PS5

2024-09-18

Stat 5000 Homework #4
Fall 2024 due Fri, October 11th @ 11:59 pm Name:
Collaborators: **The Hatman**

Progress Report

- Q1: DONE
- Q2: QUESTION (g)
- Q3: QUESTION (d)

Q1

Exercise 19 (page 143) at the end of Chapter 5 in The Statistical Sleuth describes an observational study of competition among species of birds and rodents for nesting sites in cavities of rocks and trees (data from Donald Youkey, Oregon State University Department of Fisheries and Wildlife). The areas of entrances to nesting cavities were measured for 294 nesting sites for nine common species of birds and rodents in Oregon. The boxplots from Display 5.22 on page 144 are shown below.

The boxplots suggest that variation in areas of cavity entrances is greater for species for which the average areas of entrances are larger. There is also some indication that the distributions of entrance areas are right skewed for most species. This suggests the need for a transformation of the data to promote symmetry and reduce the relationship between the means and variances of entrance areas for the nine species. On a logarithmic scale, the variances are more nearly the same for the nine species and scenes of the distributions of entrance sizes is greatly reduced. Consequently, the following analysis is done with the natural logarithms of the cavity entrance areas. Sample sizes, sample means and sample standard deviations are shown below with samples means and sample standard deviations computed from the natural logarithms of the observed areas of cavity entrances.

There is no file with the actual data. You will need to answer the following questions using only the values of the summary statistics show in the table.

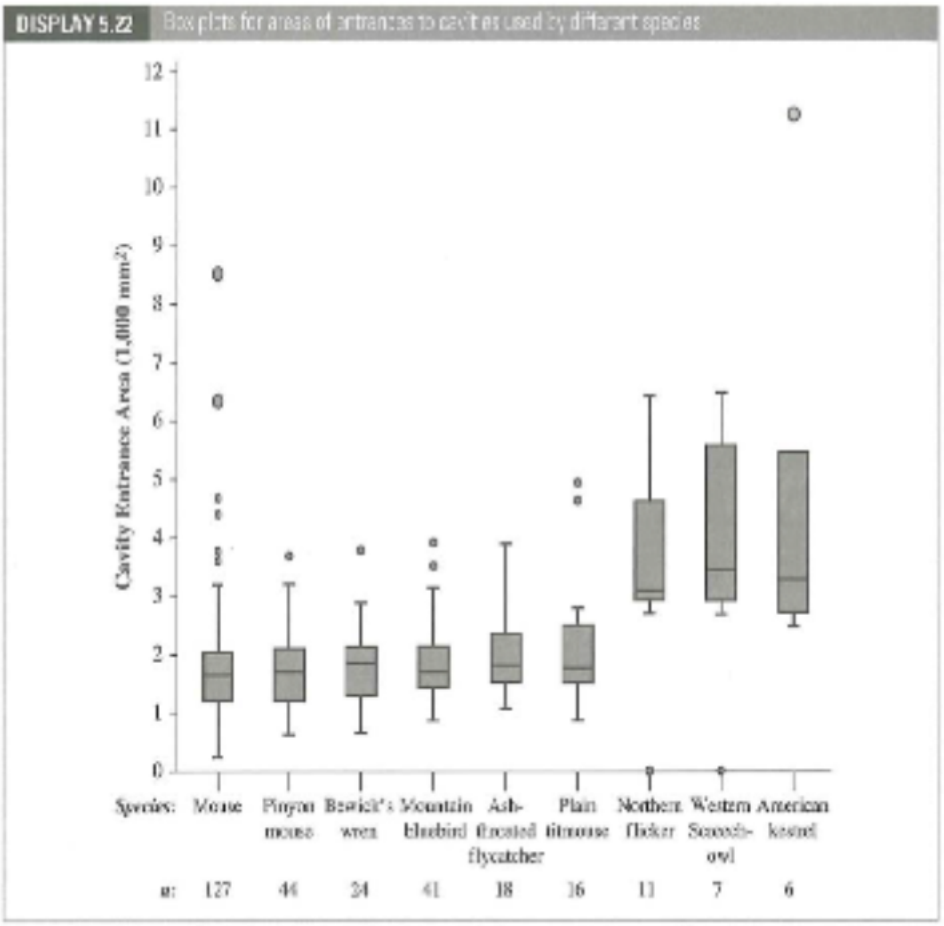


Figure 1: img 1

Summary Statistics for Natural Logarithm of Areas of Nesting Cavity Entrances

Species	*Sample Size (n)*	*Sample Mean Log(1000 mm²)*	*Sample Std. Dev. Log(1000 mm2)*
Mouse	127	7.347	0.4979
Pinyon mouse	44	7.369	0.4235
Bewick's wren	24	7.428	0.3955
Mountain bluebird	41	7.487	0.3181
Ash-throated flycatcher	18	7.563	0.3111
Plain titmouse	16	7.568	0.4649
Northern flicker	11	8.214	0.2963
Western Screech-owl	7	8.272	0.3242
American kestrel	6	8.297	0.5842

Figure 2: img 2

(a)

Compute the pooled estimate of variance for the log-transformed data.

For $N = 294$, $R = 9$, the formula is:

$$S_p^2 = \frac{\sum_{i=1}^{9}(n_i - 1)S_i^2}{294 - 9}$$

```
denom <- 294 - 9
num1 <- (127-1) * (0.4979^2)
num2 <- (44-1) * (0.4235^2)
num3 <- (24-1) * (0.3955^2)
num4 <- (41-1) * (0.3181^2)
num5 <- (18-1) * (0.3111^2)
num6 <- (16-1) * (0.4649^2)
num7 <- (11-1) * (0.2963^2)
num8 <- (7-1) * (0.3242^2)
num9 <- (6-1) * (0.5842^2)
numeratorTotal <- num1 + num2 + num3 + num4 + num5 + num6 + num7 + num8 + num9
pooledEstimate <- numeratorTotal / denom
pooledEstimate
```

```
## [1] 0.1919143
```

$$S_p^2 = 0.1919$$

(b)

Construct an analysis of variance (ANOVA) table for the log transformed data.

N = Total Observations $= 127 + 44 + 24 + 41 + 18 + 16 + 11 + 7 + 6 = 294$

$df_{total} = N - 1 = 294 - 1 = 293$

$df_{model} = k - 1 = 9 - 1 = 8$

$df_{error} = df_{total} - df_{model} = 293 - 8 = 285$

$MS_{model} = \sigma^2 + \frac{1}{9-1}\sum_{i=1}^{9} n_i(\mu_i - \bar{\mu})^2$

Where

$\bar{\mu} = \frac{1}{294}\sum_i n_i\mu_i$

```
nmu1 <- 127 * 7.347
nmu2 <- 44 * 7.369
nmu3 <- 24 * 7.428
nmu4 <- 41 * 7.487
nmu5 <- 18 * 7.563
nmu6 <- 16 * 7.568
nmu7 <- 11 * 8.214
nmu8 <- 7 * 8.272
```

```r
nmu9 <- 6 * 8.297

numerator <- sum(nmu1, nmu2, nmu3, nmu4, nmu5, nmu6, nmu7, nmu8, nmu9)
denom <- 294

barMu <- numerator / denom
barMu
```

```
## [1] 7.475531
```

```r
model1 <- 127 * ((7.347 - barMu)^2)
model2 <- 44 * ((7.369 - barMu)^2)
model3 <- 24 * ((7.428 - barMu)^2)
model4 <- 41 * ((7.487 - barMu)^2)
model5 <- 18 * ((7.563 - barMu)^2)
model6 <- 16 * ((7.568 - barMu)^2)
model7 <- 11 * ((8.214 - barMu)^2)
model8 <- 7 * ((8.272 - barMu)^2)
model9 <- 6 * ((8.297 - barMu)^2)
numeratorModel <- sum(model1, model2, model3, model4, model5, model6, model7, model8, model9)
denomModel <- 8

sumModel <- numeratorModel / denomModel

msModel <- pooledEstimate + sumModel
msModel
```

```
## [1] 2.369372
```

$SS_{model} = df_{model} * MS_{model} = 8 * 2.369372 = 18.95498$ $SS_{error} = df_{error} * MS_{error} = 285 * 0.1919143 = 54.69558$ $SS_{total} = SS_{model} + SS_{error} = 54.69558 + 18.95498 = 73.65056$ $MS_{total} = SS_{total}/df_{total} = 73.65056/293 = 0.2513671$

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Model	8	18.955	2.369
Error	285	54.696	0.192
Total	293	73.651	0.251

(c)

Perform an F-test of the null hypothesis that means of the natural logarithm of cavity entrance areas are the same for all nine species. Report the value of the F-statistics, its degrees of freedom, and the corresponding p-value.

```r
MS_model <- 2.369
MS_error <- 0.192

F_statistic <- MS_model / MS_error

df_model <- 8
```

```r
df_error <- 285

p_value <- 1 - pf(F_statistic, df_model, df_error)
p_value
```

```
## [1] 3.552714e-15
```

$F - Statistic = MS_{model}/MS_{error} = 2.369/0.192 = 12.33854 \approx 12.339$ $F - Statistic_{df_{model},df_{error}} = F_{8,285} \approx 12.339$ p-value: $= P(F_{r-1,N-r} > F) = 3.552714e-15 \approx 3.55e-15$ (very small!)

(d)

Interpret the test result in the context of the study.

Given the relatively large F-statistic and its associated (very small!) p-value, we have evidence to reject the null hypothesis that the average size of log-transformed entrances to nesting cavities are the same across all groups. The p-value is significant at the $\alpha = 0.01$ level. This provides evidence in support of the alternative hypothesis that at least one of the the average size of log-transformed entrances to nesting cavities is different from the the average size of log-transformed entrances to nesting cavities of all other groups (species of birds and rodents).

Q2

Researchers were interested in the effect of a dietary supplement on weight gain in hogs. A total of 12 hogs were used for the experiment. Researchers randomized the 12 hogs so that 3 hogs were assigned to each of 4 treatment groups. Each treatment involved adding a certain amount of the dietary supplement to the daily feed given to the hogs. The 4 amounts considered were 0g, 20g, 40g and 60g of the supplement per kg of feed. The amount of weight gained by each hog was measured in kilograms after 6 weeks. The observed mean weight gains in pounds are provided in the following table.

(a)

Fill in the missing entries in the following ANOVA table.

Amount of Supplement	0	20	40	60
Number of Hogs	3	3	3	3
Mean Weight Gain	22	25	29	32

N = Total Observations $= 4 * 3 = 12$

$df_{total} = N - 1 = 12 - 1 = 11$

$df_{model} = k - 1 = 4 - 1 = 3$

$df_{error} = df_{total} - df_{model} = 11 - 3 = 8$

$MS_{model} = SS_{model}/df_{model} = 174/3 = 58$

$SS_{error} = df_{error} * MS_{error} = 8 * 28$

$SS_{total} = SS_{model} + SS_{error} = 174 + 224 = 398$

$MS_{total} = SS_{total}/df_{total} = 398/11 = 36.18182$

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Model	3	174	58
Error	8	224	28
Total	11	398	36.18

(b)

For the cell means model $Y_{ij} = \mu_i + \epsilon_{ij}$, where $i = 1, 2, 3, 4$ and $j = 1, 2, 3$, what is the parameter vector β, the corresponding design matrix X, and the estimated $\hat{\beta}$ vector?

$$\beta = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\hat{\beta} = \begin{bmatrix} 22 \\ 25 \\ 29 \\ 32 \end{bmatrix}$$

(c)

For the effects model $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, where $i = 1, 2, 3, 4$ and $j = 1, 2, 3$, using the baseline constraint $\alpha_4 = 0$, what is the parameter vector β, the corresponding design matrix X, and the estimated $\hat{\beta}$ vector?

$$\beta = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

For reference, given our prior $\hat{\beta}$,

$$\hat{\beta} = \begin{bmatrix} 22 \\ 25 \\ 29 \\ 32 \end{bmatrix}$$

$\hat{\mu} = (22 + 25 + 29 + 32)/4 = 27$

Our new $\hat{\beta}$ for this problem is:

$$\hat{\beta} = \begin{bmatrix} 27 \\ 22 - 27 \\ 25 - 27 \\ 29 - 27 \end{bmatrix} = \begin{bmatrix} 27 \\ -5 \\ -2 \\ 2 \end{bmatrix}$$

(d)

For the model and constraint in part (c), what does α_2 represent in the context of the study?

This α_2 represents the difference in mean weight gain between the group 2 (group receiving Supplement 20) and the baseline group that receives Supplement 60 (group 4).

(e)

For the effects model $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, where $i = 1, 2, 3, 4$ and $j = 1, 2, 3$, using the **sum-to-zero constraint** $\sum_{i=1}^{r} \alpha_i = 0$, what is the parameter vector β, the corresponding design matrix X, and the estimated $\hat{\beta}$ vector?

$$\beta = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \end{bmatrix}$$

Again, we have:

$\hat{\mu} = (22 + 25 + 29 + 32)/4 = 27$

Making our estimated $\hat{\beta}$ as follows:

$$\hat{\beta} = \begin{bmatrix} 27 \\ 22 - 27 \\ 25 - 27 \\ 29 - 27 \end{bmatrix} = \begin{bmatrix} 27 \\ -5 \\ -2 \\ 2 \end{bmatrix}$$

(f)

For the model and constraint in part (e), what does α_2 represent in the context of the study?

This α_2 represents the deviation of the mean weight for the group 2 (group receiving Supplement 20) from the overall mean weight gain across all groups.

(g)

For the models and/or constraints given in parts (b), (c), and (e), the ANOVA table will be the same provided that $X\hat{\beta}$ is the same. Verify that this is true for all 3 models/constraints.

If we assume that $X\hat{\beta}$ is the same, then the sums of squares for the model and error depend on the differences between the observed values and the fitted values. Since the observed values stay the same, and we presume the fitted values are equal, then their resultant values in the ANOVA table will all be equal. Furthermore, under each of the parametrizations, we retain the same degrees of freedom, meaning that the associated mean squared values (model, error, and total) and their resultant F-statistics remain the same. As all elements of the ANOVA table are equal under the assumption of $X\hat{\beta}$, we see they are one and the same.

As:

$$SS_{\text{error}} \equiv SS_{\text{within groups}} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} e_{ij}^2$$

and

$$SS_{model} \equiv SS_{\text{within groups}} = \sum_{i=1}^{r} n_i (\bar{Y}_i - \bar{Y})^2$$

Q3

The data table below gives the fuel economy of 8 different vehicles of the same car class from four different car companies.

The data can also be found in the car.csv file posted in Canvas.

```
library(readr)
car <- read_csv("C:/Users/samue/OneDrive/Desktop/Iowa_State_PS/STAT 5000/PS/PS5/car.csv")
```

```
## Rows: 32 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): company
## dbl (1): economy
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

(a)

Construct an ANOVA table for these data in R using the aov() function and provide a screenshot of the output

```
anovaFit <- aov(economy ~ company, data = car)
anovaFit
```

```
## Call:
##    aov(formula = economy ~ company, data = car)
##
## Terms:
##                    company Residuals
## Sum of Squares  2167.0859   14.2888
## Deg. of Freedom         3        28
##
## Residual standard error: 0.7143616
## Estimated effects may be unbalanced
```

(b)

Examine the associated parameter estimates using the $coefficients variable as described during the lab. Which model constraint does R use?

```
unique(car$company)
```

```
## [1] "M" "B" "P" "V"
```

```
anovaFit$coefficients
```

```
## (Intercept)    companyM    companyP    companyV
##     19.3000      5.4000      4.8875     21.8000
```

This uses the baseline constraint, where Company B (alphabetically the first group) is treated as the baseline.

(c)

Report the value of the F-statistic and the corresponding p-value for testing the null hypothesis of equal means for the four car companies: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$. Interpret the test result in the context of the study.

```
summary(anovaFit)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## company      3 2167.1   722.4    1416 <2e-16 ***
## Residuals   28   14.3     0.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F-statistic: 1416 p-value: p-value < 2e-16 (very small!) Interpretation: The large F-statistic corresponds to a small p-value (well below the $\alpha = 0.01$ significance threshold). This provides evidence to support rejecting the null hypothesis that each company has the same average fuel economy and to consider supporting the alternative hypothesis that at least one company has a different average fuel economy compared to the other companies (groups, in this study).

(d)

Check the ANOVA assumptions in R and include screenshots of any relevant figures. Are any of the assumptions not satisfied for this study?

Assumptions

- Normality
- Independence within and between groups
- Equal variance

Assessment:

Overall: We have reason to suspect our normality assumption and equal variances assumption are violated. We have some apprehension about within group independence being violated as well.

The independence and equal variances assumptions possibly being violated are especially problematic for the use of ANOVA.

Normality: Mean is not equal to median, but they are fairly close to one another. Furthermore, when inspecting the QQ plot of the overall data, the data does visually appear normally distributed, with minor deviations at the tails (comparing points to the reference line).

When reviewing the data at the group level (company), upon review of Shapiro-Wilk we see the following p-values:

- company == "M", p-value = 0.07217
- company == "B", p-value = 0.4919
- company == "P", p-value = 0.6419
- company == "V", p-value = 0.1982

We do not have evidence to reject the null hypothesis of the data being normally distributed for companies "B" and "P", and at $\alpha = 0.2$ we would reject the null hypothesis of normality for company "V", and at $\alpha = 0.075$ we would reject the null hypothesis of normality for company "M".

Equal Variance: With use of Levene's test, we may directly assess the equal variances assumption between all groups. With a corresponding p-value of 0.09095 and a null hypothesis that all groups have equal (the same) variance, we would reject the null hypothesis at the $\alpha = 0.1$ level.

Upon inspecting the residual plot, we would have further reason to suspect the equal variances assumption is violated, as larger fitted values have a greater spread of residuals (non-constant variance).

Also, upon reviewing the ratio of sample standard deviations, we see a ratio between the max and min of 2.54951, which leads us to suspect potential impact and violation of this equal variances assumption.

Independence:

We know that the cars in the dataset all come from the same class of car, but, in part because I am not a car expert by any stretch, we do not know whether this means they are the same car, i.e. same make and model for a given year.

As such, it would be reasonable to suspect independence is being violated in this scenario. We don't have any factors being controlled, such as differences in year or model, the latter of which when not controlled could be problematic for independence within groups to be violated.

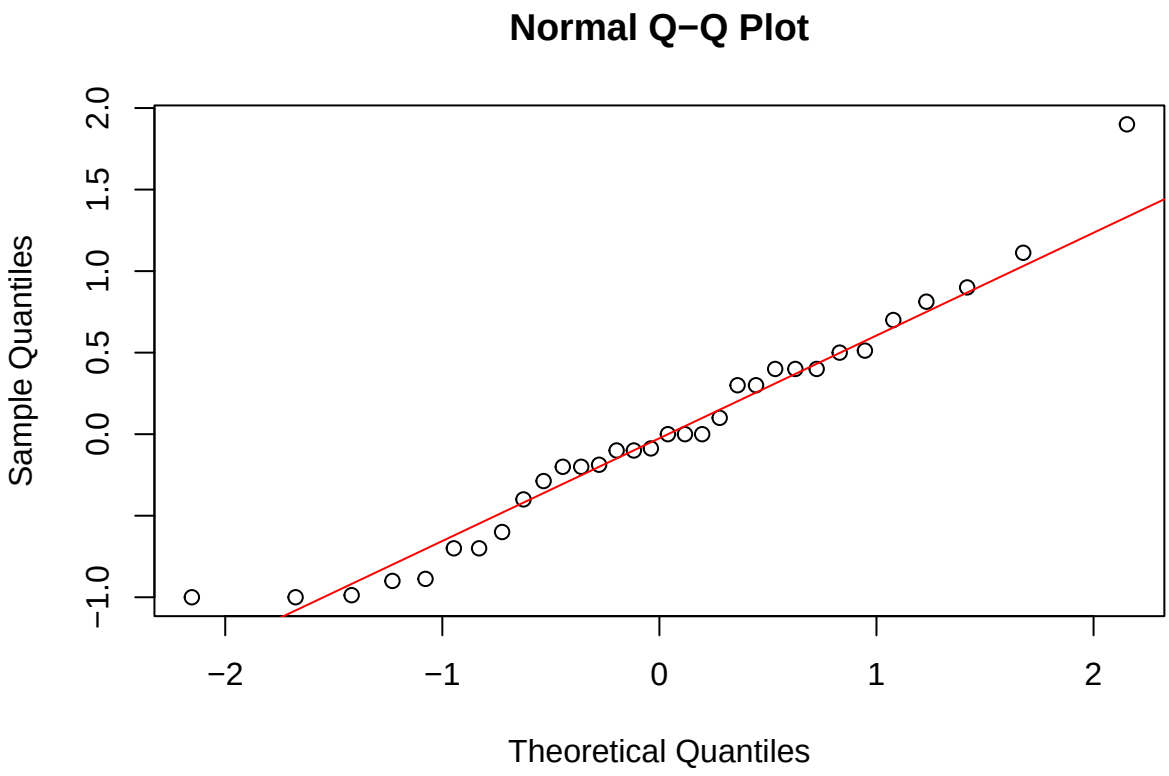
Furthermore, we have no information about whether certain parts are used by car models between and within groups, and if an especially important part, like an engine, is in use by multiple companies, then we may suspect that between group independence may also be violated.

Taken together, I'd say we have reason to suspect within group independence to be violated, but have less reason (though still some!) to suspect that between group independence is not violated.

```
summary(anovaFit$residuals)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1.00000 -0.45000 -0.04375  0.00000  0.40000  1.90000
```

```
qqnorm(anovaFit$residuals)
qqline(anovaFit$residuals, col = "red")
```




```
x1 <- sd(car$economy[car$company=="M"])
x2 <- sd(car$economy[car$company=="B"])
x3 <- sd(car$economy[car$company=="P"])
x4 <- sd(car$economy[car$company=="V"])

maxCar <- max(x1, x2, x3, x4)
minCar <- min(x1, x2, x3, x4)
maxCar/minCar
```

```
## [1] 2.54951
```

```
shapiro.test(car$economy[car$company=="M"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  car$economy[car$company == "M"]
## W = 0.83819, p-value = 0.07217
```

```
shapiro.test(car$economy[car$company=="B"])
```

```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  car$economy[car$company == "B"]
## W = 0.92731, p-value = 0.4919
```

```
shapiro.test(car$economy[car$company=="P"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  car$economy[car$company == "P"]
## W = 0.94311, p-value = 0.6419
```

```
shapiro.test(car$economy[car$company=="V"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  car$economy[car$company == "V"]
## W = 0.88234, p-value = 0.1982
```

```
library(car)
```

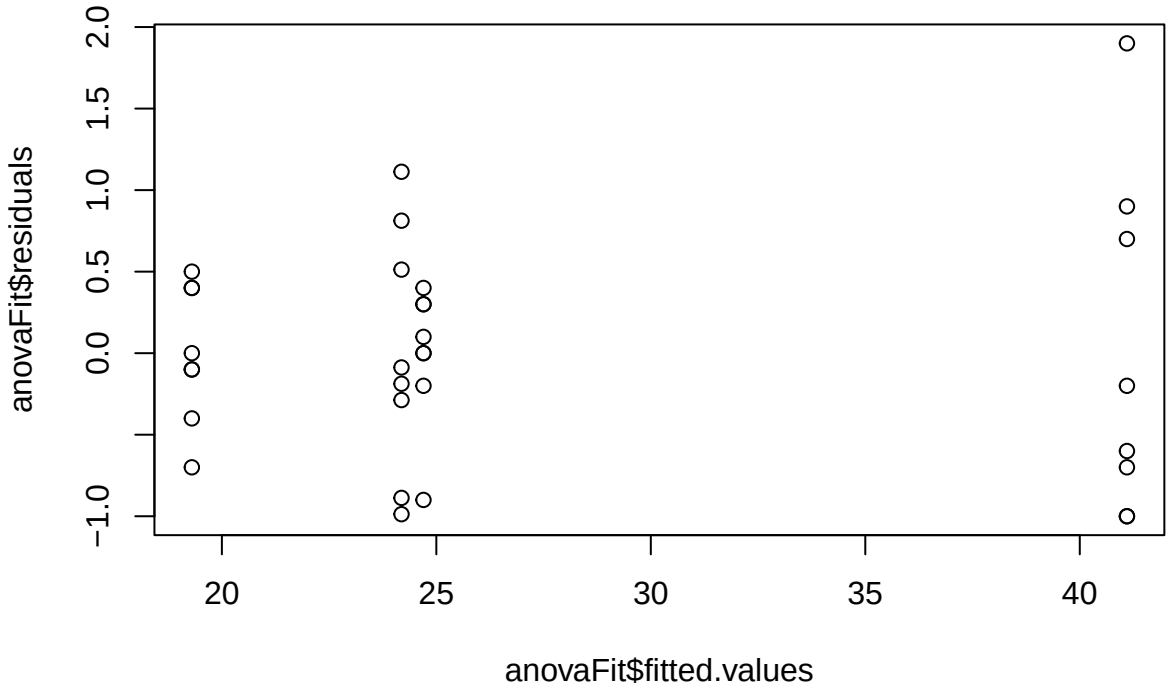
```
## Loading required package: carData
```

```
car::leveneTest(economy ~ company, data = car)
```

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

## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value  Pr(>F)
## group  3  2.3792 0.09095 .
##       28
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(anovaFit$fitted.values, anovaFit$residuals)
```



(e)

Perform a Kruskal-Wallis non-parameteric test in R for the equality of the fuel economy distributions for the four car companies. For the test, report the null and alternative hypotheses, test statistic, p-value, and interpret the test result in the context of the study.

```
kruskal.test(economy ~ company, data = car)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  economy by company
## Kruskal-Wallis chi-squared = 26.645, df = 3, p-value = 6.989e-06
```

Null: The distributions of each company is the same as the distribution of the other companies, i.e., all companies have the same distribution.
Alternative: At least one company has a different distribution than the other companies. Test Statistic: 26.645 p-value: 6.989e-06 (again, very small!)

Interpretation: The large chi-squared statistic corresponds to a small p-value (well below the $\alpha = 0.01$ significance threshold). This provides evidence to support rejecting the null hypothesis that each company has the same distribution in fuel economy and to consider supporting the alternative hypothesis that at least one company has a different distribution in fuel economy compared to the other companies (groups, in this study).