Data Analysis Report

Trees

a) Investigate whether the tree type influences volume by performing ANOVA, without taking diameter and height into account. Can a t-test be related to the above ANOVA test? Estimate the volumes for the two tree types.

```
# Load the data and calculate the mean volume for each tree type

data <- read.table("resources/treeVolume.txt", header=TRUE)
mean_vol_beech <- mean(data[data$type == "beech", "volume"])
mean_vol_oak <- mean(data[data$type == "oak", "volume"])
```

We use the aov() function to perform ANOVA:

```
aov_result <- aov(volume ~ type, data=data)
summary(aov_result)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## type         1    380   379.5   1.898  0.174
## Residuals   57  11395   199.9
```

The output of the summary function shows the results of the ANOVA test. When the p-value is below the given confidence level (0.05), we can conclude that there is a significant difference in volumes between the two tree types.

We can see that the p-value is 0.174. Therefore we can conclude that there is not a significant difference in volumes between the two tree types.

It is possible to relate the ANOVA test to a t-test by performing a two-sample t-test between the volumes of the two tree types. However, the ANOVA test is more appropriate when we have more than two groups to compare.

It is possible to relate an ANOVA test to a t-testby performing a two-sample t-test between the volumes of the two tree types. However, the ANOVA test is more appropriate when we have more than two groups to compare.

To perform a t-test, we can use the t.test() function:

```
t_test_result <- t.test(data[data$type == "beech", "volume"], data[data$type == "oak", "volume"])
t_test_result
```

```
##
##  Welch Two Sample t-test
##
## data:  data[data$type == "beech", "volume"] and data[data$type == "oak", "volume"]
## t = -1.4051, df = 52.804, p-value = 0.1659
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -12.32992   2.17186
```

```
## sample estimates:
## mean of x mean of y
##  30.17097  35.25000
```

The p-value from the t-test is 0.1659. Relatively similar result to ANOVA performed earlier. The t-test shows us if there is a significant difference in mean volume between the two tree types. With the result we can again say that there is not a significant difference in volumes between the two tree types.

To calculate the mean value of each group we estimate the volumes for each of the tree types.

```
# calculate mean volumes for each tree type
mean_volumes <- tapply(data$volume, data$type, mean)
mean_volumes
```

```
##    beech      oak
## 30.17097 35.25000
```

b) Now include diameter and height as explanatory variables into the analysis. Investigate whether the influence of diameter on volume is similar for the both tree types. Do the same for the influence of height on volume. (Consider at most one (relevant) pairwise interaction per model.) Comment.

To investigate the influence of diameter and height on volume, we can perform a multiple linear regression analysis in R. We can start by creating a model that includes both diameter and height as explanatory variables:

```
# create multiple linear regression model
reg_model <- lm(volume ~ type + diameter + height, data = data)
summary(reg_model)
```

```
##
## Call:
## lm(formula = volume ~ type + diameter + height, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1859 -2.1396 -0.0871  1.7208  7.7010
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -63.78138    5.51293 -11.569 2.33e-16 ***
## typeoak      -1.30460    0.87791  -1.486    0.143
## diameter      4.69806    0.16450  28.559  < 2e-16 ***
## height        0.41725    0.07515   5.552 8.42e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.243 on 55 degrees of freedom
## Multiple R-squared:  0.9509, Adjusted R-squared:  0.9482
## F-statistic: 354.9 on 3 and 55 DF,  p-value: < 2.2e-16
```

This code will create a multiple linear regression model using the lm() function in R and display the results using the summary() function. The output will show the coefficients for each variable and their significance.

To investigate whether the influence of diameter on volume is similar for the both tree types, we can include a diameter-tree type interaction term in the model:

```
# create multiple linear regression model with interaction term
reg_model_interaction <- lm(volume ~ type * diameter + height, data = data)
summary(reg_model_interaction)
```

```
##
## Call:
## lm(formula = volume ~ type * diameter + height, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3500 -2.1940 -0.1413  1.7012  8.1765
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -63.87254    5.53859 -11.532 3.47e-16 ***
## typeoak           -4.96300    5.14936  -0.964    0.339
## diameter           4.60813    0.20701  22.261  < 2e-16 ***
## height             0.43412    0.07903   5.493 1.09e-06 ***
## typeoak:diameter   0.25886    0.35897   0.721    0.474
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.257 on 54 degrees of freedom
## Multiple R-squared:  0.9513, Adjusted R-squared:  0.9477
## F-statistic:   264 on 4 and 54 DF,  p-value: < 2.2e-16
```

This code will create a new regression model with an interaction term between tree type and diameter, and display the results using the summary() function. We can see that the interaction term is not significant ($0.474 > 0.05$), therefore, the relationship between diameter and volume does not depend on the tree type.

We can perform a similar analysis to investigate the influence of height on volume:

```
# create multiple linear regression model with interaction term
reg_model_interaction <- lm(volume ~ type + diameter + height * type, data = data)
summary(reg_model_interaction)
```

```
##
## Call:
## lm(formula = volume ~ type + diameter + height * type, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2298 -2.1127 -0.1612  1.8006  8.1648
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -57.5514     7.1114  -8.093 6.99e-11 ***
## typeoak        -17.4710    11.8262  -1.477  0.14540
## diameter         4.7787     0.1735  27.547  < 2e-16 ***
## height           0.3212     0.1023   3.140  0.00274 **
## typeoak:height   0.2117     0.1545   1.371  0.17613
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.217 on 54 degrees of freedom
## Multiple R-squared:  0.9525, Adjusted R-squared:  0.949
## F-statistic: 270.9 on 4 and 54 DF,  p-value: < 2.2e-16
```

This code will create a new regression model with an interaction term between tree type and height, and display the results using the summary() function. We can see that the interaction term is not significant

(0.17613 > 0.05), therefore, the relationship between height and volume does not depend on the tree type.

In conclusion, by including diameter and height as explanatory variables, we can see that both variables do not have a significant influence on volume, and that the relationships between these variables and volume do not depend on the tree type.

c) Using the results from c), investigate how diameter, height and type influence volume. Comment. Using the resulting model, predict the volume for a tree with the (overall) average diameter and height?

Using the results from b), we can fit a linear regression model for volume using diameter, height, and type as explanatory variables:

```
lm_all <- lm(volume ~ diameter + height + type, data=data)
summary(lm_all)
```

```
##
## Call:
## lm(formula = volume ~ diameter + height + type, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1859 -2.1396 -0.0871  1.7208  7.7010
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -63.78138    5.51293 -11.569 2.33e-16 ***
## diameter      4.69806    0.16450  28.559  < 2e-16 ***
## height        0.41725    0.07515   5.552 8.42e-07 ***
## typeoak      -1.30460    0.87791  -1.486    0.143
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.243 on 55 degrees of freedom
## Multiple R-squared:  0.9509, Adjusted R-squared:  0.9482
## F-statistic: 354.9 on 3 and 55 DF,  p-value: < 2.2e-16
```

The results show that diameter and height have a significant influence on volume. Type does not have a significant influence on volume.

To predict the volume for a tree with the overall average diameter and height, we can calculate the mean values for diameter and height and use them in the model:

```
mean_diam <- mean(data$diameter)
mean_height <- mean(data$height)

predicted_vol <- predict(lm_all, newdata=data.frame(diameter=mean_diam, height=mean_height, type="beech
predicted_vol
```

```
##        1
## 33.20049
```

d) Propose a transformation of the explanatory variables that possibly yields a better model (verify this). (Hint: think of a natural link between the response and explanatory variables.)

One possible transformation of the explanatory variables is to take the natural logarithm of the diameter and height. This transformation can be useful when the relationship between the explanatory variables and the response is not linear but instead follows a logarithmic pattern. We can apply this transformation by creating new variables in the data frame:

```r
data$log_diameter <- log(data$diameter)
data$log_height <- log(data$height)
```

Then, we can build a new model using these transformed variables:

```r
model2 <- lm(volume ~ log_diameter + log_height + type, data = data)
summary(model2)
```

```
##
## Call:
## lm(formula = volume ~ log_diameter + log_height + type, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0171 -2.3492 -0.7925  1.4275 13.2661
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -245.841     33.151  -7.416 7.86e-10 ***
## log_diameter   64.035      3.280  19.522  < 2e-16 ***
## log_height     25.941      8.070   3.214  0.00219 **
## typeoak        -2.258      1.260  -1.792  0.07864 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.596 on 55 degrees of freedom
## Multiple R-squared:  0.9013, Adjusted R-squared:  0.8959
## F-statistic: 167.4 on 3 and 55 DF,  p-value: < 2.2e-16
```

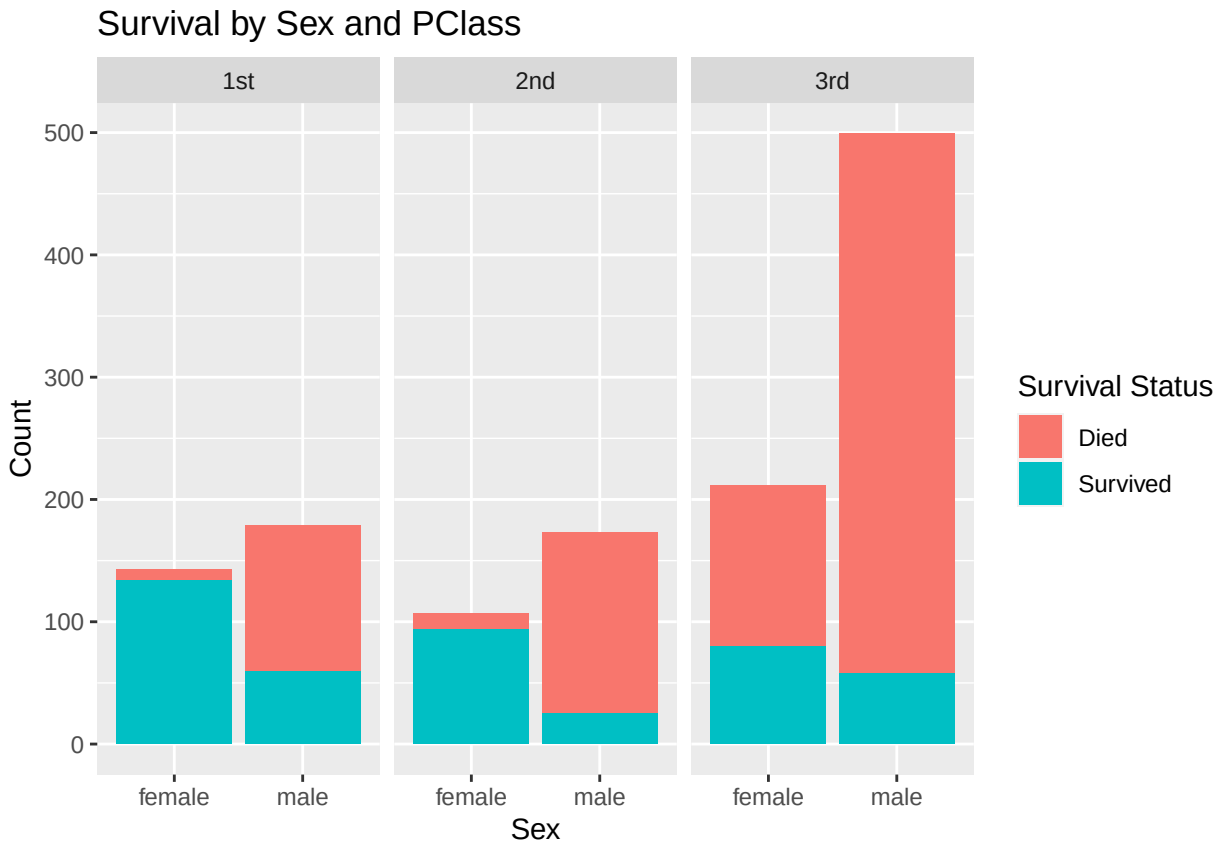
We can compare the R-squared values of the two models to see which one is a better fit. If the R-squared value for model2 is higher, then the logarithmic transformation improved the model fit. If not, then the original model may be the better choice.

Titanic

a) Overview + liniar model for Survival

Survival based on Sex and PClass

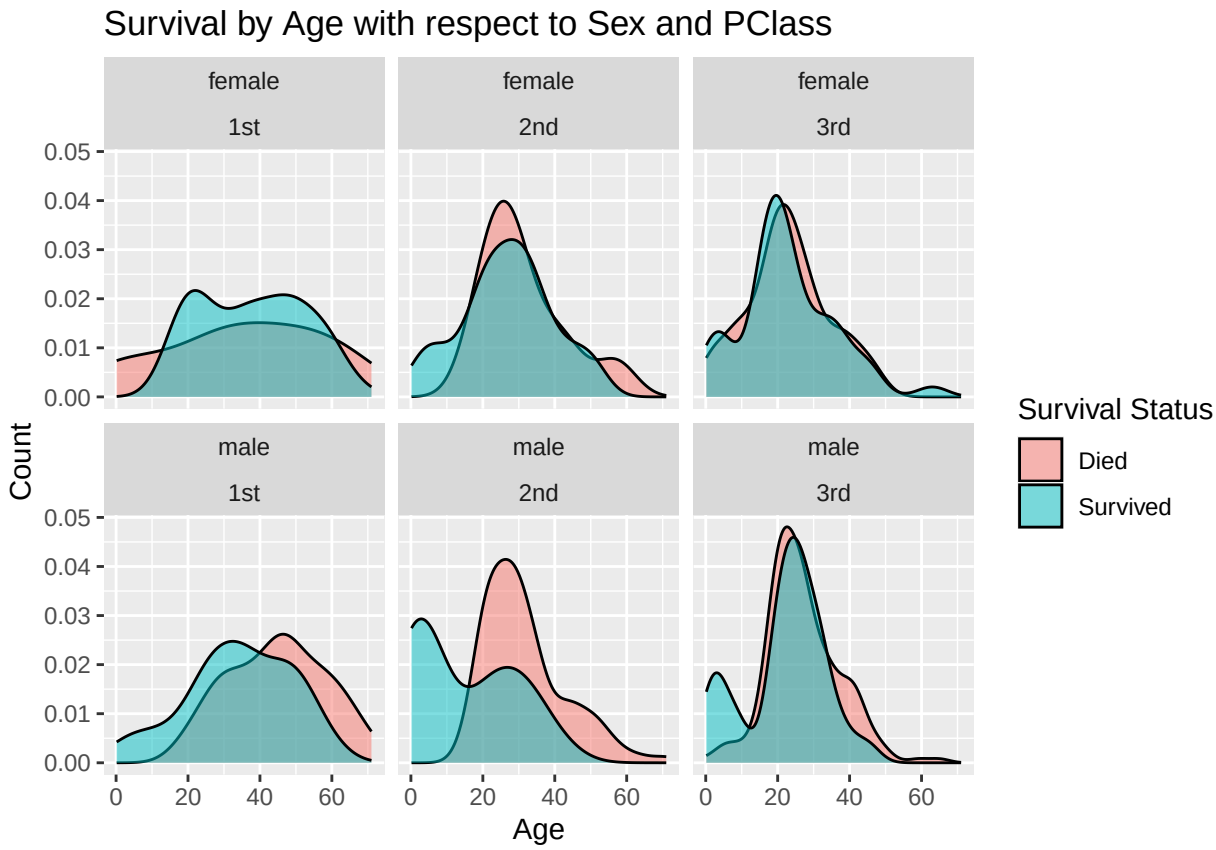
```r
titanic %>%
  mutate(Survived = factor(Survived, labels = c("Died", "Survived"))) %>%
  ggplot(aes(x = Sex, fill = Survived)) +
  geom_bar() +
  facet_wrap(~PClass) +
  labs(
    title = "Survival by Sex and PClass",
    x = "Sex",
    y = "Count",
    fill = "Survival Status"
  )
```



Survival based on Age with respect to Sex and PClass

```
titanic %>%
  mutate(Survived = factor(Survived, labels = c("Died", "Survived"))) %>%
  ggplot(aes(x = Age, fill = Survived)) +
  geom_density(alpha = 0.5) +
  facet_wrap(~Sex + PClass) +
  labs(x = "Age",
       y = "Count",
       title = "Survival by Age with respect to Sex and PClass",
       fill = "Survival Status")
```

```
## Warning: Removed 557 rows containing non-finite values (`stat_density()`).
```



We will fit a liniar model, having `Survived` as the effect and, independently, `PClass`, `Age` and `Sex` as possible causes.

```
model <- glm(Survived ~ PClass + Age + Sex, data = titanic, family = binomial)
model_summary <- summary(model)$coefficients
model_summary
```

```
##                Estimate  Std. Error    z value     Pr(>|z|)
## (Intercept)  3.75966210 0.397567324   9.456668 3.179129e-21
## PClass2nd   -1.29196240 0.260075781  -4.967638 6.777324e-07
## PClass3rd   -2.52141915 0.276656805  -9.113888 7.948131e-20
## Age         -0.03917681 0.007616218  -5.143868 2.691392e-07
## Sexmale     -2.63135683 0.201505379 -13.058494 5.684093e-39
```

```
age_effect_estimate <- summary(model)$coefficients["Age", "Estimate"]
age_p_value <- summary(model)$coefficients["Age", "Pr(>|z|)"]

pclass2nd_p_value <- summary(model)$coefficients["PClass2nd", "Pr(>|z|)"]
pclass3nd_p_value <- summary(model)$coefficients["PClass3rd", "Pr(>|z|)"]
sexmale_p_value <- summary(model)$coefficients["Sexmale", "Pr(>|z|)"]
```

We observe a low p_value (2.6913921×10^{-7}) for Age, therefore were reject the initial hypothesis and conclude that age has an effect on the Survival. Apparently, the chances of survival change with (-0.0391768) with each year.

We observe a low p_values for PClass2nd (6.7773237×10^{-7}), PClass3rd ($7.9481311 \times 10^{-20}$), Sexmale ($5.6840932 \times 10^{-39}$), therefore were reject the initial hypothesis and conclude that being 2nd class, 3rd class or being a male has an effect on the Survival.

b) Interactions + Studying a 55 years old person chances to survive

Interactions

```
model_age_pclass <- glm(Survived ~ Age:PClass, data = titanic, family = "binomial")
age_first_class = summary(model_age_pclass)$coefficients["Age:PClass1st", "Pr(>|z|)"]
age_second_class = summary(model_age_pclass)$coefficients["Age:PClass2nd", "Pr(>|z|)"]
age_third_class = summary(model_age_pclass)$coefficients["Age:PClass3rd", "Pr(>|z|)"]
```

Age * PClass interaction Big P values (> 0.05) for Age:PClass1st (0.0724887) and Age:PClass3rd (5.972008×10^{-14}) suggest there is no strong evidence of an interaction between age and those classes.

Low P value ($« 0.05$) for Age:PClass2nd (5.8083981×10^{-6}) suggest therestrong evidence of an interaction between age and 2nd class of passengers.

```
model_age_sex <- glm(Survived ~ Age:Sex, data = titanic, family = "binomial")
age_female = summary(model_age_sex)$coefficients["Age:Sexfemale", "Pr(>|z|)"]
age_male = summary(model_age_sex)$coefficients["Age:Sexmale", "Pr(>|z|)"]
```

Age * Sex interaction Low P values ($« 0.05$) for Age:Sexfemale (7.2489359×10^{-7}) and Age:Sexmale (age_male) suggest there is strong evidence of an interaction between age and the sex of the passenger

Hypothetical of a 55 years old person

We build a linear model to predict `Survival` with respect to `Age`, `PClass`, `Sex` and their interactions.

```
model <- glm(Survived ~ Age*PClass*Sex, data = titanic, family = "binomial")
male_first <- predict(model, data.frame(Age = 55, PClass = "1st", Sex = "male"), type = "response")
male_second <- predict(model, data.frame(Age = 55, PClass = "2nd", Sex = "male"), type = "response")
male_third <- predict(model, data.frame(Age = 55, PClass = "3rd", Sex = "male"), type = "response")

female_first <- predict(model, data.frame(Age = 55, PClass = "1st", Sex = "female"), type = "response")
female_second <- predict(model, data.frame(Age = 55, PClass = "2nd", Sex = "female"), type = "response")
female_third <- predict(model, data.frame(Age = 55, PClass = "3rd", Sex = "female"), type = "response")
```

- Male
  - 55 years old `male` of `1st` class has a change of 18.5658098% to survive.
  - 55 years old `male` of `2nd` class has a change of 0.2056975% to survive.
  - 55 years old `male` of `3rd` class has a change of 3.5889482% to survive.
- Female
  - 55 years old `female` of `1st` class has a change of 96.0046036% to survive.
  - 55 years old `female` of `2nd` class has a change of 77.6731219% to survive.
  - 55 years old `female` of `3rd` class has a change of 44.5486448% to survive.

Based on previous predictions we argue that a 1st class 55 years old female 96.0046036% was very likely to survive, while a 2nd class 55 years old male (96.0046036%) would have likely died.

c) Survival status predictor + quality measures

I would use a random forest classifier as a predictor.

First I would fill in the missing data with following heuristic:

- Average age for the missing values on `Age` column.
- 50% male/female for missing values on `Sex` column.

We could use cross-validation to split the dataset into train/test datasets.

Python implementation is pretty straightforward with the usage of sklearn.ensemble.RandomForestClassifier class. As quality measures we could use:

- precision [`TP / (TP + FP)`]
- recall [`TP/ (TP + FN)`]
- accuracy [`(TP + TN) / ALL`]

Due to the stochastic nature of the heuristic chosen for filling in `Sex` column and cross-validation dataset splits, I argue the following:

```
* Multiple experiments should be  with respect to cross-validation heuristic.
* The deviation of the `precision`, `recall` and `accuracy` values should be analysed.
```

d) Contingency table test

We will perform two `chi-squared` tests of independence:

Association between `Survival` and `PClass`

```
class_survival_p_value = chisq.test(table(titanic$Survived, titanic$PClass))$p.value
```

We observe a small ($\ll 0.05$) P value ($3.8523155 \times 10^{-38}$) for our test, therefore, we reject the null hypothesis and conclude that there is evidence of an association between the `Survival` and `PClass`.

Association between `Survival` and `Sex`

```
sex_survival_p_value = chisq.test(table(titanic$Survived, titanic$Sex))$p.value
```

We observe a small ($\ll 0.05$) P value ($1.0404031 \times 10^{-73}$) for our test, therefore, we reject the null hypothesis and conclude that there is evidence of an association between the `Survival` and `Sex`.

e) `random forest classifier` vs `chi-squared` independence test.

`chi-squared` test provides a statistical significance test to check association between `Survival` and (`PClass` | `Sex`). It was useful and fast to get insights over the dataset and to statistically derive the association between (`PClass` | `Sex`) and `Survival`.

Arguably, `random forest classifier` offers a predictive model which can be used for extensive analysis.

Both methods serve their purposes in a complementary way.

Military Coups

a) Perform Poisson regression on the full data set, taking miltcoup as response variable. Comment on your findings.

To perform Poisson regression in R, we first need to load the coups.txt file into a data frame. Then, we can use the glm() function to fit a Poisson regression model with miltcoup as the response variable and all other variables as predictors.

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

# Fit the Poisson regression model
model <- glm(miltcoup ~ oligarchy + pollib + parties + pctvote + popn + size + numelec + numregim, data=
```

```r
# or model <- glm(miltcoup ~ ., data = coups, family = "poisson")

# Print the model summary
summary(model)
```

```
##
## Call:
## glm(formula = miltcoup ~ oligarchy + pollib + parties + pctvote +
##     popn + size + numelec + numregim, family = "poisson", data = coups)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3443  -0.9542  -0.2587   0.3905   1.6953
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.5102693  0.9053301  -0.564  0.57301
## oligarchy    0.0730814  0.0345958   2.112  0.03465 *
## pollib      -0.7129779  0.2725635  -2.616  0.00890 **
## parties      0.0307739  0.0111873   2.751  0.00595 **
## pctvote      0.0138722  0.0097526   1.422  0.15491
## popn         0.0093429  0.0065950   1.417  0.15658
## size        -0.0001900  0.0002485  -0.765  0.44447
## numelec     -0.0160783  0.0654842  -0.246  0.80605
## numregim     0.1917349  0.2292890   0.836  0.40303
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 65.945  on 35  degrees of freedom
## Residual deviance: 28.668  on 27  degrees of freedom
## AIC: 111.48
##
## Number of Fisher Scoring iterations: 6
```

The family argument specifies the type of model to fit. In this case, we want a Poisson regression model, so we set family = "poisson".

The output of the summary() function will show us the estimated coefficients, standard errors, z-values, and p-values for each predictor variable in the model. We can use these values to interpret the effect of each predictor on the number of successful military coups.

Based on the p-values, we see that the variables oligarchy, pollib, and parties are significant at the 0.05 level. This means that these variables are likely to be important predictors of the number of military coups.

The coefficient for the variable "oligarchy" is positive and statistically significant (i.e., the p-value is less than 0.05). Therefore, we can conclude that countries with more years ruled by a military oligarchy are more likely to have experienced successful military coups. Moreover, the coefficient for the variable "pollib" is negatively and statistically significant. We can therefore conclude that countries with more political liberalization (i.e., full civil rights) are less likely to have experienced successful military coups.

The coefficient for parties is negative and statistically significant, which means that as the number of legal political parties increases, the number of military coups tends to decrease.
