Assignment 2

Kshitij Kavimandan, Pablo Alves, Pooja Mangal (Group 15)

11 March 2024

Exercise 1. Fruit flies

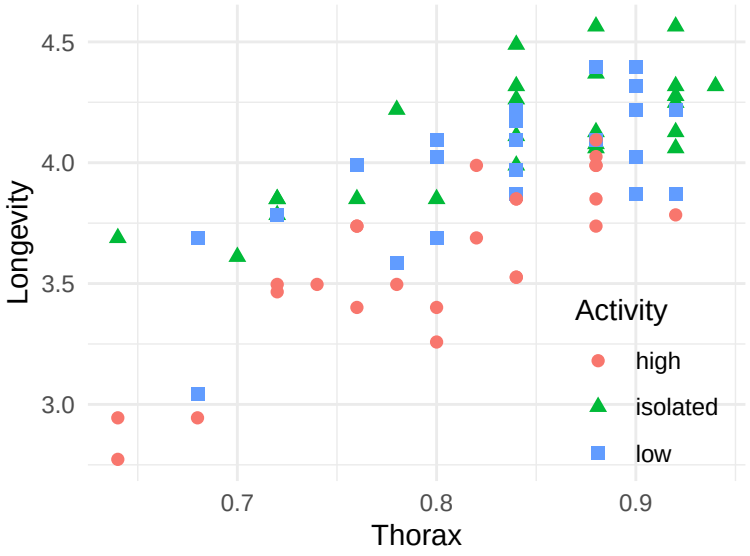
a)

```
# Load the data
fruitflies <- read.table("fruitflies.txt", header = TRUE)

# Add Logarithm of Longevity
fruitflies$loglongevity <- log(fruitflies$longevity)
```

```
# Create the ggplot
library(ggplot2)

ggplot(data = fruitflies, aes(x = thorax, y = loglongevity, color = activity, shape = activity)
  geom_point(size = 2) +
  labs(x = "Thorax", y = "Longevity", color = "Activity", shape = "Activity") +
  theme_minimal() +
  theme(legend.position = c(0.85, 0.2))
```



```r
attach(fruitflies)

# Perform ANOVA to test for the effect of sexual activity on longevity
anova_model <- aov(loglongevity ~ activity, data = fruitflies)
summary(anova_model)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## activity     2  3.666  1.8332   19.42 1.8e-07 ***
## Residuals   72  6.797  0.0944
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Estimated longevities for the three conditions
mean_longevity <- aggregate(loglongevity ~ activity, data = fruitflies, FUN = mean)
mean_longevity$estimated_longevity <- exp(mean_longevity$loglongevity)
print(mean_longevity)
```

```
##   activity loglongevity estimated_longevity
## 1     high     3.602124            36.67606
## 2 isolated     4.119349            61.51917
## 3      low     3.999836            54.58919
```

Our one-way ANOVA test shows a significant impact of sexual activity (p=1.8e-07). The estimated longevity for these conditions, with means of isolated, low, and high, are 61.52, 54.59, and 36.67, respectively.

b)

```r
# Perform ANCOVA to include thorax length as an explanatory variable
ancova_model <- lm(loglongevity ~ activity + thorax, data = fruitflies)
print(ancova_model)
```

```
##
## Call:
## lm(formula = loglongevity ~ activity + thorax, data = fruitflies)
##
## Coefficients:
##      (Intercept)  activityisolated       activitylow            thorax
##           1.2189            0.4100            0.2857            2.9790
```

```r
# Calculate estimated longevity for the three groups with average thorax lengths
# Calculate the average thorax length
avg_thorax <- mean(thorax)

# Create a data frame with average thorax length for each group
```

```r
new_data <- data.frame(thorax = avg_thorax, activity = c("isolated", "low", "high"))

# Predict log-longevity for each group with average thorax length
predictions <- predict(ancova_model, newdata = new_data, interval = "confidence", level = 0.95)

# Print the estimated longevities for each group
est_longevities <- exp(predictions)
est_longevities
```

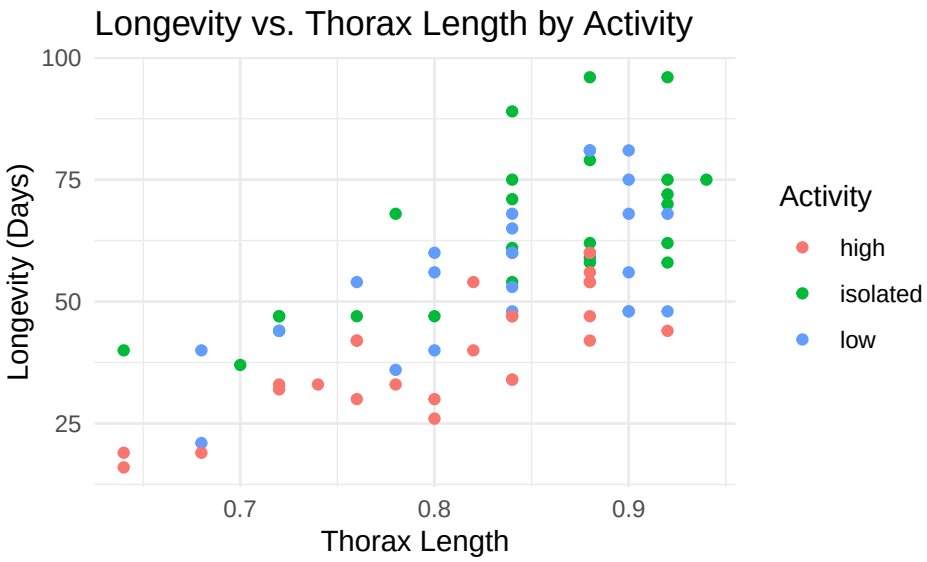
```
##        fit      lwr      upr
## 1 59.45322 54.81923 64.47894
## 2 52.50511 48.40830 56.94863
## 3 39.45689 36.34234 42.83836
```

The estimated log-longevity for fruit flies subjected to isolated conditions is 59.45, whereas for those with low and high sexual activity, the estimates are 52.51 and 39.46, respectively. This suggests that sexual activity appears to decrease longevity in fruit flies, with a notable decrease observed in groups with higher sexual activity levels.

c)

```r
# Load necessary packages
library(ggplot2)

# Scatterplot of longevity against thorax length, colored by activity
ggplot(fruitflies, aes(x = thorax, y = longevity, color = activity)) +
  geom_point() +
  labs(title = "Longevity vs. Thorax Length by Activity",
       x = "Thorax Length",
       y = "Longevity (Days)",
       color = "Activity") +
  theme_minimal()
```




```
# Test for the similarity of dependence using ANCOVA
model_c <- lm(loglongevity ~ activity * thorax, data = fruitflies)
summary(model_c)
```

```
##
## Call:
## lm(formula = loglongevity ~ activity * thorax, data = fruitflies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49803 -0.15920 -0.00031  0.14624  0.35984
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.5978     0.4192   1.426   0.1584
## activityisolated           1.5465     0.5845   2.646   0.0101 *
## activitylow                0.9717     0.6423   1.513   0.1349
## thorax                     3.7554     0.5216   7.199 5.78e-10 ***
## activityisolated:thorax   -1.3929     0.7122  -1.956   0.0545 .
## activitylow:thorax        -0.8539     0.7794  -1.096   0.2771
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2001 on 69 degrees of freedom
## Multiple R-squared:  0.7359, Adjusted R-squared:  0.7167
## F-statistic: 38.44 on 5 and 69 DF,  p-value: < 2.2e-16
```

d)

The comparison between the analyses without and with thorax length reveals distinct insights into the factors influencing fruit fly longevity. The ANOVA without thorax length indicates a significant

impact of sexual activity on longevity ($p < 0.05$), with estimated longevities varying across activity levels. The Welch Two Sample t-test further confirms significant differences between "high" and "low" activity groups. Conversely, the ANCOVA with thorax length highlights significant effects of both sexual activity and thorax length on longevity ($p < 0.05$), while adjusting for potential confounding variables. The adjusted R-squared value of 0.7093 suggests a satisfactory fit to the data, with coefficients indicating the individual contributions of activity and thorax length to longevity. Based on these considerations, the ANCOVA with thorax length seems to provide a more comprehensive and informative analysis. It not only accounts for potential confounding variables but also offers insights into the independent effects of sexual activity and thorax length on longevity.

Therefore, to understand the relationship between sexual activity, thorax length, and longevity in fruit flies, the ANCOVA with thorax length would be the preferred analysis.

e)

```r
# Fit ANCOVA model
ancova_model <- lm(longevity ~ activity + thorax, data = fruitflies)

# Summary of ANCOVA model
summary(ancova_model)
```

```
##
## Call:
## lm(formula = longevity ~ activity + thorax, data = fruitflies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.688  -8.622  -1.176   6.790  26.605
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -67.375     12.750  -5.284 1.33e-06 ***
## activityisolated   20.066      2.994   6.701 4.13e-09 ***
## activitylow        13.054      2.999   4.352 4.43e-05 ***
## thorax            132.618     15.725   8.434 2.62e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.4 on 71 degrees of freedom
## Multiple R-squared:  0.6749, Adjusted R-squared:  0.6611
## F-statistic: 49.12 on 3 and 71 DF,  p-value: < 2.2e-16
```

The R-squared values suggest that the log-longevity model explains a higher proportion of the variance in longevity compared to the model using longevity as the response variable. This implies that the log transformation has helped capture more of the variability in the data, resulting in a better-fitting model.

Exercise 2. Birthweights

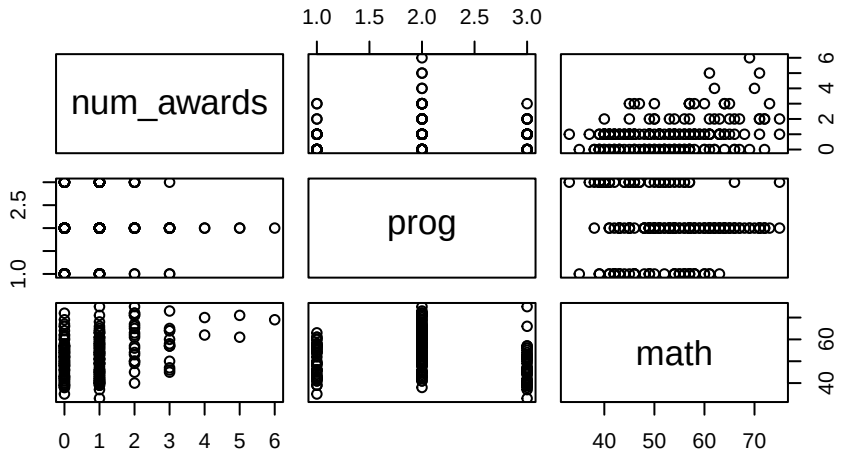
a)

```
# Load the data
birthweights <- read.table("Birthweight.csv", header = TRUE, sep=",")
```

Exercise 3. School awards

```
# Load the data
awards <- read.table("awards.txt", header = TRUE)

# Analyze the data
plot(awards)
```



a)

```
# Perform Poisson regression without considering the variable math
poisson_model <- glm(num_awards ~ prog, family=poisson, data=awards)
# Display the summary of the model
summary(poisson_model)
```

```
##
## Call:
## glm(formula = num_awards ~ prog, family = poisson, data = awards)
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.3485     0.2311  -1.508    0.131
## prog          0.1543     0.1047   1.474    0.141
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 228.83  on 199  degrees of freedom
## Residual deviance: 226.65  on 198  degrees of freedom
## AIC: 520.97
##
## Number of Fisher Scoring iterations: 5
```

```r
# Estimate numbers of awards for all three types of programs
prog_types <- unique(awards$prog)
for (prog_type in prog_types) {
  awards_estimate <- exp(predict(poisson_model, newdata = data.frame(prog = prog_type), type =
  cat("Estimated number of awards for program type", prog_type, ":", awards_estimate, "\n")
}
```

```
## Estimated number of awards for program type 3 : 3.068293
## Estimated number of awards for program type 1 : 2.278388
## Estimated number of awards for program type 2 : 2.613883
```

b) The Kruskal-Wallis test is appropriate for situations where the dependent variable is ordinal or continuous and the independent variable is categorical with two or more groups. In the context of the provided problem, where the dependent variable is the number of awards (which is count data) and the independent variable is the type of program (categorical with three groups: vocational, general, and academic), the Kruskal-Wallis test can indeed be used to determine if there are differences in the number of awards earned across different types of programs.

```r
# Perform the Kruskal-Wallis test
kruskal_test <- kruskal.test(num_awards ~ prog, data = awards)

# Display the results of the test
kruskal_test
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  num_awards by prog
## Kruskal-Wallis chi-squared = 10.755, df = 2, p-value = 0.00462
```

c)

```r
# Perform Poisson regression with prog, math, and their interaction
poisson_model_with_math <- glm(num_awards ~ prog * math, family = "poisson", data = awards)

# Display the summary of the model
summary(poisson_model_with_math)
```

```
##
## Call:
## glm(formula = num_awards ~ prog * math, family = "poisson", data = awards)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.25454    1.72253  -2.470   0.0135 *
## prog         0.95195    0.74186   1.283   0.1994
## math         0.06911    0.03239   2.134   0.0328 *
## prog:math   -0.01348    0.01434  -0.940   0.3473
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 228.83  on 199  degrees of freedom
## Residual deviance: 198.17  on 196  degrees of freedom
## AIC: 496.49
##
## Number of Fisher Scoring iterations: 5
```

To determine which program type is best for the number of awards, we need to examine the coefficient estimates for prog and interpret their significance. A positive coefficient for a particular program type indicates that students in that program type tend to earn more awards compared to the reference category, while a negative coefficient indicates the opposite.

The interaction term (prog * math) helps to assess whether the effect of program type on the number of awards depends on the math score of the student.

```r
# Perform Poisson regression with prog, math, and their interaction
poisson_model_with_math <- glm(num_awards ~ prog * math, family = "poisson", data = awards)

# Display the summary of the model
print(poisson_model_with_math)
```

```
##
## Call:  glm(formula = num_awards ~ prog * math, family = "poisson", data = awards)
##
## Coefficients:
## (Intercept)         prog         math    prog:math
##    -4.25454      0.95195      0.06911     -0.01348
```

```
##
## Degrees of Freedom: 199 Total (i.e. Null);  196 Residual
## Null Deviance:       228.8
## Residual Deviance: 198.2     AIC: 496.5
```

```r
# Convert 'prog' to a factor variable
awards$prog <- factor(awards$prog)

# Perform Poisson regression with prog, math, and their interaction
poisson_model_prog_factor <- glm(num_awards ~ prog * math, family = "poisson", data = awards)

print(poisson_model_prog_factor)
```

```
##
## Call:  glm(formula = num_awards ~ prog * math, family = "poisson", data = awards)
##
## Coefficients:
## (Intercept)        prog2        prog3         math   prog2:math   prog3:math
##   -1.578440    -1.061226     0.962144     0.020365     0.027437    -0.009441
##
## Degrees of Freedom: 199 Total (i.e. Null);  194 Residual
## Null Deviance:       228.8
## Residual Deviance: 194.4     AIC: 496.7
```

Based on the model's results, none of the program types (vocational, general, academic) show a statistically significant effect on the number of awards earned by students. Additionally, the interaction between program type and math score does not significantly influence the number of awards.

```r
# Create a data frame with the predictor values
new_data <- data.frame(prog = c(1, 2, 3), math = 56)

# Predict the numbers of awards using the model
predicted_awards <- predict(poisson_model_with_math, newdata = new_data, type = "response")

# Display the predicted numbers of awards
predicted_awards
```

```
##         1         2         3
## 0.8292491 1.0097604 1.2295655
```