Practice 3: MULTIPLE LINEAR REGRESSION

### METHODOLOGY: Linear Regression

****

Preliminary univariate and bivariate exploratory analysis:

• Histogram / Boxplot and numerical statistics of the variables

• Diagram of scatterplot with the explanatory variable on the abscissa axis and the response to the ordinate axis

Relationship identification:

• Determination of the appropriate functional dependence (linear, quadratic, ...) and, if necessary, transformations

Estimation of the regression line:

• Coefficients and expression of the adjusted line

• Assessment of the degree of adjustment: Coefficient of determination (R2) and R2 Adjusted

• Inference on the parameters of the model: Meaning of the coefficients and ANOVA table

Validation of the premises of the model (Waste Analysis):

• Linearity: waste bivariate plot vs. predictions

• Normal: Normal probability plot (qqnorm)

• Homocedasticity (constant variance): plot bivariate residues vs predictions

• Independence

• Identification of atypical data (outliers) and influential data (a priori and a posteriori)

In the case of multiple linear regression, it should be borne in mind that there is more than one explanatory variable, and therefore new elements and procedures need to be introduced in the phases of model identification, estimation, and validation. Identification:

• Missing data analysis and treatment

• Possible correlation between regressors (multicolineality): VIFs (squared root means inflation). For GVIF normalized for df, squared root is not necessary.

• Selection of the best subset of regressors: compromise between fit and parsimony

Estimation:

• Coefficients of the fitted line and the fitted line on the standardized variables

• Partial correlation coefficient

Validation:

• Residual bivariate plot vs explanatory variables in the model

• Residual bivariate plot vs explanatory variables excluded from the model

# Case 1. smokCancer

**Name: smokCancer (Smoking and Cancer)**

**Reference: J.F. Fraumeni, "Cigarette Smoking and Cancers of the Urinary Tract: Geographic Variations in the United States," Journal of the National Cancer Institute, 41, 1205-1211, (1968)**

**Description:**

Number of cigarettes consumed (sold) per capita in 43 states and the District of Columbia in 1960 along with death rates per 100,000 population for different types of cancer.

**Number of cases: 44**

**Variables:**

1. IGC = Number of cigarettes consumed (hundreds per capita)

2. BLAD = Died per 100K inhabitants from bladder cancer

3. LONG = Deaths per 100K inhabitants from lung cancer

4. KID = Died per 100K inhabitants from kidney cancer

5. LEUK = Died per 100K inhabitants from leukemia

STATE CIG BLAD LUNG KID LEUK

AL 18.20 2.90 17.05 1.59 6.15

AZ 25.82 3.52 19.80 2.75 6.61

AR 18.24 2.99 15.98 2.02 6.94

CA 28.60 4.46 22.07 2.66 7.06

CT 31.10 5.11 22.83 3.35 7.20

DE 33.60 4.78 24.55 3.36 6.45

DC 40.46 5.60 27.27 3.13 7.08

FL 28.27 4.46 23.57 2.41 6.07

ID 20.10 3.08 13.58 2.46 6.62

IL 27.91 4.75 22.80 2.95 7.27

IN 26.18 4.09 20.30 2.81 7.00

IO 22.12 4.23 16.59 2.90 7.69

KS 21.84 2.91 16.84 2.88 7.42

KY 23.44 2.86 17.71 2.13 6.41

LA 21.58 4.65 25.45 2.30 6.71

ME 28.92 4.79 20.94 3.22 6.24

MD 25.91 5.21 26.48 2.85 6.81

MA 26.92 4.69 22.04 3.03 6.89

MI 24.96 5.27 22.72 2.97 6.91

MN 22.06 3.72 14.20 3.54 8.28

MS 16.08 3.06 15.60 1.77 6.08

MO 27.56 4.04 20.98 2.55 6.82

MT 23.75 3.95 19.50 3.43 6.90

NB 23.32 3.72 16.70 2.92 7.80

NE 42.40 6.54 23.03 2.85 6.67

NJ 28.64 5.98 25.95 3.12 7.12

NM 21.16 2.90 14.59 2.52 5.95

NY 29.14 5.30 25.02 3.10 7.23

ND 19.96 2.89 12.12 3.62 6.99

OH 26.38 4.47 21.89 2.95 7.38

OK 23.44 2.93 19.45 2.45 7.46

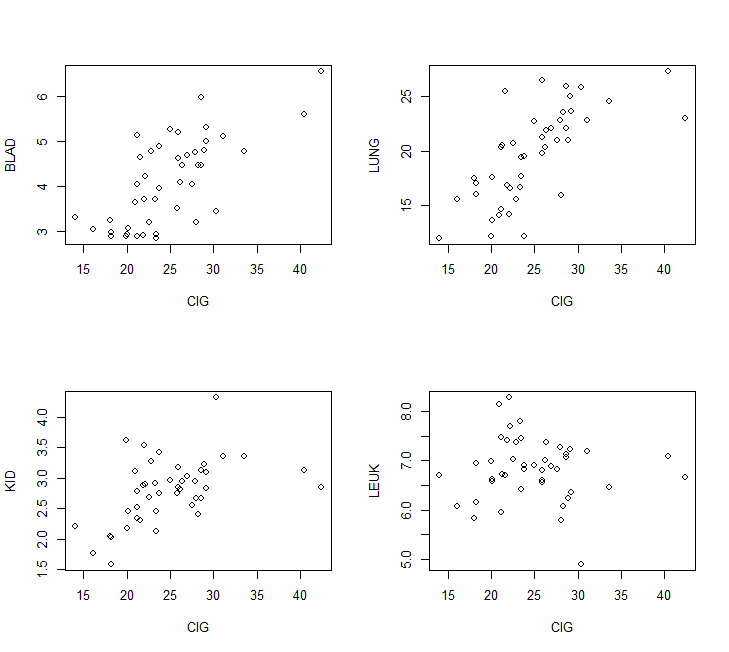
PE 23.78 4.89 12.11 2.75 6.83

**Aim:** To study the relationship of the tobacco factor with any of the 4 reported types of cancer.

# Exercise 1: smokCancer: Is there any relationship between smoking and cancer?

1. Apply simple linear regression to explore the relationship between tobacco and other types of cancer.

2. For the four situations, specify: the variables involved, the linear relationship based on the adjusted model, the measures to assess this relationship and the analysis of waste to validate the model.

3. What conclusions can be drawn from the analysis of this data?

# Case 2. Aliments

**Nom: Aliments**

**Reference:** What explains the calories in food?

**Description:**

Nutrition labeling information for 58 foods was taken.

**Number of observations:** 58

**Variables:**

1. Producte: Product name
2. Kcalories: Energy for 100g product
3. Proteines: Protein content (g)
4. HidratsCarboni: Carbohydrate Content (g)
5. Sucres: Carbohydrate Content in the Form of Sugar (g)
6. Greixos: Fat Content (g)
7. GreixosSaturats: Saturated Fat Content (g)
8. Fibra: Fiber Content (g)
9. Sodi: Sodium content (g)

Producte Kcalories Proteines HidCarb Sucres Greixos GreixSat Fibra Sodi

Zumo de naranja 45 0,6 10 10 0,1 0 0,2 0,002

Leche entera 62 3 4,5 4,5 3,6 2,43 0 0,05

Bebida de soja 33 3,6 0,7 0,6 1,8 0,3 1 0,04

Margarina vegetal ligera 370 2 0,1 0,1 40 11 0 0,29

Te sin azucar 6 0 0 0 0 0 0 0

Cereales con chocolate 445 6 75 29 13 4,6 2 0,25

Pasta con vegetales 344 12 67 4 2 0,5 5 0,02

Chocolate soluble 378 6,9 78 70 2,4 1,5 8,4 0,1

Miel de flores 290 0,6 71,8 71,8 0 0 0 0

Edulcorante granulado 351 3,2 84,6 84,6 0 0 0,6 0,66

Pan rayado 365 9,6 77,3 3,1 1,2 0,3 3,2 0

Espinacas frescas 23 2,4 1,9 0 0,2 0 2,1 0,06

Galletas fibra-chocolate 498 6,1 64,2 23,6 23,5 12,7 2,8 0,55

Yogourt griego con fresa 147 1,9 17,2 13,2 7,8 5,2 0 0,04

Panzerotti ricota-espinacas 280 11 40 1,3 8 3,8 2,3 0,47

Danonino 109 6,4 14,2 13,4 2,7 1,7 1,3 0,04

Chocolate negro intenso 590 8 30,7 26,4 44,9 26,6 15,3 8

Galletas chocolate 482 6 69 36 20 10 1 0,2

Galletas clasicas 472 5,9 71,5 18,7 18 9,9 2,4 0,15

Bizcochito fruta-fibra 383 4 58 38 13 1 9 0,18

Pan tostado 390 10 74 4,9 6 2,6 4,3 0,42

Chocolate soluble light 0,2 0 0 0 0 0 0 0

Preparado lácteo 77 2,7 13,5 11 1,4 0,9 0 0,04

Reductor colesterol 83 3 15,1 13 1,2 0,6 0,3 0,2

Queso fresco Quark 132 11,1 4,4 4,4 7,7 4,6 0 0,05

Mousse 113 3,2 10,1 10,1 6,6 4 0 0,02

Yogourt light 51 5,1 5,5 5,5 0,5 0,26 0 0,07

Queso rayado Emmental 365 28,3 0 0 28 19,9 0 0,29

Lomo embutido 212 38 0 0 6,7 2,3 0 1,17

Queso fresco de burgos 199 10,7 3,6 3,6 15,5 10,8 1,5 0,31

Gofres 444 5 55,4 29,6 22,3 11,7 0,9 0,23

Filetes de Anchoa 179 24,6 0,7 0 8,6 2,2 0 6,08

Chocolate blanco 540 5 62,5 59,7 30 19,4 0 0,09

Patatas fritas light 458 7,3 59 0,8 20,1 2,3 6,1 0,9

Tortitas de maiz 406 9 87 1 2 0,3 2 0,3

Atún claro 247 22,6 0,2 0 17,3 2,7 0 0,51

Cereales con chocolate 431 6,6 69,7 31,3 12,6 2 6,2 0,13

Almohadillas choco 397 10 64 19 9 5 10 0,02

Galletas 447 7,5 68,5 26,4 15 7,8 4,3 0,5

Lentejas 96 7,9 9,4 1,8 1,8 0,9 5,3 0,15

Zumo de naranja 41 0,6 9,5 8,9 0,1 0 0,6 0

Queso cremoso 202 8,3 4 4 17 11,2 2 0,5

Reductor colesterol 45 3,2 4,7 4,3 1,1 0,1 0,1 0,05

Arroz integral 183 4,9 30,6 0,5 3,8 0,7 1,5 0,2

Lomo embuchado 223 35 0,5 0,5 9 3 0 1

Soja 45 3,7 2,5 2,5 2,2 0,4 0,6 0,06

Sardinas 212 19,3 0 0 15 3,1 0 0,2

Uva en almíbar 67 0,4 15,5 15,3 0,1 0,1 1 0,006

Crema de verduras 70 1 6 3,5 4,5 2 1,5 0,31

Pasta con vegetales 354 11 72 3,5 1,5 0,3 4 0,03

Piña en almíbar 69 0,4 15,5 0,1 0,1 0 0,8 0,001

Tostadas integrales 377 10,6 71 4,3 4,7 1,9 4,3 0,57

Arroz blanco 343 6,8 78 0 0,45 0,1 1 0,04

Galletas Maria 436 7,4 77 23 11 2,1 2 0,31

Mezcla de setas 16 1,2 1,5 0 0,2 0,1 1,5 0,02

Merluza 71 16,9 0 0 0,4 0,1 0 0,2

Preparado de paella 75 7,8 4,3 1,8 2,6 0,5 1,9 0,57

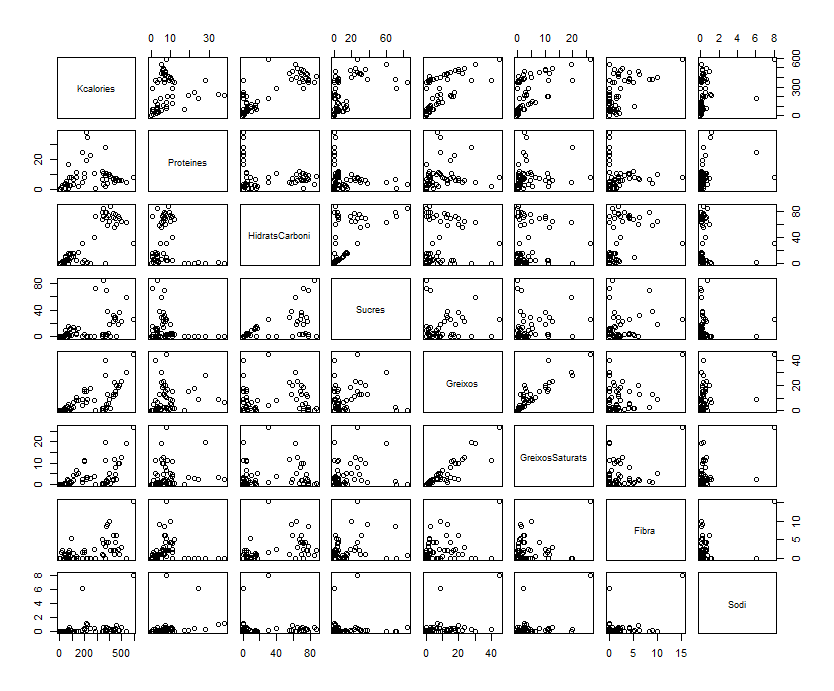
Cuajada 137 7,94 3,58 3,1 10,13 6,6 0,5 0,1

# Exercise 2: Aliments. Which components determine the energy?

1. Analyze the relationship between the different food components and the energy (in Kcal) of food.

2. Make a univariate study for each variable, using simple linear regression. Are the models validated?

3. Obtain the corresponding multiple linear regression model and validate it.



# Case 3. Attitude

**Name: attitude** (Chatterjee–Price Attitude Data)

**Reference:** Chatterjee, S. and Price, B. (1977) *Regression Analysis by Example*. New York: Wiley. (Section 3.7, p.68ff of 2nd ed.(1991).

**Description:**

Based on a survey of office employees of a large financial institution, data are aggregated based on questionnaires from approximately 35 employees for each of the 30 (randomly selected) departments. The numbers indicate the proportion of favorable answers to a set of questions in each department and the rating variable it contains.

**Number of observations:** 30

**Variables:**

1. rating: Overall rating of job being done by supervisor
2. complaints: Handling of employee complaints
3. privileges: Does not allow special privileges
4. learning : Opportunity to learn
5. raises: Raises based on performance
6. critical: Too critical of poor performance
7. advance: Advancement (Rate of advancing to better jobs)

rating complaints privileges learning raises critical advance

1 43 51 30 39 61 92 45

2 63 64 51 54 63 73 47

3 71 70 68 69 76 86 48

4 61 63 45 47 54 84 35

5 81 78 56 66 71 83 47

6 43 55 49 44 54 49 34

7 58 67 42 56 66 68 35

8 71 75 50 55 70 66 41

9 72 82 72 67 71 83 31

10 67 61 45 47 62 80 41

11 64 53 53 58 58 67 34

12 67 60 47 39 59 74 41

13 69 62 57 42 55 63 25

14 68 83 83 45 59 77 35

15 77 77 54 72 79 77 46

16 81 90 50 72 60 54 36

17 74 85 64 69 79 79 63

18 65 60 65 75 55 80 60

19 65 70 46 57 75 85 46

20 50 58 68 54 64 78 52

21 50 40 33 34 43 64 33

22 64 61 52 62 66 80 41

23 53 66 52 50 63 80 37

24 40 37 42 58 50 57 49

25 63 54 42 48 66 75 33

26 66 77 66 63 88 76 72

27 78 75 58 74 80 78 49

28 48 57 44 45 51 83 38

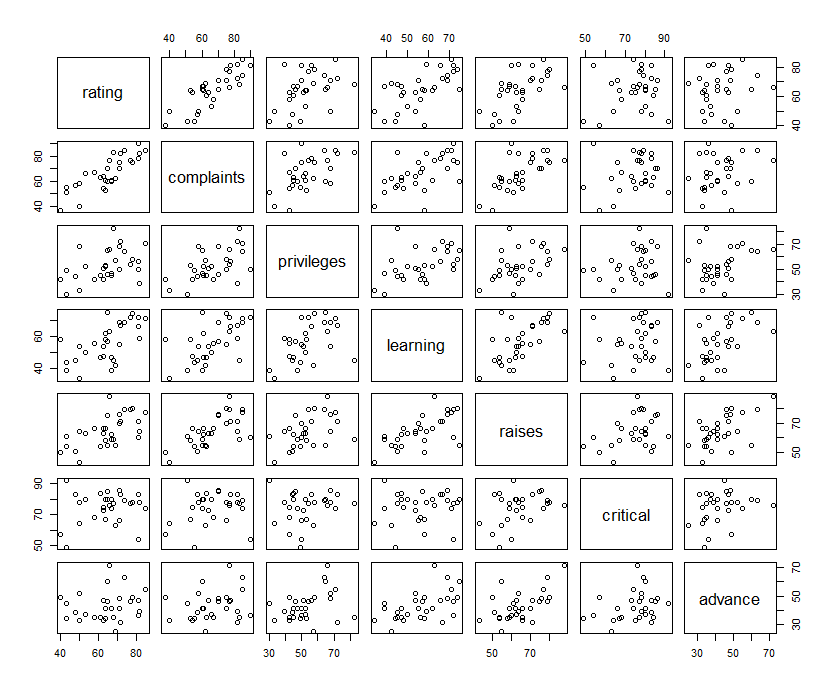
29 85 85 71 71 77 74 55

30 82 82 39 59 64 78 39

**Objective: To study which factors determine the overall assessment (supervisor rating)**

# Exercise 3: attitude. Which are the most associated variables to global score ?

1. Determine a multiple linear regression model to explain the overall score obtained



## Case 4: Prestige of Canadian Occupations in data.frame Prestige in car library for R (Fox and Weisber 2011)

Description: The Prestige data frame has 102 rows and 6 columns. The observations are occupations. This data frame contains the following columns:

|  |  |
| --- | --- |
| Education | Average education of occupational incumbents, years, in 1971. |
| Income | Average income of incumbents, dollars, in 1971. |
| Women | Percentage of incumbents who are women. |
| Prestige | Pineo-Porter prestige score for occupation, from a social survey conducted in the mid-1960s. |
| Census | Canadian Census occupational code. |
| Type | Type of occupation. A factor with levels (note: out of order): bc, Blue Collar; prof, Professional, Managerial, and Technical; wc, White Collar. |

**Source**

Canada (1971) Census of Canada. Vol. 3, Part 6. Statistics Canada [pp. 19-1–19-21].

**> summary(Prestige)**

**education income women prestige census type**

**Min. : 6.380 Min. : 611 Min. : 0.000 Min. :14.80 Min. :1113 bc :44**

**1st Qu.: 8.445 1st Qu.: 4106 1st Qu.: 3.592 1st Qu.:35.23 1st Qu.:3120 prof:31**

**Median :10.540 Median : 5930 Median :13.600 Median :43.60 Median :5135 wc :23**

**Mean :10.738 Mean : 6798 Mean :28.979 Mean :46.83 Mean :5402 NA's: 4**

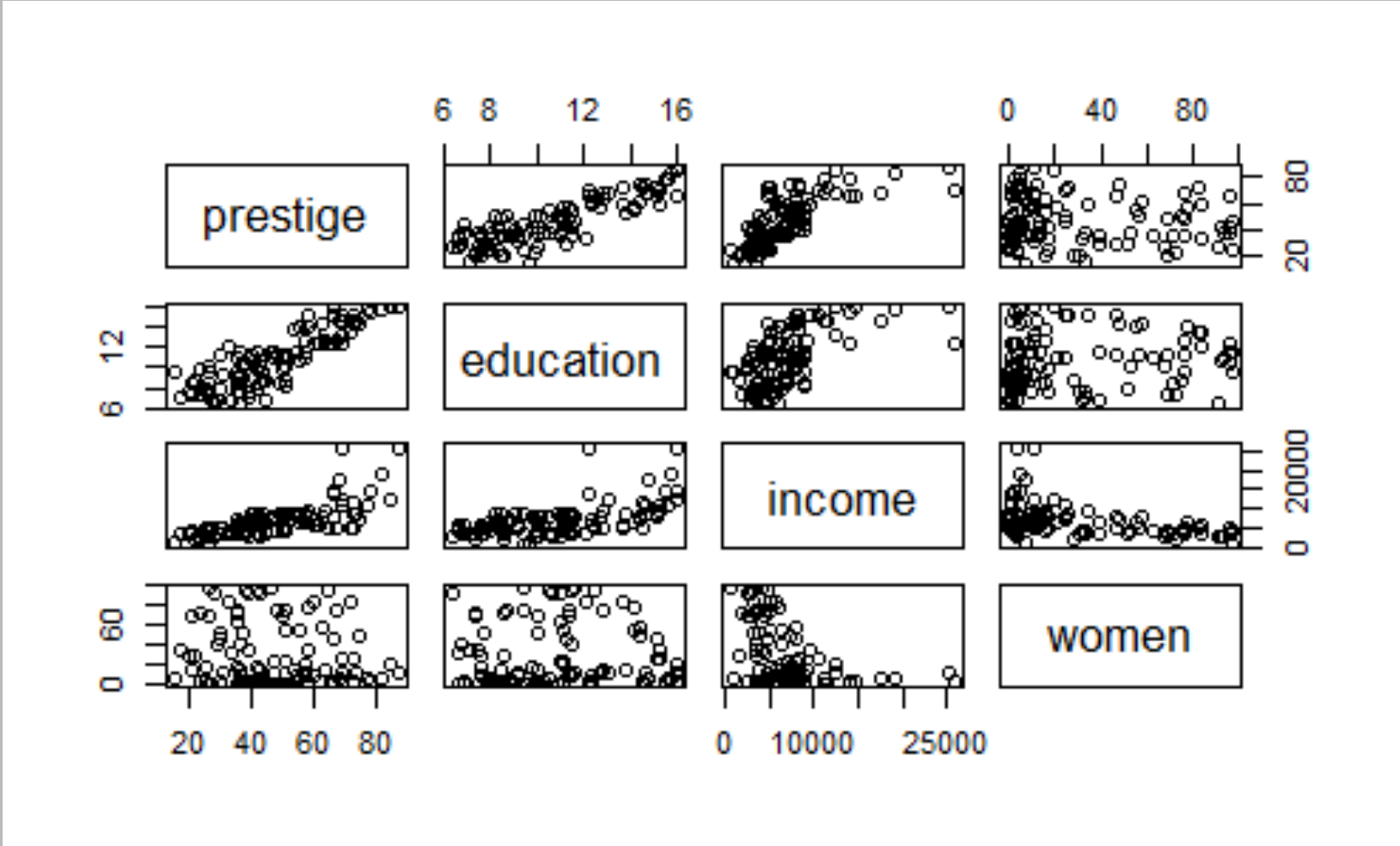
**3rd Qu.:12.648 3rd Qu.: 8187 3rd Qu.:52.203 3rd Qu.:59.27 3rd Qu.:8312**

**Max. :15.970 Max. :25879 Max. :97.510 Max. :87.20 Max. :9517**

**>**

* Exploratory Data Analysis. Prestige is the response variable.

1. Matrix of 2 by 2 scatterplots.



1. Correlation matrix in R, cor(Prestige[,numvar], use="pairwise.complete.obs"). Study correlations between numeric variables appearing in the work space

* Step( ) method in R, base on AIC (*Akaike information criteria*) can be used to assess the best model consistent to data. Interpret the resulting model and goodness of fit.

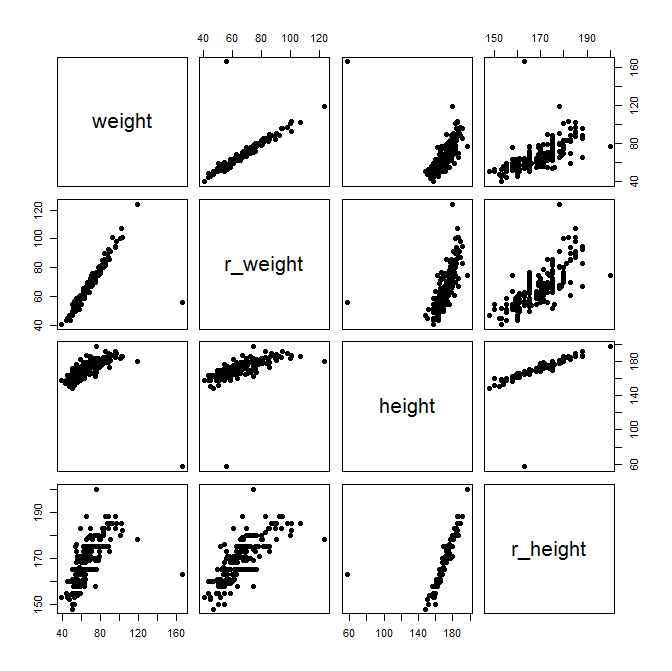
1. Study if the average of the response (prestige) depends on the levels of factor type. Use graphical and inferential tools.
2. Incorporate factor type in the multiple linear model developed so far considering interaction between numerical explicative variables and factor levels. Is it worth? (test statistical significance). Interpret the resulting model.
3. Check outliers in residuals and influent data in the selected model. Compute histogram of studentized residuals (rstudent(model)), leverage (hatvalues(model)) and Cook’s distance (cooks.distance(model)).
4. R2 and global regression test .
5. Residual analysis:

* Detection of o*utliers*.
* Scatterplot of studentitzed residual *vs.* .
* Scatterplot of studentitzed residual *vs.* *vs.* .
* Detection of *a priori* and *a posterior influent data*.
* Scatterplot of studentitzed residual *vs.* *leverage*.
* Scatterplot of studentitzed residual *vs.* Cook’s distance.
* residualPlots(model, layout=c(1, 2)) in library(car)
* influencePlot(model,id.n=5,id.method=abs(rstudent(model))) in library(car)
* marginalModelPlots(model,labels=row.names(df),id.method=abs(cooks.distance(model)), id.n=3) in library(car)

## Case 5: weight vs height in Davis

The Davis data frame has 200 rows and 5 columns. The subjects were men and women engaged in regular exercise. There are some missing data. This data frame contains the following columns:

* sex: A factor with levels: F, female; M, male.
* weight:Measured weight in kg.
* height: Measured height in cm.
* r\_weight : Reported weight in kg.
* r\_height : Reported height in cm.



Firstly, we examine the relationship between the reported weight and the actual weight in order to assess how data behaves. Pay attention to outliers.

Secondly, we focus on the classical relationship between weight (Y) and height (X): does a quadratic fit hold? Why?

### Suggested steps

* Correlation matrix in R, cor(Davis, use="pairwise.complete.obs" )

1. Matrix of 2 by 2 scatterplots.
2. Multiple regression weight (Y) vs r\_weight (Y). Interpret the regression equation and quality of the fit
3. Multiple regression weight (Y) vs height (X). Interpret the regression equation and quality of the fit
4. Multiple regression weight (Y) vs poly(height,2) (X). Can you Interpret the regression equation and quality of the fit?

# Case 6: Air Fares – This is your first assignment Course 21-22

The data set to perform this exercise belongs to the stata website (http://www.stata.com/texts/eacsap/), there are 4596 individual observations of US air fares documented by Jeffrey M. Wooldridge, 2000 "Airfare" Instructional Stata datasets for econometrics”, Boston College Department of Economics.

-----------------------------------------------------------------------------

storage display value

variable name type format label variable label

-----------------------------------------------------------------------------

year int %9.0g 1997, 1998, 1999, 2000

origin str21 %21s flight's origin

destin str24 %24s flight's destination

id int %9.0g route identifier

**dist int %9.0g distance, in miles**

**passen int %9.0g avg. passengers per day**

**fare int %9.0g avg. one-way fare, $**

**bmktshr** float %9.0g fraction market, biggest carrier

ldist float %9.0g log(distance)

y98 byte %9.0g =1 if year == 1998

y99 byte %9.0g =1 if year == 1999

y00 byte %9.0g =1 if year == 2000

lfare float %9.0g log(fare)

ldistsq float %9.0g ldist^2

concen float %9.0g = bmktshr

lpassen float %9.0g log(passen)

-----------------------------------------------------------------------------

Sorted by: id year

The file contains the price paid by passengers for internal flights in 48 US states between 1997 and 2000. The original response variable is fare. The key explanatory variable is the market share of the majority company (bmktshr or concen), the distance should also be included in the model and the daily demand (passen). **You only have to work in this exercise with the data for the year 2000.**

1. Determine if the response variable (airfare) has an acceptably normal distribution.
2. Value the descriptive statistics of the variables fare (fare), market share of the largest operator (bmktshr), distance (dist) and number of daily passengers (passen).
3. Indicate by exploration of the data which are apparently the variables most associated with the response variable (use only the indicated variables).
4. Define a polyeconomic factor f.concen for the covariate of the market share of the main operator (concen) according to its quartiles and argue if the average rate depends on the level of concentration of the path. Statistically justify the answer.
5. Calculate and interpret the anova model of a route that explains the rate according to the concentration factor of the main operator (concentration factor created in point 4).
6. Do you think that the variability of the fare depends on the degree of monopoly that is operated on the route?
7. Calculate the linear regression model that explains the fare from the distance of the path: interpret the regression line and assess its quality.
8. What is the percentage of the rate variability that is explained by the length of the air route?
9. Do you think it is necessary to introduce a quadratic term in the equation that relates the fare to the length of the air route?
10. After controlling for the length of the path, indicate whether the additive effect of the monopoly factor on the route is statistically significant.
11. Indicate whether the relationship between the air fare and the length of the route depends on the monopoly factor on the route? Calculate the models involved, make the appropriate hypothesis tests and interpret them statistically.
12. Select the best model available so far. Interpret the equations that relate the explanatory variables to the answer (rate).
13. Study the model that relates the logarithm of the rate to the linear and quadratic terms of the logarithm of distance.
14. Graphically assess the models obtained in Point 9 and in Point 13.

**A new logarithmic scale model of the fare is constructed with explanatory variables distance (linear) and number of passengers (both in logarithmic scale) and the monopoly dichotomous factor that indicates 1 if the market share of the main operator exceeds 80% . There are interactions in the proposed model. The results are presented below and on this model the questions of residue analysis and prediction will be developed.**

> summary(lfit.8)

Call:

lm(formula = lfare ~ (ldist) \* lpassen + f.concen1 \* lpassen +

(ldist) \* f.concen1, data = air)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7.48657 0.73393 10.201 < 2e-16 \*\*\*

ldist -0.30180 0.10856 -2.780 0.00552 \*\*

lpassen -0.79735 0.11546 -6.906 8.28e-12 \*\*\*

f.concen1>80% 1.14684 0.37247 3.079 0.00213 \*\*

ldist:lpassen 0.11136 0.01716 6.489 1.29e-10 \*\*\*

lpassen:f.concen1>80% -0.08684 0.03117 -2.787 0.00542 \*\*

ldist:f.concen1>80% -0.10155 0.04859 -2.090 0.03685 \*

Residual standard error: 0.3185 on 1142 degrees of freedom

Multiple R-squared: 0.4196, Adjusted R-squared: 0.4166

F-statistic: 137.6 on 6 and 1142 DF, p-value: < 2.2e-16

1. Assess the presence of outliers in the studentized residuals at a 99% confidence level. Indicate what those observations are.
2. Study the presence of *a priori* influential data observations, indicating their number according to the criteria studied in class.
3. Study the presence of *a posteriori* influential values, indicating the criteria studied in class and the actual atypical observations.
4. Given a flight of a distance of 300km, with a daily average of 1000 passengers / day and absence of monopolistic behavior in the service, what would be the expected fare with a 95% confidence interval?
5. Summarize what you have learned by working with this interesting real data game.