STAT 5385: Homework 4

Willliam Ofosu Agyapong

12/04/2022

Problem 6.5: Brand Preference

As a first step, we start off by importing the `brand preference` data set. Below are the variables of interest:

- X_1: Moisture content
- X_2: Sweetness
- Y: Degree of brand liking (brand preference)

Henceforth, we shall use X_1, X_2, and Y where necessary without any ambiguity.

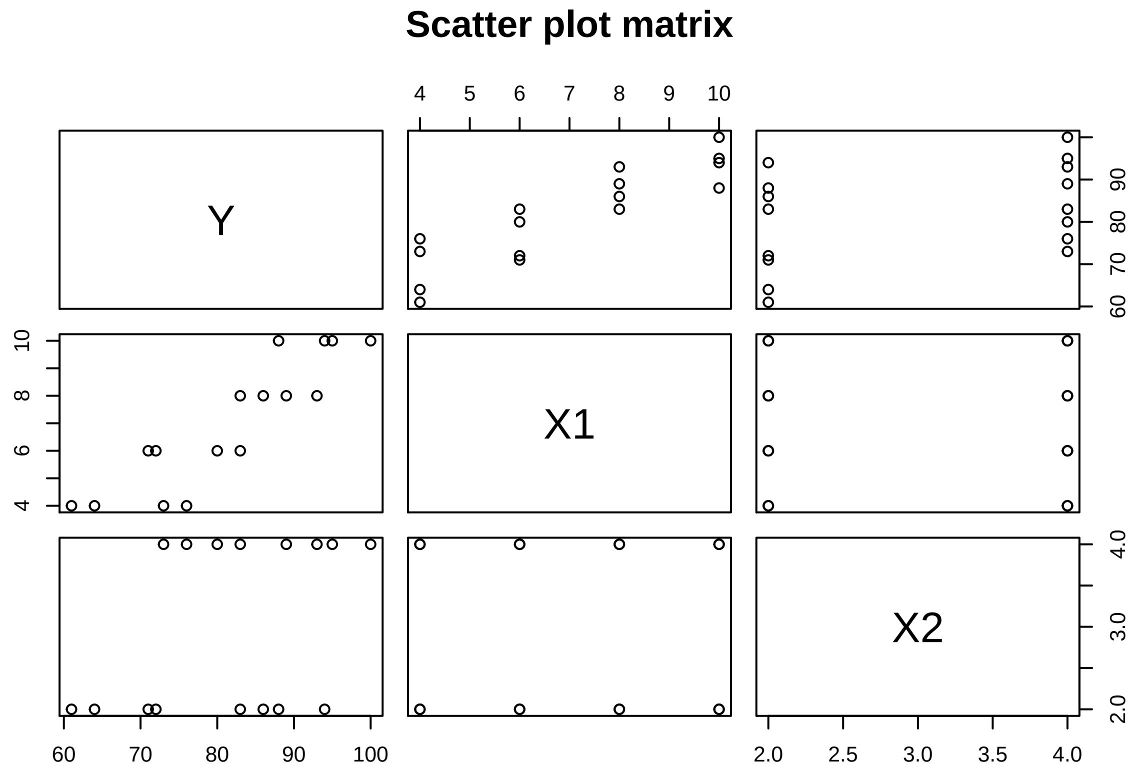
```r
brand=read.table("../Data Sets/Chapter  6 Data Sets/CH06PR05.txt")
colnames(brand)=c("Y","X1","X2")
kable(head(brand), caption = "First 6 observations")
```

Table 1: First 6 observations

Y	X1	X2
64	4	2
73	4	4
61	4	2
76	4	4
72	6	2
80	6	4

Part (a): Scatter plot matrix and correlation matrix

```r
pairs(brand, main="Scatter plot matrix")
```



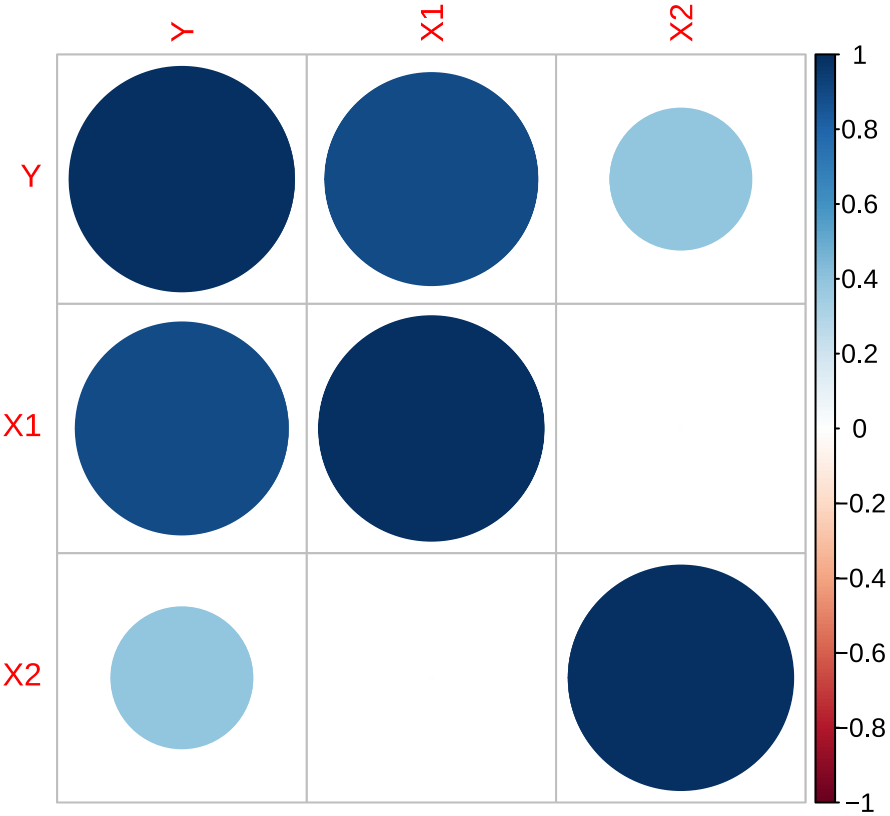
```
brand_cmat <- cor(brand)
```

Correlation matrix with associated correlation plot

```
kable(brand_cmat)
```

	Y	X1	X2
Y	1.0000	0.8924	0.3946
X1	0.8924	1.0000	0.0000
X2	0.3946	0.0000	1.0000

```
corrplot(brand_cmat)
```



The above diagrams provide insight into the direction and the strength of pairwise association existing among the underlying variables. For instance, it is clear that, with correlation coefficient of `0.8924` and by visual inspection of the correlation plot and scatter plots, there is a strong positive linear relationship between X_1 and Y. On the other hand, X_2 is weakly linearly related to Y, with a slight upward trend. The predictors X_1 and X_2 are not related or associated so multicollinearity would not be an issue here. In fact, the correlation coefficient between X_1 and X_2 is zero.

Part (b): Fitting a first order regression model to the data

```r
# muiltiple linear regression model
brand_mod <- lm(Y ~ ., data = brand)

# obtain model outputs for the estimates
brand_mod %>%
  tidy() %>%
  kable(caption = "Parameter estimates")
```

Table 3: Parameter estimates

term	estimate	std.error	statistic	p.value
(Intercept)	37.650	2.9961	12.566	0
X1	4.425	0.3011	14.695	0
X2	4.375	0.6733	6.498	0

From the above table, the estimated regression function is

$$\hat{Y} = 37.650 + 4.425X_1 + 4.375X_2$$

where X_1, X_2, and Y are as defined above.

$\hat{\beta}_1 = 4.425$ signifies that for every 1 unit increase in the moisture content we would expect the degree of liking of the product to increase by `4.425` on average while holding X_2 constant.

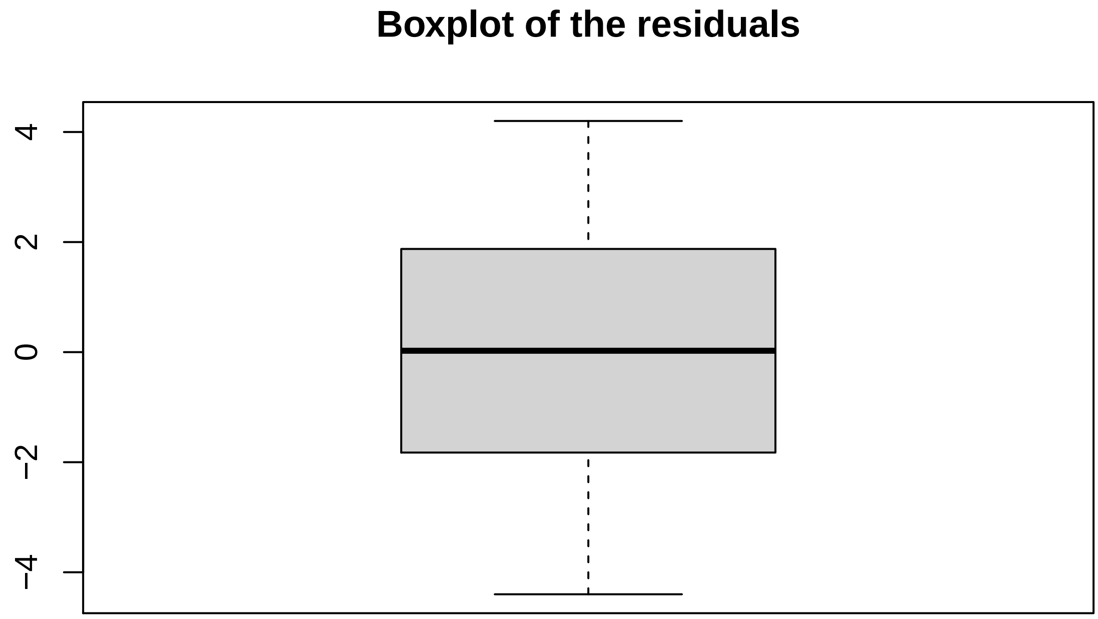
Part (c)

```r
residuals <- resid(brand_mod)
resid_df <- cbind(residuals)
kable(t(resid_df), caption = "Table of Residuals")
```

Table 4: Table of Residuals

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
residuals	-0.1	0.15	-3.1	3.15	-0.95	-1.7	-1.95	1.3	1.2	-1.55	4.2	2.45	-2.65	-4.4	3.35	0.6

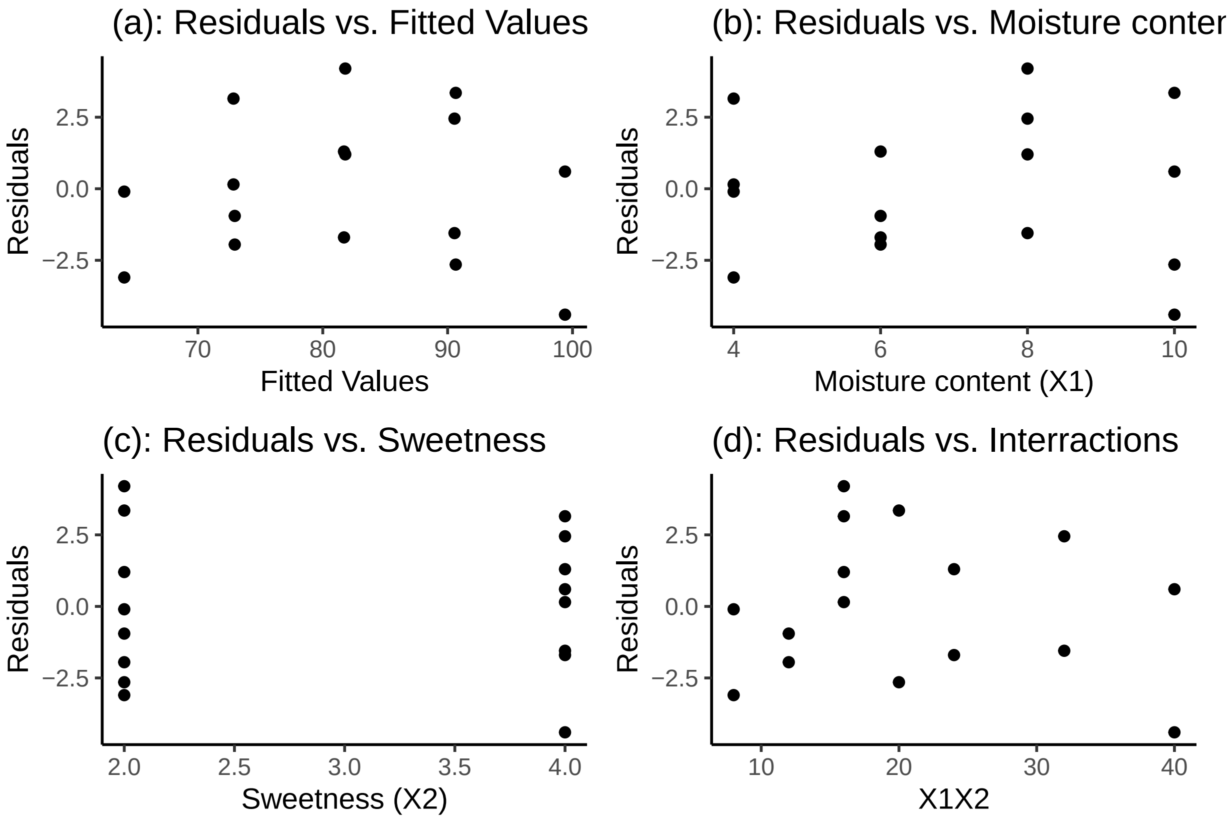
```r
boxplot(residuals, main="Boxplot of the residuals")
```



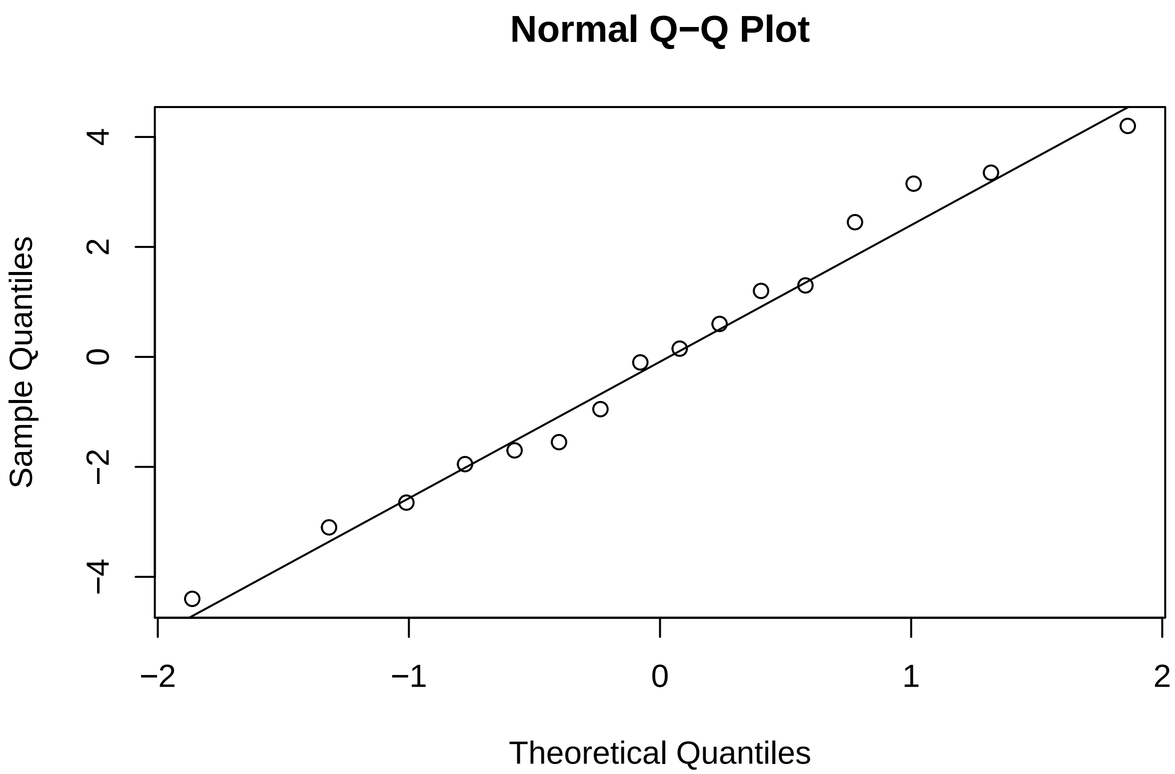
The boxplot tells us that the distribution of the residuals are symmetrical or normal with no outlying observations.

Part (d)

```r
brand_aug <- data.frame(brand, fitted = fitted(brand_mod), resid = resid(brand_mod))

# par(mfrow = c(3,2))
((ggplot(brand_aug, aes(fitted, resid)) +
  geom_point()  +
  labs(x="Fitted Values", y="Residuals", title = " (a): Residuals vs. Fitted Values")) +

(ggplot(brand_aug, aes(X1, resid)) +
  geom_point()  +
  labs(x="Moisture content (X1)", y="Residuals", title = "(b): Residuals vs. Moisture content"))) /

((ggplot(brand_aug, aes(X2, resid)) +
  geom_point()  +
  labs(x="Sweetness (X2)", y="Residuals", title = "(c): Residuals vs. Sweetness")) +
(ggplot(brand_aug, aes(X1*X2, resid)) +
  geom_point()  +
  labs(x="X1X2", y="Residuals", title = "(d): Residuals vs. Interractions")))
```



```
qqnorm(brand_aug$resid)
qqline(brand_aug$resid)
```



From figure (a), we observe that there is no systematic deviations from the response plane, indicating constant variance of the error terms and the appropriateness of a linear model. (b) and (c) also provide evidence of a linear relationship between the response variable and moisture content and sweetness, respectively. We can as well assume constant error variance with respect to the two predictor variables. Figure (d) does not exhibit any systematic pattern; hence, suggesting no interaction effect between the moisture content and sweetness of the product. Finally, the normal probability plot shows that the error terms are fairly normally distributed.

Part (e): Breusch-Pagan test

```r
library(lmtest)
bptest(brand_mod, student = FALSE)
```

```
##
##  Breusch-Pagan test
##
## data:  brand_mod
## BP = 1, df = 2, p-value = 0.6
```

```r
# qchisq(1 - 0.01, df = 2) # degrees of freedom corresponds to the number of predictors in the model
```

- Significance level, $\alpha = 0.01$.
- The two alternative hypotheses are:

$$H_0 : \text{error variance is constant}$$

$$H_A : \text{error variance is not constant}$$

- Decision rule: If p-value $\leq \alpha = 0.01$, conclude H_a, otherwise, conclude H_0.

- Since the the p-value of `0.6` is greater than `0.01`, we fail to reject H_0 and conclude that the error variance is constant.

Part (f): A formal test for lack of fit

```r
# Using an approach suggested by Laurent (stackexchange.com, 2017). Please refer to the reference section.
brand_mod2 <- lm(Y ~ factor(X1) * factor(X2), data=brand)

# Lack of fit test
kable(anova(brand_mod, brand_mod2), caption = "Lack of fit test results")
```

Table 5: Lack of fit test results

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
13	94.3	NA	NA	NA	NA
8	57.0	5	37.3	1.047	0.453

- Desired significance level, $\alpha = 0.01$.
- We are interested in testing the following two alternative hypotheses:

$$H_0 : \text{the regression function is linear (there is no lack of fit)}$$

$$H_a : \text{the regression function is not linear (there is lack of fit)}$$

- The decision rule is to reject H_0 if p-value of the test $\leq \alpha = 0.01$.
- The F test-statistic turns out to be `1.047` with corresponding p-value `0.45`. Since this p-value is greater than `0.01`, we will fail to reject H_0 and conclude that the regression function is linear. That is, there is no lack of fit.

Problem 6.6: In reference to Brand Preference (Problem 6.5)

Part (a): Is there a regression relation?

```r
# computing the necessary measures for the test.
brand_mod %>% glance() %>%
  select(F_statistic = statistic, p.value, df, df.residual) %>%
  kable(caption = "F test resutls")
```

Table 6: F test resutls

F_statistic	p.value	df	df.residual
129.1	0	2	13

- Significance level, $\alpha = 0.01$.
- Given the first-order regression model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, i = 1, 2, \cdots, 16$$

  where β_1 and β_2 are the true parameter coefficients associated with the moisture content and the sweetness of the product, respectively. We are therefore interested in testing the following two competing hypotheses:

$$\begin{aligned} &H_0 : \beta_1 = \beta_2 = 0 \\ &H_a : \text{at least one } \beta_k \neq 0, k = 1, 2 \end{aligned}$$

- From the model summary output, the test statistic is

$$F^* = \frac{MSR}{MSE} = 129 \text{ on 2 and 13 degrees of freedom, with associated p-value} = 2.66e - 09 \approx 0$$

- The decision rule is to reject H_0 if p-value $\leq \alpha = 0.01$.
- Here, we reject H_0 since the p-value is less than `0.01`, and conclude that at least one of the independent variables is helpful in predicting the degree of brand liking. That is, brand preference is related to the moisture content and the sweetness of the product.
- Thus, the test implies that at least one of β_1 and β_2 is significant.

Part (b)

The p-value for the test in part (a) is $2.66e - 09 \approx 0$.

Part (c):

```r
kable(confint(brand_mod, level = (1-0.01/2))) # g = 2
```

	0.25 %	99.75 %
(Intercept)	27.546	47.754
X1	3.409	5.441
X2	2.104	6.646

From the above output, the joint confidence intervals for the slope parameters by the Bonferroni procedure using 99% family confidence coefficient are

$$3.409 \leq \beta_1 \leq 5.441, \quad \text{and}$$
$$2.104 \leq \beta_2 \leq 6.645$$

Therefore, with family confidence coefficient 0.99, we conclude that β_1 falls between `3.409` and `5.441` while β_2 falls between `2.104` and `6.645`.

Problem 6.7: In reference to Brand Preference Problem 6.5

Part (a): coefficient of multiple determination (R^2)

```r
# Computing the required sums of squares
 brand_mod %>%
  glance() %>%
  select(r.squared, sigma, F.statistic=statistic, p.value, df, df.residual) %>%
  kable(caption = "Model level results from the original model")
```

Table 8: Model level results from the original model

r.squared	sigma	F.statistic	p.value	df	df.residual
0.9521	2.693	129.1	0	2	13

$$R^2 = \frac{SSR}{SST} = 0.952$$

This value means that, approximately 95.2% of the variation in brand preference (Y) can be explained by using the independent variables, moisture content and sweetness of the product, to estimate Y.

Part (b):

To answer this question, we fit a new regression model of Y as a function of $\hat{Y}$, where $\hat{Y}$ is the fitted values from the ariginal model.

```
# Computing the required sums of squares
 (lm(Y ~ fitted, data = brand_aug)) %>%
  glance() %>%
  select(r.squared, sigma, F.statistic=statistic, p.value, df, df.residual) %>%
  kable(caption = "Model level results")
```

Table 9: Model level results

r.squared	sigma	F.statistic	p.value	df	df.residual
0.9521	2.595	278	0	1	14

From above table, the coefficient of simple determination r^2 between Y_i and $\hat{Y}_i$ is

$$r^2 = \frac{SSR}{SST} \approx 0.952$$

Yes, the coefficient of simple determination here is equal to the coefficient of multiple determination in part (a).

Problem 6.8

Part (a)

Obtain an interval estimate of $E[Y_h]$ when $X_{h1} = 5$ and $X_{h2} = 4$, with a 99% confidence coefficient.

```
kable(predict(brand_mod, newdata = data.frame(X1=5, X2=4), interval = "confidence", level = 0.99))
```

fit	lwr	upr
77.28	73.88	80.67

The required 99% confidence interval estimate of $E\{Y_h\}$ is: $73.88 \leq E\{Y_h\} \leq 80.67$. Thus, we conclude with 99% confidence that the mean response for brand preference lies between `73.88` and `80.67` when the moisture content and sweetness of the product are 5 and 4, respectively.

Part (b)

Obtain a prediction interval for a new observation $Y_{h(new)}$ when $X_{h1} = 5$ and $X_{h2} = 4$?, with a 99% confidence coefficient.

```
kable(predict(brand_mod, newdata = data.frame(X1=5, X2=4), interval = "prediction", level = 0.99))
```

fit	lwr	upr
77.28	68.48	86.07

The required 99% prediction interval is: $68.48 \leq Y_{h(new)} \leq 86.07$. Meaning, we predict with 99% confidence that the new observatin will be somewhere between `68.48` and `86.07`.

Reference

- Kutner, M. H., Nachtsheim, C. J., Neter, J., & Wasserman, W. (2004). Applied linear regression models (Vol. 4). New York: McGraw-Hill/Irwin..
- Stéphane Laurent (https://stats.stackexchange.com/users/8402/st%c3%a9phane-laurent), F-test for lack of fit using R, URL (version: 2017-07-05): https://stats.stackexchange.com/q/288976