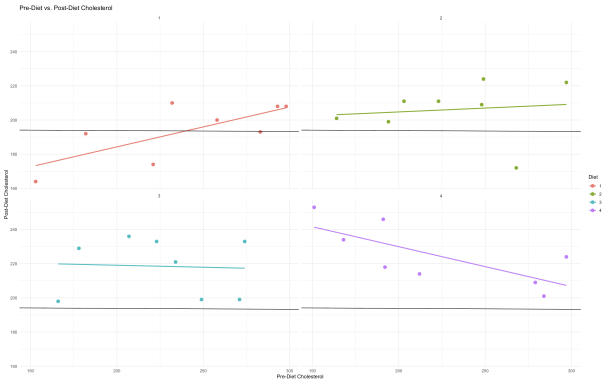

Figure 7: ANCOVA Scatter Plot

The main thing I take away from the scatter plot in figure 7 is that our assumption of equal slopes is probably false. In figure 7 we can see a scatter plot of the pre- and post-diet cholesterol values, the fitted lines for these points (in the specific colour), and the fitted line of the linear regression model (in black). We can see that the black line for diet 1 and diet 4 does not fit at all, leading me to the above mentioned conclusion.

```
Diet emmean   SE df lower.CL upper.CL
1       194 6.20 27      181      206
2       206 6.16 27      193      219
3       218 6.16 27      206      231
4       225 6.18 27      212      238

Confidence level used: 0.95
```

Figure 8: Least Squares Means and Confidence Intervals for ANCOVA (equal slopes)

In figure 8, we see the least squares means for the corrected post-diet cholesterol levels from the ANCOVA analysis. "Corrected" in this case meaning that we account for the fact that different diet groups might have started at different cholesterol levels. The interesting thing here is that there is basically no difference (ignoring some rounding errors) to the ANOVA means. This leads me to believe that the pre-diet cholesterol level has the same effect on all diets. Therefore we can conclude that all diets work equally well/bad for different levels of pre-diet cholesterol.

Code:

```
# ANCOVA analysis, equal slopes
ancova_equal_result <- aov(PostChol ~ PreChol + Diet, data = diet)
ancova_equal_summary <- summary(ancova_equal_result)

lm_ancova_equal_output <- lm(PostChol ~ PreChol + Diet, data = diet)
lm_ancova_equal_result <- anova(lm_ancova_equal_output)

intercept <- coef(lm_ancova_equal_output)[1]
slope <- coef(lm_ancova_equal_output)[2]

diet_scatterplot <- ggplot(diet,
    aes(x = PreChol,
        y = PostChol,
        color = Diet)) +
  geom_point(size = 3) +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE) +
  geom_abline(intercept = intercept, slope = slope, color = "black") +
  facet_wrap(~ Diet) +
  theme_minimal() +
  labs(title = "Pre-Diet vs. Post-Diet Cholesterol",
       x = "Pre-Diet Cholesterol",
       y = "Post-Diet Cholesterol")
print(diet_scatterplot)

ancova_post_means <- emmeans(lm_output_equal,
    ~ Diet,
    at = list(PreChol = mean(diet$PreChol)))
ancova_post_means_summary <- summary(ancova_post_means)
```

ANCOVA Analysis (Unequal Slopes)

Table of ANCOVA analysis with unequal slopes:

```
             Df Sum Sq Mean Sq F value  Pr(>F)
PreChol       1    130   130.4   0.535 0.47170
Diet          3   4465  1488.4   6.103 0.00308 **
PreChol:Diet  3   2335   778.3   3.191 0.04173 *
Residuals    24   5853   243.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 9: ANCOVA Analysis with Unequal Slopes

Like in figure 6 for equal slopes, in figure 9 we can see the outputs of the ANCOVA test but this time with unequal slopes. Especially interesting are again the p-values. Like with the ANCOVA test for equal slopes, we can see a non-significant effect of the pre-diet cholesterol levels on the post-diet cholesterol levels and a significant effect of the diet on the post-cholesterol levels. Additionally we now see a significant interaction between pre-cholesterol levels and the diet. This means that the effect of the pre-cholesterol levels on the post-cholesterol levels varies depending on the diet group. This also suggests that we should be cautious about interpreting the main effects of pre-cholesterol levels and diet in isolation because the effect of pre-cholesterol levels depends on the diet groups (suggesting unequal slopes).

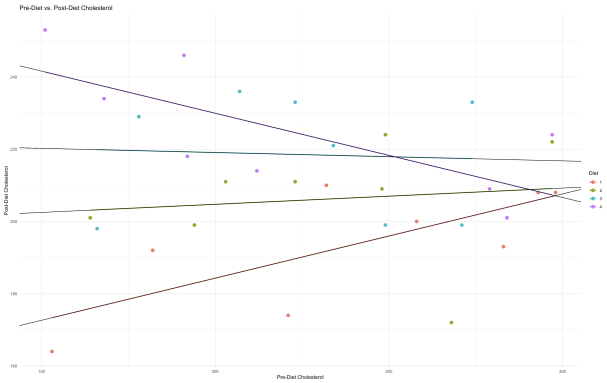

Figure 10: ANCOVA Scatter Plot (unequal slopes)

In figure 10 we can see that the lines from the regression model with unequal slopes (in black) coincide with the "actual" lines we get when we just fit a line from the individual points of the diets. This suggest that the unequal slopes model more accurately represents the actual data.

```
 Diet emmean   SE df lower.CL upper.CL
 1       191 5.64 24      180      203
 2       206 5.52 24      195      217
 3       218 5.56 24      207      230
 4       223 5.58 24      212      235

Confidence level used: 0.95
```

Figure 11: Least Squares Means and Confidence Intervals for ANCOVA (unequal slopes)

In figure 11 we can see the least squares means and the confidence intervals for post-diet cholesterol means in the ANCOVA analysis with unequal slopes. We again see, like with the ANCOVA analysis with equal slopes, that we don't have a meaningful difference in means when compared to the ANOVA analysis. This in turn let us assume that the conclusion we drew from the means in the ANCOVA analysis with equal slopes still holds true, even when accounting for different interactions between pre-diet cholesterol levels and Diet.

When talking about corrected least squares means of post-diet cholesterol levels in this context, I already discussed that we mean corrected in terms of pre-diet cholesterol. The value we adjusted for, is the average pre-diet cholesterol level (the mean value of the column PreChol in our sample).

Code:

```
# ANCOVA analysis, unequal slopes
ancova_unequal_result <- aov(PostChol ~ PreChol * Diet, data = diet)
ancova_unequal_summary <- summary(ancova_unequal_result)

lm_ancova_unequal_output <- lm(PostChol ~ PreChol * Diet, data = diet)
lm_ancova_unequal_result <- anova(lm_ancova_unequal_output)

coefficients <- coef(lm_ancova_unequal_output)
intercept_1 <- coefficients[1]
slope_1 <- coefficients[2]
intercept_2 <- coefficients[1] + coefficients[3]
slope_2 <- coefficients[2] + coefficients[6]
intercept_3 <- coefficients[1] + coefficients[4]
slope_3 <- coefficients[2] + coefficients[7]
intercept_4 <- coefficients[1] + coefficients[5]
slope_4 <- coefficients[2] + coefficients[8]

diet_scatterplot_2 <- ggplot(diet,
    aes(x = PreChol, y = PostChol, color = Diet)) +
  geom_point(size = 3) +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE) +
  geom_abline(intercept = intercept_1,
    slope = slope_1, color = "black") +
  geom_abline(intercept = intercept_2,
    slope = slope_2, color = "black") +
  geom_abline(intercept = intercept_3,
    slope = slope_3, color = "black") +
  geom_abline(intercept = intercept_4,
    slope = slope_4, color = "black") +
  theme_minimal() +
  labs(title = "Pre-Diet vs. Post-Diet Cholesterol",
       x = "Pre-Diet Cholesterol",
       y = "Post-Diet Cholesterol")
print(diet_scatterplot_2)
```

```
ancova_post_means_unequal <- emmeans(lm_ancova_unequal_output,
    ~ Diet,
    at = list(PreChol = mean(diet$PreChol)))
ancova_post_means_summary <- summary(ancova_post_means_unequal)
```