Practice 4: General Linear Model

### METHODOLOGY: ANOVA and ANCOVA

Multiple regression and general models are computed by method lm() in R:

****

1. Formulation and interpretation of 1 Way Anova (Y ~ A):



1. Formulation and interpretation TWO-WAY ANOVA model with interactions (Y ~ A\*B) :

# .

1. Formulation and interpretation ANCOVA model with interactions (Y ~ A\*X) :



Preliminary univariate and bivariate exploratory analysis:

• Histogram / Boxplot and numerical statistics of the variables for each level of the categorical variable

Linear model estimation:

• Coefficients and expression of the model obtained, from the corresponding dummy variables for the categorical

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

• Inference about the parameters of the model: Meaning of the coefficients

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)

ANOVA table of factors

• Significance of the factor within the model

• For meaningful interactions: plot of interactions for interpretation

TRANSFORMATIONS

We propose Box-Cox type transformations (logarithm, square root, square, inverse, etc.…) when…

• ... the relationship is clearly non-linear between the answer and some explanatory variable

• ... there is an increase / decrease in variability as the level of explanatory variables increases

• ... there are values ​​with a very high leverage for some explanatory variable

## Case 0. 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**

**>**

* Create a new binary factor (women) indicating if there are mostly women professions (women percentage greater than 50) (Factor B).
* Does average prestige depend on factor type? Use inferential test presented in topic ‘Computational Statistical Inference’. Pay attention to parametric/nonparametric test requeriment.
* Does average prestige depend on factor f.femenin? Use inferential test. Pay attention to parametric/nonparametric test requeriment.
* Does prestige variability depend on factor type? Use inferential test presented in topic ‘Computational Statistical Inference’. Pay attention to parametric/nonparametric test requeriment.
* Does prestige variability depend on factor f.femenin? Use inferential test. Pay attention to parametric/nonparametric test requeriment.
* Use model building by standard multiple regression: lm(.) in R. Interpret model estimates and compute the prediction for prestige in the groups defined by type.
* Build a general linear model using lm() for prestige considering factors type and new binary factor femenin. Consider interaction model (A\*B), additive model (A+B) and oneway models (A or B). Select the best model using inferential results based on testing nested models with anova(model1,model2) method in R.
* Nul model: 
* OneWay ANOVA for factor A: 
* OneWay ANOVA for factor B: 
* TwoWay ANOVA additive model for factors A and B: 
* TwoWay ANOVA interactive model for factors A and B:: 
* Consider the family of ANCOVA models concerning explicative variables education, income, women and factor type. Select the best model using inferential results based on testing nested models with anova(model1,model2) method in R.
* Nul model: 
* OneWay ANOVA for factor A: 
* Simple linear regression for X: 
* Additive ANCOVA MODEL: 
* Interaction ANCOVA model : 
* Validate the best available model.

# Case 1. InsectSprays

**Nom: InsectSprays (Efectivitat d’insecticides)**

**Reference:** Beall, G., (1942) The Transformation of data from entomological field experiments, *Biometrika*, 29, 243–262

**Description:** Count of insects in experimental units treated with different insecticides**.**

**Number of observations:** 72

**Variables:** count = Number of observations

spray = Insecticide type

count spray

10 A

7 A

20 A

14 A

14 A

12 A

10 A

23 A

17 A

20 A

14 A

13 A

11 B

17 B

21 B

11 B

16 B

14 B

17 B

17 B

19 B

21 B

7 B

13 B

0 C

1 C

7 C

2 C

3 C

1 C

2 C

1 C

3 C

0 C

1 C

4 C

3 D

5 D

12 D

6 D

4 D

3 D

5 D

5 D

5 D

5 D

2 D

4 D

3 E

5 E

3 E

5 E

3 E

6 E

1 E

1 E

3 E

2 E

6 E

4 E

11 F

9 F

15 F

22 F

15 F

16 F

13 F

10 F

26 F

26 F

24 F

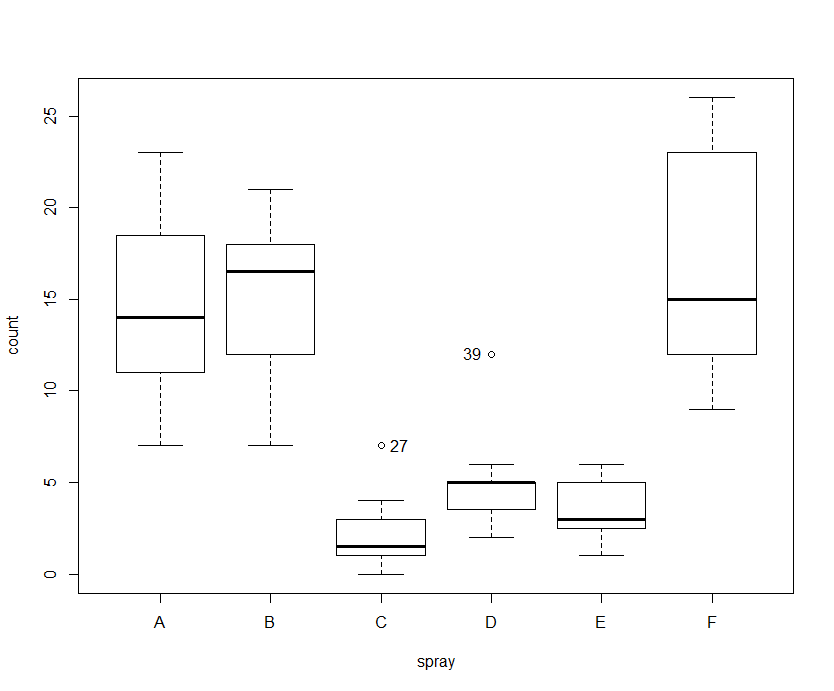
13 F

# Exercise 1: InsectSprays: Are there any significant difference among insecticides?

1. If the ANOVA technique is used directly to compare the number of individuals left by each insecticide, what objections can be made to this technique? Are there any premises that are not met?

2. Consider a monotonous transformation of the response variable that involves an improvement in the validation of the method premises. Apply the ANOVA of a factor for the transformed variable.

3. Clearly represent the associated linear model and estimate its coefficients, considering two possible contrasts (definition of the dummy variables associated with the factor): contr.treatment and contr.sum. What are the differences between the two estimated models?



# Case 2. Machines

**Nom: Machines (package nlme)**

**Referència:** Milliken, G. A. and Johnson, D. E. (1992), *Analysis of Messy Data, Volume I: Designed Experiments*, Chapman and Hall, London

**Description:** Data from an experiment to compare three types of machines used in an industrial process. We also want to analyze the performance of six workers who operate each machine three times. The answer is the overall productivity indicator for each run

**Number of observations:** 54

**Variables:**

1. Worker: Worker ID
2. Machine: Machine type
3. score: Productivity score

Worker Machine score

1 A 52.0

1 A 52.8

1 A 53.1

2 A 51.8

2 A 52.8

2 A 53.1

3 A 60.0

3 A 60.2

3 A 58.4

4 A 51.1

4 A 52.3

4 A 50.3

5 A 50.9

5 A 51.8

5 A 51.4

6 A 46.4

6 A 44.8

6 A 49.2

1 B 62.1

1 B 62.6

1 B 64.0

2 B 59.7

2 B 60.0

2 B 59.0

3 B 68.6

3 B 65.8

3 B 69.7

4 B 63.2

4 B 62.8

4 B 62.2

5 B 64.8

5 B 65.0

5 B 65.4

6 B 43.7

6 B 44.2

6 B 43.0

1 C 67.5

1 C 67.2

1 C 66.9

2 C 61.5

2 C 61.7

2 C 62.3

3 C 70.8

3 C 70.6

3 C 71.0

4 C 64.1

4 C 66.2

4 C 64.0

5 C 72.1

5 C 72.0

5 C 71.1

6 C 62.0

6 C 61.4

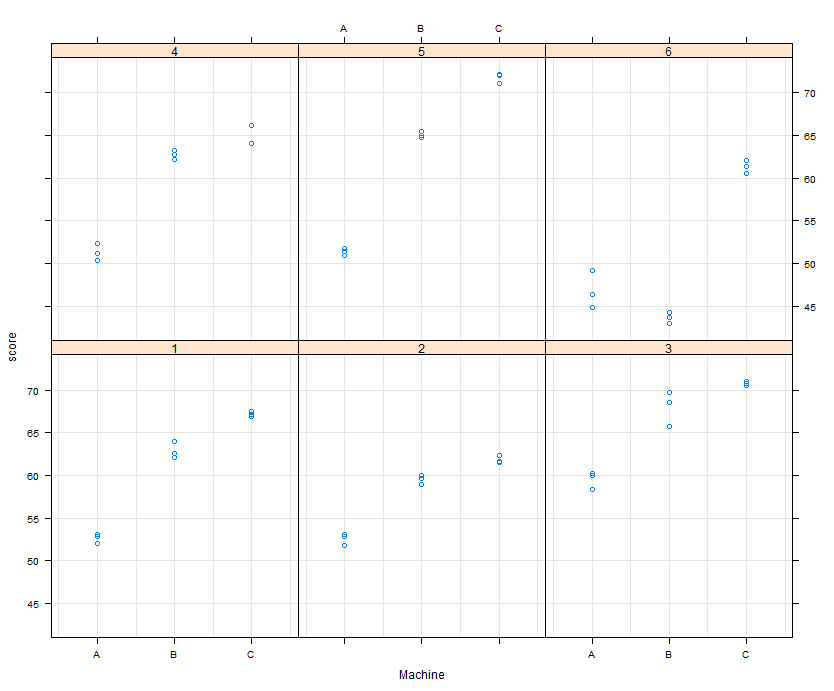
6 C 60.5

# Exercise 2: Machines. Which are the factors affecting high productivity ?

1. Construct the model to analyze the two cross-factors without interaction, and solve the ANOVA table model to determine if there are differences between operators and between machines.

2. Now consider the model with interaction and determine if the interaction is significant. Interpret the corresponding interaction graph.

3. For each model obtained, validate the premises and try to formulate it alternately by changing the contrasts.



# Case 3. BirthWeight

**Name: BirthWeight**

**Reference:** Hosmer, D.W. and Lemeshow, S. (1989) Applied Logistic Regression. New York: Wiley

**Description:** Subset of data on the weight of babies at birth and two factors to analyze: the mother's race and whether or not she smoked during pregnancy. The original dataset contains many more variables. In this case, two categories have been selected.

**Cases:** 189

**Variables:**

bwt: Weight at birth in grams

smoke: Was the mother smoking during pregnancy (0=no, 1 =yes)

race: Mother’s race (0=White, 1= Afroamerican, 2=Other)

bwt smoke race bwt smoke race bwt smoke race

2523 0 2 3234 0 1 4174 0 1

2551 0 3 3260 1 3 4238 1 1

2557 1 1 3274 0 3 4593 0 1

2594 1 1 3274 0 3 4990 0 1

2600 1 1 3303 1 3 709 1 3

2622 0 3 3317 0 1 1021 0 1

2637 0 1 3317 0 2 1135 1 2

2637 0 3 3317 1 1 1330 0 3

2663 1 1 3321 1 3 1474 0 3

2665 1 1 3331 1 3 1588 0 3

2722 0 3 3374 0 2 1588 0 3

2733 0 3 3374 1 1 1701 0 2

2751 0 3 3402 0 2 1729 0 3

2750 0 3 3416 0 1 1790 1 1

2769 1 1 3430 1 1 1818 1 1

2769 1 1 3444 1 2 1885 1 1

2778 0 2 3459 0 1 1893 0 3

2782 1 1 3460 0 1 1899 0 3

2807 0 3 3473 0 1 1928 1 1

2821 1 1 3544 0 3 1928 1 1

2835 0 3 3487 0 3 1928 0 2

2835 0 1 3544 0 3 1936 1 1

2836 0 1 3572 1 3 1970 0 3

2863 0 3 3572 0 3 2055 0 3

2877 0 3 3586 0 1 2055 0 3

2877 0 1 3600 0 1 2082 0 1

2906 1 1 3614 0 1 2084 1 1

2920 0 1 3614 0 1 2084 1 1

2920 1 2 3629 0 3 2100 0 1

2920 0 2 3629 1 1 2125 1 3

2920 0 2 3637 1 1 2126 1 2

2948 1 1 3643 1 1 2187 1 1

2948 1 2 3651 0 1 2187 0 2

2977 0 1 3651 0 1 2211 1 3

2977 0 2 3651 1 1 2225 1 1

2977 1 1 3651 1 1 2240 0 3

2977 1 1 3699 0 1 2240 0 3

2922 1 1 3728 0 1 2282 0 3

3005 1 1 3756 1 1 2296 1 2

3033 1 1 3770 0 3 2296 1 1

3042 1 2 3770 0 1 2301 0 3

3062 0 1 3770 0 3 2325 0 3

3062 0 2 3790 0 2 2353 1 1

3062 0 1 3799 0 1 2353 0 1

3062 1 1 3827 0 1 2367 1 2

3062 1 1 3856 1 1 2381 1 2

3080 0 1 3860 0 2 2381 1 2

3090 0 3 3860 0 1 2381 0 3

3090 0 1 3884 0 3 2410 0 3

3090 1 3 3884 1 1 2410 1 1

3100 0 1 3912 0 1 2410 1 1

3104 0 3 3940 1 1 2414 1 1

3132 1 1 3941 0 3 2424 1 1

3147 1 1 3941 0 1 2438 0 2

3175 0 3 3969 0 3 2442 0 3

3175 0 3 3983 0 1 2450 0 3

3203 1 3 3997 0 3 2466 1 1

3203 0 3 3997 0 1 2466 1 3

3203 0 3 4054 0 3 2466 1 1

3225 0 3 4054 0 1 2495 0 3

3225 0 3 4111 0 1 2495 1 3

3232 0 3 4153 0 1 2495 0 2

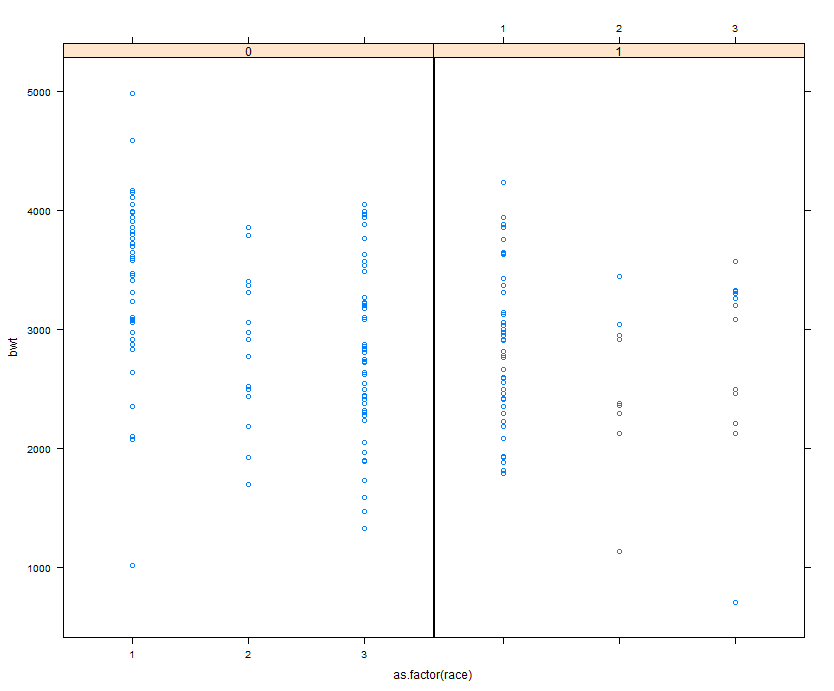
3232 0 3 4167 0 1 2495 1 1

# Exercise 3: BirthWeight. Are race and mother smoking status affecting weight at birth?

1. Analyze whether the relationship between the proposed factors and the response variable is significant. Initially, consider the model with interaction. Is the interaction significant? Analyze the corresponding plot

2. Obtain the design matrix of the linear model associated with treatment-type contrasts for factor variables. Interpret the coefficients obtained

3. Validate the model premises. What objections can be made in these analyzes?



# Cas Pràctic 1. Oftalmic

**Nom: Oftàlmic (Index de capacitat visual)**

# Referència: Perros, P. (1993), Age and gender influence the severity of thyroid-associated ophthalmopathy: a study of 101 patients attending a combined thyroid-eye clinic. *Clinical Endocrinology*, 38, 367-372

**Description:** Visual ability indicator in adults on the first visit (Large values ​​associated with loss of visual ability). Data on the age and sex of patients are added

**Number of cases: 101**

**Variates:** OI = Index Oftàlmic

age = Edat del pacient

Sex= Sexe del pacient

OI age Sex

5 65 2

3,75 40 2

7,6 52 2

2,45 45 2

5,4 72 2

10,87 64 2

6,15 67 2

5,15 66 2

2,15 47 2

2,45 44 2

5,95 46 2

9,95 52 2

6,6 55 2

9,35 64 2

5,15 62 2

7,1 37 2

8,45 39 2

2,35 45 2

2,65 39 2

3,35 54 2

5,8 37 2

0,98 34 2

9,8 48 2

6,25 66 2

6,7 39 2

5,9 47 2

4,75 39 2

7,7 48 2

6,85 57 2

4,85 55 2

6,75 61 2

8,35 29 2

3,75 27 2

12,75 62 2

2,65 37 2

8,1 63 2

6,8 58 2

6,75 49 2

6 29 2

3,05 15 2

4,75 35 2

3,35 60 2

4,05 37 2

4,05 26 2

6,1 60 2

6,4 40 2

1,85 31 2

4,05 41 2

3,75 30 2

4,25 26 2

13,95 56 2

6,9 46 2

5,25 65 2

3,55 26 2

4,95 52 2

2,95 65 2

4,8 64 2

7,55 61 2

9,4 47 2

2,95 52 2

4,3 54 2

2,75 14 2

9 54 2

3,55 25 2

9,85 58 2

5,1 52 2

6,4 45 2

6 48 2

3,35 46 2

14,55 73 2

3,4 43 2

4,95 56 2

2,95 61 2

3 54 2

3,55 62 2

4,55 55 2

3,05 59 2

9,95 54 2

4,65 26 2

2,8 34 2

6,05 56 2

9,55 69 1

5,45 65 1

2,95 25 1

8,4 53 1

7,75 39 1

11,55 67 1

8,4 53 1

7,9 64 1

2,25 48 1

8,1 62 1

5,95 24 1

5,65 51 1

9,65 46 1

7,6 56 1

9,65 67 1

8,9 57 1

16,55 56 1

4,65 53 1

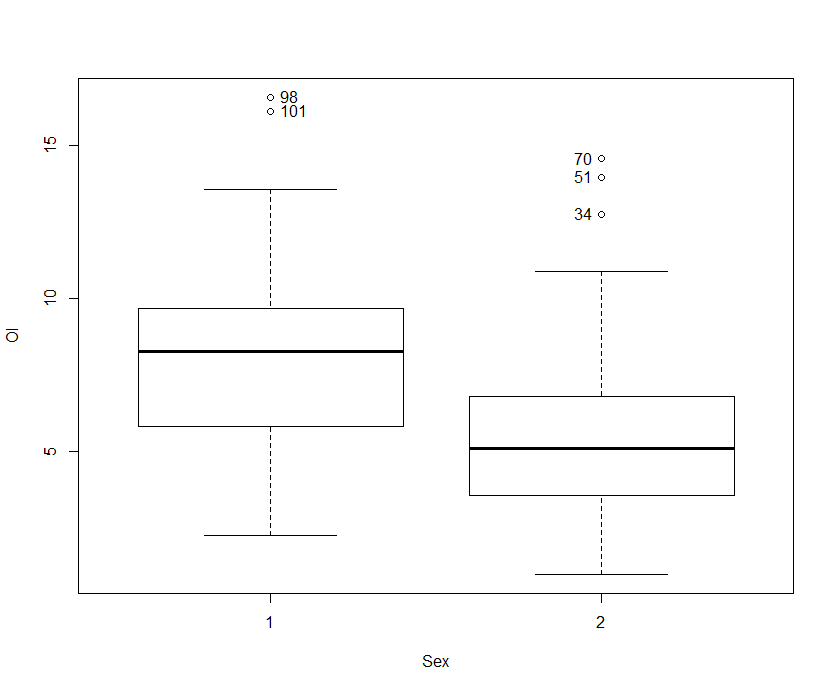
13,55 56 1

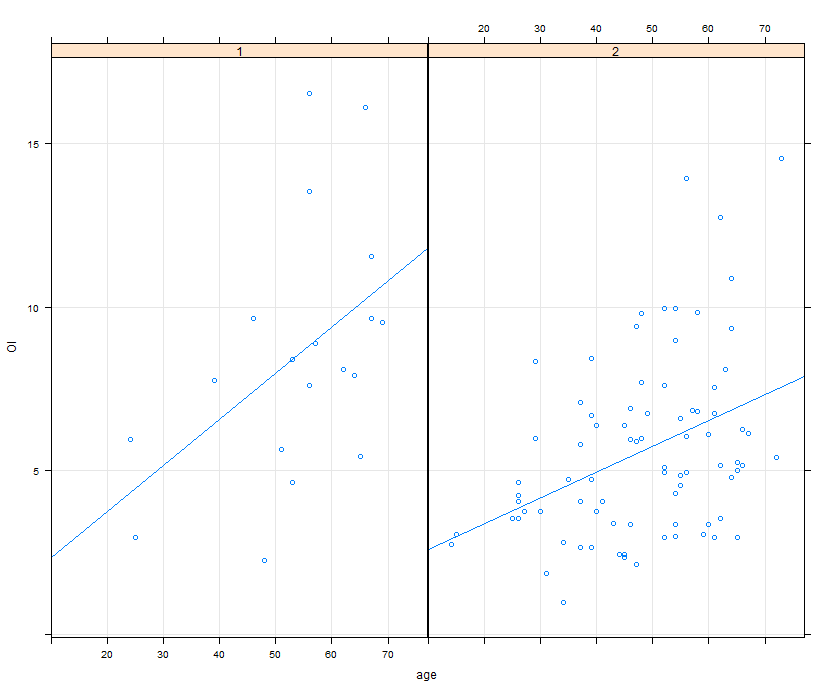
16,1 66 1

# Exercise 4: Oftalmic. Are there any gender differences? Which?

1. We can apply the ANOVA technique to compare the value of the index between the two sexes. If we ignore the continuous covariant information, the test will allow us to compare the response between the different levels of the factor as if the observations were random. What conclusion would we reach?

2. If we relate the response to the continuous covariant age, it is observed that there may be a linear relationship between both variables. Can it be convenient to perform some transformation to improve the model?

3. When looking at the influence of the factor (Sex) on the response (OI) once adjusted by the covariant (age), we can propose different scenarios: a) common linear model, b) parallel lines for both groups, c) divergent lines with the same origin, and d) totally divergent lines. Adjust each of the models and compare them to decide what the appropriate situation is in this case. Validate and interpret the final model. Apply the ANCOVA technique to confirm the result obtained.



# Case 5. WWheat

**Name: WWheat (SASmixed package)**

**Reference:** Littel, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. (1996), SAS System for Mixed Models, SAS Institute (Data Set 7.2).

**Description:** We select 10 varieties of a type of cereal adapted to extremely dry climatic conditions (Winter Wheat) that are planted randomly in 6 plots each more or less homogeneous variety (with a total of 60 plots). Moisture from each plot is also collected, as it is thought to have affected the degree of germination. The answer is the production obtained

**Number of cases: 60**

**Variables:**

* Variety: Variety considered
* Yield: Total production (faneges per acre)
* Moisture: Degree of humidity

Variety Yield Moisture

1 41 10

1 69 57

1 53 32

1 66 52

1 64 47

1 64 48

2 49 30

2 44 21

2 44 20

2 46 26

2 57 44

2 42 19

3 69 50

3 62 40

3 50 23

3 76 58

3 48 21

3 55 30

4 48 22

4 60 40

4 45 17

4 47 21

4 62 44

4 43 13

5 65 49

5 63 44

5 71 57

5 68 51

5 52 27

5 68 52

6 76 55

6 46 11

6 45 11

6 67 43

6 65 38

6 79 60

7 35 17

7 37 20

7 30 11

7 30 10

7 57 48

7 49 36

8 75 57

8 64 41

8 46 15

8 54 28

8 52 23

8 52 23

9 51 26

9 63 44

9 42 13

9 61 40

9 67 48

9 69 53

10 60 37

10 73 58

10 66 44

10 71 53

10 67 48

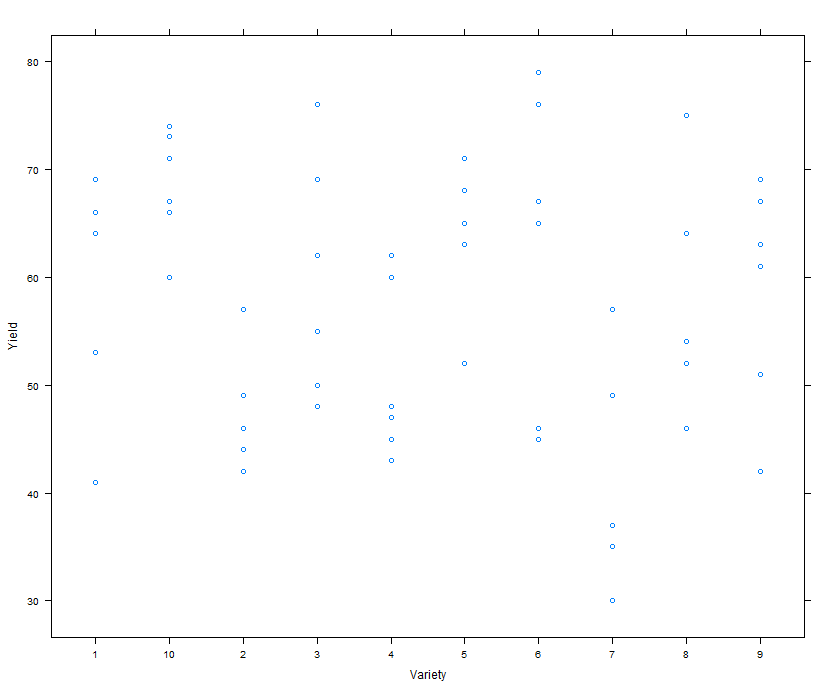
10 74 59

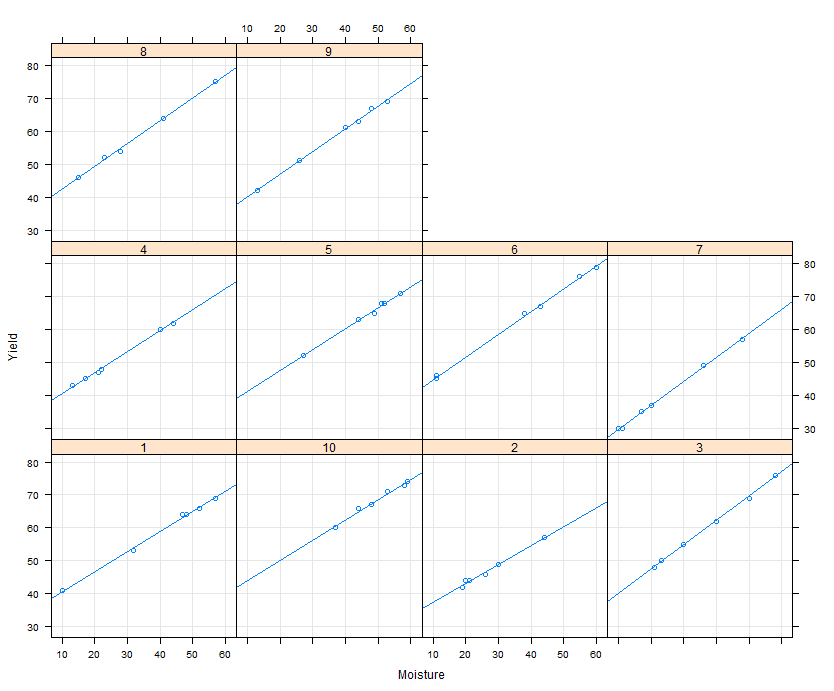
# Exercise 5: WWheat. Are there any differences in varieties once humidity is considered?

1. Construct the model to analyze whether there are differences between varieties. What is the conclusion? Is the model correct?

2. See the effect of the covariant on the response, at each factor level. Is a linear relationship between humidity and production permissible?

3. Apply the ANCOVA technique to determine whether the effect of moisture on production depends additively or interactively (on the slope) depending on the variety.





# Case 6. PIBsp

**Name: GDPsp** (GDP in Spain at the provincial level in 2008)

**Reference:** National Statistics Institute (www.ine.es)

**Description:** Demographic and Economic Data by Province in 2008

**Number of cases: 52**

**Variables:**

* Province: Province identifier
* GDP: Accumulated GDP (year 2008)
* Town: Population
* Area: Area (km2)
* Unemployment: Unemployment rate
* SAU: Useful Agricultural Area (%)
* Tourists: Number of tourists per year
* Immigrants: Percentage of immigrant population
* Birth rate: Birth rate (per 1000 inhabitants)
* Companies: Number of companies
* Mortgages: Number of mortgages granted on urban and rural properties
* Coast: Province located on the coast (0 = n0, 1 = yes)
* Border: Province bordering another country (0 = n0, 1 = yes)

Provincia PIB Pob Sup Atur SAU Turistes Imm. Nat. Empresas Hipotecas Costa Frontera

A Coruña 23891 1145488 7950 11,6 1906,19 1745491 3,3 8,76 82729 15259 1 0

Álava 10669 313819 3038 11,3 802,01 334761 8,3 10,45 21675 7330 0 0

Albacete 6707 400891 14926 20 6001,34 334950 8,2 10,36 27256 4742 0 0

Alicante 37501 1917012 5817 22,4 1501,35 3450368 24,2 10,82 132814 27140 1 0

Almería 14489 684426 8775 26,8 2000,03 1175254 21 13,58 42546 16611 1 0

Asturias 23497 1085289 10604 13,4 2637,39 1441377 4,3 7,76 70362 12084 1 0

Ávila 3423 171680 8050 18,5 3421,79 289106 7,2 9,01 11414 2986 0 0

Badajoz 11014 688777 21714 21,4 11078,45 561574 3,3 10,67 39739 8955 0 1

Barcelona 151555 5487935 7728 16,2 1360,41 9473981 14,6 12,1 458918 61861 1 0

Burgos 10007 375563 14291 12,7 6211,31 712477 9,2 9,66 25397 6250 0 0

Cáceres 6803 413633 19868 19 9358,54 595575 3,5 8,74 25834 5584 0 1

Cádiz 22835 1230594 7436 26,9 2292,1 2218720 3,7 12,03 61056 18510 1 0

Cantabria 13867 589235 5321 12 1171,99 1063145 6,5 10,26 39024 8638 1 0

Castellón 14771 602301 6632 20 1647,82 979155 18,5 12,33 40797 10028 1 0

Ceuta 1709 78674 19 18,9 0,03 81035 4,5 16,1 3615 768 1 1

Ciudad R. 9482 527273 19813 19,9 8863,21 394222 8,5 10,97 31627 5220 0 0

Córdoba 13114 803998 13771 26,1 6587,89 878939 3,1 11,06 48249 10641 0 0

Cuenca 3952 217363 17141 15,9 7510,98 316393 12,5 8,73 14236 1917 0 0

Girona 20655 747782 5910 18,2 1063,41 3246409 21,4 12,98 58240 13063 1 1

Granada 15370 907428 12647 26,3 4871,22 2132366 6,8 11,73 58319 12020 1 0

Guadalaj. 4799 246151 12212 13,5 3068,06 241289 15,4 13,6 13494 4399 0 0

Guipúzcoa 22230 705698 1980 9 489,46 900471 5,8 10,46 60580 9448 1 1

Huelva 9800 513403 10128 22,6 2406,65 1088706 7,8 12,29 26277 9050 1 1

Huesca 5456 228409 15636 10,6 4691,13 697744 12 9,67 16550 3669 0 1

Il.Balears 27846 1095426 4992 18 1674,24 7555293 21,7 12,01 89562 19044 1 0

Jaén 10220 669782 13496 23,2 6379,8 485660 2,9 10,79 35758 7416 0 0

La Rioja 8045 321702 5045 12,8 1217,14 507458 14,6 11,21 23190 6791 0 0

Las Palmas 23359 1083502 4066 28 227,84 4020934 14 10,12 70889 14408 1 0

León 10616 500169 15581 14,6 3930,96 750223 5 7,26 33126 7064 0 0

Lleida 11950 436402 12172 11,4 3513,52 786200 17,8 12,25 36598 6456 0 1

Lugo 6636 355195 9856 9,3 2705,39 590852 3,8 6,37 24714 4700 1 0

Madrid 195197 6386932 8028 14 2431,62 9812454 16,7 12,62 503501 80008 0 0

Málaga 28927 1593068 7308 26,4 2803,57 4066685 16,8 12,28 110291 24181 1 0

Melilla 1550 73460 13 24,2 0,2 47669 10,3 18,05 3655 859 1 1

Murcia 28064 1446520 11313 20,7 3531,13 1038209 16,3 13,55 92196 24059 1 0

Navarra 18618 630578 10390 10,9 3650,22 726604 11,2 11,52 42347 7535 0 1

Ourense 5721 335642 7273 10,3 950,63 294892 4,6 6,57 23100 2699 0 1

Palencia 4035 173306 8053 14,1 4057,92 214153 4 7,48 10825 1930 0 0

Pontevedra 19205 959764 4495 15,5 742,57 1167410 4,1 9,55 68331 14551 1 1

Salamanca 6880 354608 12350 15,5 7728,55 872200 4,9 8,34 23376 5954 0 1

Segovia 3955 164854 6923 12,2 2976,64 385071 13 9,5 11597 2546 0 0

Sevilla 35886 1900224 14036 23,6 5469,75 2350553 3,8 13,69 116083 30380 0 0

Soria 2100 95101 10307 10,4 3275,51 171322 10 8,3 5885 1476 0 0

Tarragona 21595 803301 6303 17 2051,11 2273130 18,6 12,84 55914 12122 1 0

Tenerife 20225 1020490 3381 24,2 266,1 3107808 14,7 9,93 65065 12448 1 0

Teruel 3641 146751 14810 10,3 4595,33 324605 12,5 8,53 9388 1530 0 0

Toledo 12996 689635 15370 20,1 8025,88 686915 12,1 12,82 45223 10290 0 0

Valencia 56117 2575362 10806 20,7 2871,61 2581368 12,2 11,87 178755 35616 1 0

Valladolid 13405 532575 8110 12,4 5043,24 601581 6,1 9,7 35254 9821 0 0

Vizcaya 33510 1152658 2217 12,3 537,91 1006307 5,7 9,53 87527 17201 1 0

Zamora 3673 195665 10561 14,1 4619,14 234294 4,1 6,25 12098 2474 0 1

Zaragoza 25274 970313 17274 13,7 6970,72 1147583 13 10,95 66267 14092 0 0

# Case 6: PIB. Which are factors associated to high GP (gross product)?

1. Obtain the appropriate bivariate representation to analyze the relationships between the numerical variables. Do you consider the linear relationships to be clear? Analyze whether a clear transformation results in clearer linear relationships

2. Adjust a linear model to predict GDP based on other factors. Validate the model and analyze if multicollinearity exists

3. Would you propose an alternative analysis, given that both the random variable and some of the explanatory variables are highly correlated with population?

# Case 7. CPS1985

Determinants of Wages Data (CPS 1985). Cross-section data originating from the May 1985 Current Population Survey by the US Census Bureau (random sample drawn for Berndt 1991). A data frame containing 534 observations on 11 variables.

**Usage:** data("CPS1985") after loading AER Package.

|  |  |
| --- | --- |
| Variable | Description |
| wage | Wage (in dollars per hour). |
| education | Number of years of education. |
| experience | Number of years of potential work experience (age - education - 6). |
| age | Age in years. |
| ethnicity | Factor with levels "cauc", "hispanic", "other". |
| region | Factor. Does the individual live in the South? |
| gender | Factor indicating gender. |
| occupation | Factor with levels "worker" (tradesperson or assembly line worker), "technical" (technical or professional worker), "services" (service worker), "office" (office and clerical worker), "sales" (sales worker), "management" (management and administration). |
| sector | Factor with levels "manufacturing" (manufacturing or mining), "construction", "other". |
| union | Factor. Does the individual work on a union job? |
| married | Factor. Is the individual married? |

**Source:** StatLib.

http://lib.stat.cmu.edu/datasets/CPS\_85\_Wages

**References**

Berndt, E.R. (1991). The Practice of Econometrics. New York: Addison-Wesley.

Address a model exercise for wage target. Outline for today’s exercise is the one for your first assignment.

* Create factors for qualitative variables.
* Determine if the response variable (wage) has an acceptably normal distribution.
* Address tests to discard serial correlation.
* Detect univariant and multivariant outliers, errors and missing values (if any) and apply an imputation technique if needed.
* Preliminary exploratory analysis to describe the relationships observed has to be undertaken.
* If you can improve linear relations or limit the effect of influential data, you must consider the suitable transformations for variables.
* Apart from the original factor variables, you can consider other categorical variables that can be defined from categorized numeric variables.
* You must take into account possible interactions between categorical and numerical variables.
* When building the model, you should study the presence of multicollinearity and try to reduce their impact on the model for easier interpretation.
* You should build the model using a technique for selecting variables (removing no significant predictors and/or stepwise selection of the best models).
* The validation of the model has to be done with graphs and / or suitable tests to verify model assumptions.
* You must include the study of unusual and / or influential data.
* The resulting model should be interpreted in terms of the relationships of selected predictors and its effect on the response variable.