Linear model to predict harmfullness of cigarettes

Léa MARINGER

October 2022

1 Purpose

We have data about cigarettes and we would like to explain cigarettes harmfulness (carbon monoxide amount (CO)) (Y) depending on 3 variables:

- tar quantity (X1),
- nicotine quantity (X2),
- and weight (X3).

We have n = 24 observations.

Cigarette	TAR (mg)	NICOTINE (mg)	WEIGHT (g)	CO (mg)
Alpine	14.1	0.86	0.9853	13.6
Benson&Hedges	16	1.06	1.0938	16.6
CamelLights	8	0.67	0.928	10.2
Carlton	4.1	0.4	0.9462	5.4
Chesterfield	15	1.04	0.8885	15
GoldenLights	8.8	0.76	1.0267	9
Kent	12.4	0.95	0.9225	12.3
Kool	16.6	1.12	0.9372	16.3
L&M	14.9	1.02	0.8858	15.4
LarkLights	13.7	1.01	0.9643	13
Marlboro	15.1	0.9	0.9316	14.4
Merit	7.8	0.57	0.9705	10
MultiFilter	11.4	0.78	1.124	10.2
NewportLights	9	0.74	0.8517	9.5
Now	1	0.13	0.7851	1.5
OldGold	17	1.26	0.9186	18.5
PallMallLight	12.8	1.08	1.0395	12.6
Raleigh	15.8	0.96	0.9573	17.5
SalemUltra	4.5	0.42	0.9106	4.9
Tareyton	14.5	1.01	1.007	15.9
VV	7.3	0.61	0.9806	8.5
ViceroyRichLight	8.6	0.69	0.9693	10.6
VirginiaSlims	15.2	1.02	0.9496	13.9
WinstonLights	12	0.82	1.1184	14.9

Figure 1: Dataset

We want to find a model in the following form:

$$Y = a_0 + a_1 X_1 + a_2 X_2 + a_3 X_3 + \epsilon \tag{1}$$

and we suppose that $\epsilon \sim \mathcal{N}(0_24,\, \sigma^2 I_{24})$.

Notation: If X is a matrix, X' is the transpose of X.

2 Calculation of the model with Excel

You can find calculation of the model, prediction and variance analysis on the Excel file.

2.1 A few formulas

- Coefficients estimator: $\hat{a} = (X'X)^{-1}X'Y$
- $\hat{Y} = X\hat{a}$
- $Y = X\hat{a} + \epsilon$
- $\epsilon = Y - \hat{Y}$
- â variance estimator: $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} {\epsilon_i}^2}{n-(p+1)}$
- Var(â) estimator : $\hat{Var}(\hat{a}) = \hat{\sigma}^2(X'X)^{-1}$
- Explained sum of squares: $ESS = \sum_{i=1}^{n}(\hat{Y_i} - Ymean)^2$
- Residual sum of squares: $RSS = \sum_{i=1}^{n}(Y_i - \hat{Y_i})^2$
- Total sum of squares: $TSS = \sum_{i=1}^{n}(Y_i - Ymean)^2$
- Coefficient of determination: $R^2 = \frac{ESS}{TSS}$

2.2 Results

- $\hat{a} = \begin{pmatrix} -0.5517 \\ 0.8876 \\ 0.5185 \\ 2.0793 \end{pmatrix}$
- $\hat{\sigma}^2 = 1.35$
- $\hat{Var}(\hat{a}) = \begin{pmatrix} 8.8285 & 0.0846 & -1.2632 & -9.0396 \\ 0.0846 & 0.0382 & -0.6080 & -0.0205 \\ -1.2632 & -0.6080 & 10.5777 & -0.5367 \\ -9.0396 & -0.0205 & -0.5367 & 10.1023 \end{pmatrix}$
- $R^2 = 0.935$

Conclusion: 93.5% of variation of CO quantity in cigarettes is explained by the 3 variables (X_1, X_2 *and* X_3).

3 Model and analysis with R

You can find all the R code on the R file.

3.1 Point clouds

We can see that there is a correlation between the quantity of tar and the quantity of nicotine in cigarettes, but it also seems to have a correlation between these two variables and the target variable, carbon monoxide (CO) quantity.

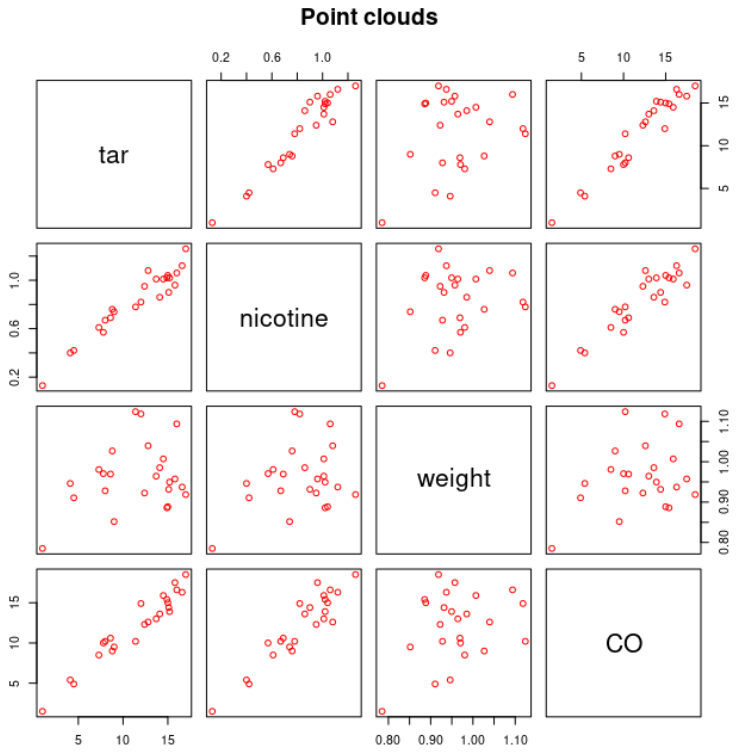

Figure 2: Point clouds 2 variables by 2 variables

3.2 Model coefficients

When we create the model and then print it, we obtain the following estimated coefficients vector:

$$\hat{a} = \begin{pmatrix} -0.5517 \\ 0.8876 \\ 0.5185 \\ 2.0793 \end{pmatrix}$$

This is, of course, exactly the same vector we found in the precedent section by calculating $\hat{a}$ on Excel.

3.3 Model analysis

First, I wanted to verify that residuals are well centered. I found $\mathbb{E}(\epsilon) = -6.13e-17$. It is indeed a very small value close to zero. We can also check that graphically (Figure 3). We also see that residuals are not correlated.

We also find a coefficient of determination of $R^2 = 0.935$, which means that 93.5% of the cigarettes harmfulness is explained by the quantity of tar and nicotine and the weight of the cigarette. We can then affirm that the model is good because R^2 is close to 1.

I calculated Student threshold with a risk $\alpha = 10\%$ and I found $t = 1.73$. That means that Student residuals which are included in $[-1.73\,;1.73]$ are outliers. We can see that there are 2 values (2 types of cigarettes) which are outliers: MultiFilter and WinstonLights.

An other way to detect outliers is the observations leverages. After calculating it, I calculated the threshold $(2\frac{(p+1)}{n})$. For a given observation, this observation is an outlier if its leverage is less than the threshold, which is $\frac{1}{3}$ here. With this method, I found one outlier: Now cigarettes.

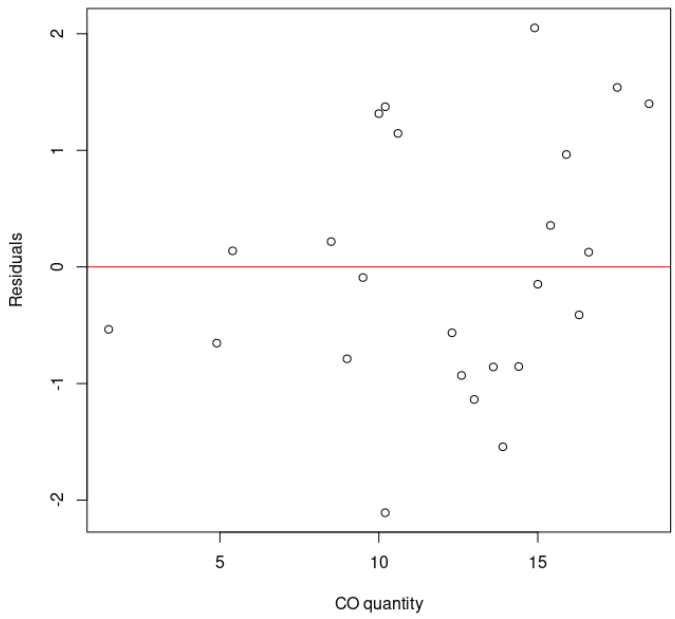

Figure 3: Residuals distribution

```
Call:
lm(formula = CO ~ tar + nicotine + weight, data = cig)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1083 -0.8046 -0.1199  1.0095  2.0501

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.5517     2.9713  -0.186 0.854569
tar           0.8876     0.1955   4.540 0.000199 ***
nicotine      0.5185     3.2523   0.159 0.874941
weight        2.0793     3.1784   0.654 0.520431
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.16 on 20 degrees of freedom
Multiple R-squared:  0.935,     Adjusted R-squared:  0.9252
F-statistic: 95.86 on 3 and 20 DF,  p-value: 4.85e-12
```

Figure 4: Model summary

We can remove those 3 outliers and create a new model. We found $R^2 = 0.938$, almost exactly the same value as before. Removing the outliers has not really improve the model.

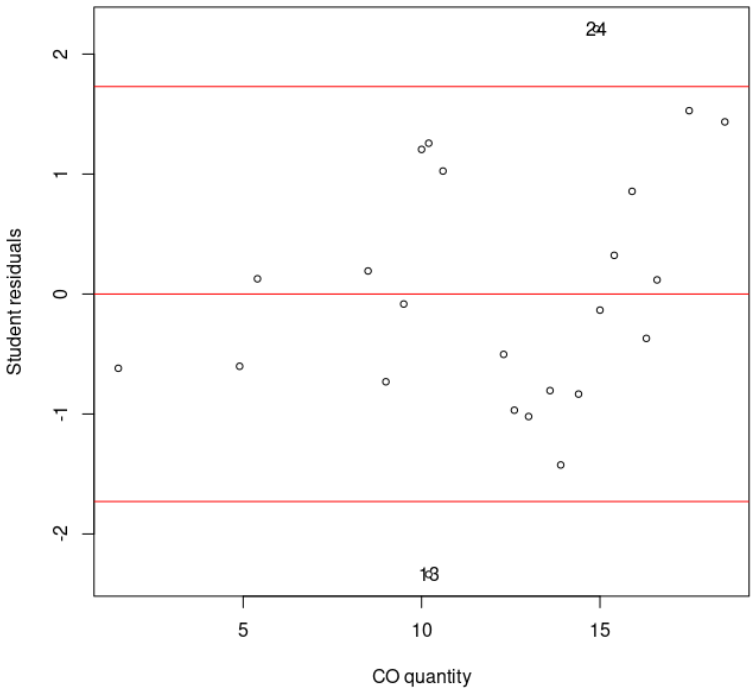

Figure 5: Student residuals