## General Linear Model: how to include factors in Multiple Regression

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) :



### 

### Duncan1 data on prestige of professions (42 observations)

**Explicative variables are income and education. Response variables is prestige.**

1. Matrix of 2 by 2 scatterplots.
2. Correlation matrix in R, cor(duncan1, use="pairwise.complete.obs"). Study correlations between numeric variables appearing in the work space

Be careful with the default order of factor levels :.

* + Reorder to simplify interpretation: factor(variable, levels=c(nivell1, …, nivellsk))
  + If factor levels are not meaningful include labels for factor levels: factor(variable, levels=c(nivell1, …, nivellsk),labels=c(nom1,…,nomk)).

Perfect collineality appears in design matrices for general linear models and reparametrization is mandatory, being baseline for the first level the default set in R:

Base-line: options(contrasts=c("contr.treatment","contr.treatment"))

Suma zero: options(contrasts=c("contr.sum","contr.sum"))

ANOVA (and ANCOVA models also) can be computed, interpreted and tested using method lm() of estimation by ordinary least squares.

* + lm(*formula, dataframe*):.
  + Contrast for 2 **nested** models computed by lm(): anova(restrictedmodel, fullmodel).

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

1. Study if the average of the response (prestige) depends on the levels of factor type. Use graphical and inferential tools.
2. Incorporate additively factor type in the multiple linear model developed so far. Is it worth? (test statistical significance). Interpret the resulting model.
3. 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). Interprete the resulting model.
4. 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)).
5. R2 and global regression test .
6. 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.

## 

## Exercise 2: Davis data

The objective is testing whether or not statistical differences in means for Weight and Height according to gender groups are present in the current data set:

* Exploratory Data Analysis.
* Use t.test and var.test :

1. Test equal dispersion hypothesis in Weight/Height according to Male/Female defined groups.
2. Test equal mean hypothesis in Weight/Height according to Male/Female defined groups.

* Use oneway.test and bartlett.test :

1. Test equal dispersion hypothesis in Weight/Height according to Male/Female defined groups.
2. Test equal mean hypothesis in Weight/Height according to Male/Female defined groups.

* Use nonparametric methods for testing :

1. Test equal dispersion hypothesis in Weight/Height according to Male/Female defined groups.
2. Test equal mean hypothesis in Weight/Height according to Male/Female defined groups.

* Use model building by standard multiple regression: lm(.) in R. Interpret model estimates and compute prediction for Weight/Height in groups defined by Gender.
* WEIGHT( ) = --------------------------------------------------------------------------------------------------
* WEIGHT ( ) = ------------------------------------------------------------------------------------------------
* HEIGHT ( ) = ---------------------------------------------------------------------------------------
* HEIGHT ( ) = ----------------------------------------------------------------------------------------
* Use model building by standard multiple regression: lm(.) in R. Build a model for Weight depending on Height and Gender. Use an additive model

( ) = ------------------------------------------------------------------------------------------------------

( ) = ------------------------------------------------------------------------------------------------------

* Use model building by standard multiple regression: lm(.) in R. Build a model for Weight depending on Height and Gender. Use an interaction model

( ) = ------------------------------------------------------------------------------------------------------

( ) = ------------------------------------------------------------------------------------------------------

* **Which is the best model? Additive or interaction ANCOVA modeling ? Use standard incremental variance in R (anova() for comparing nested models)**
* Diagnostics and Validation
* Histogram of (studentitzed) residuals. Normal P-P plot. Check for *outliers*.
* Scatterplot of (studentitzed) residuals *vs.* Fitted.values.

|  |  |
| --- | --- |
|  |  |

***WEIGHT vs GENDER Residuals vs. Fit according to GENDER***

|  |  |
| --- | --- |
|  |  |

***HEIGHT vs GENDER Residuals vs. Fit according to GENDER***

## Exercise 2: 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.
* Does prestige depend on factor type? Use inferential test presented in topic ‘Computational Statistical Inference’. 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.
* Create a new binary factor (femenin) indicating if there are mostly women professions (women percentage greater than 50%) (Factor B). 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.

1. 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)).
2. R2 and global regression test .
3. 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.

## Exercise 3: Moore and Krupat (1.971) – ANOVA TWO-WAY

The objective is testing whether or not statistical differences in means for Conformity (personal aim of satisfaction) according to STATUS and AUTHORITARISM groups are present in the current data set:

* Descriptive statistics.
* Use t.test and var.test :

1. Test equal dispersion hypothesis in CONFORMITY according to STATUS, AUTHORITARISM defined groups.
2. Test equal mean hypothesis in CONFORMITY according to STATUS, AUTHORITARISM defined groups..

* Use oneway.test and bartlett.test :

1. Test equal dispersion hypothesis in CONFORMITY according to STATUS, AUTHORITARISM defined groups.
2. Test equal mean hypothesis in CONFORMITY according to STATUS, AUTHORITARISM defined groups..

* Use nonparametric methods for testing :

1. Test equal dispersion hypothesis in in CONFORMITY according to STATUS, AUTHORITARISM defined groups..
2. Test equal means in CONFORMITY according to STATUS, AUTHORITARISM defined groups..

* Use model building by standard multiple regression in 1 way Anova models: lm(.) in R

1. Interpret model estimates and compute prediction for CONFORMITY according to STATUS, AUTHORITARISM defined groups..

CONFORMITY vs STATUS ( ) = ---------------------------------------------------------------------------

CONFORMITY vs STATUS ( ) = -----------------------------------------------------------------------

CONFORMITY vs AUTHORITARISM ( ) = --------------------------------------------------------

CONFORMITY vs AUTHORITARISM ( ) = --------------------------------------------------------

CONFORMITY vs AUTHORITARISM ( ) = --------------------------------------------------------

1. Test STATUS, AUTHORITARISM significance to explain mean CONFORMITY.

.

* Use model building by standard multiple regression in 2 way Anova models: lm(.) in R

1. Nul model: 
2. OneWay ANOVA for factor A: 
3. OneWay ANOVA for factor B: 
4. TwoWay ANOVA additive model for factors A and B: 
5. TwoWay ANOVA interactive model for factors A and B:: 

* Test significant gross/net effects and interactions for factors STATUS, AUTHORITARISM in cross-defined groups for CONFORMITY. Hint***: Compute models lm() and test multiple hypothesis with Fisher test implemented in R in anova()..***
* Interpret the additive 2 way anova model to compute the mean predicted values for CONFORMITY according joint contribution of STATUS, AUTHORITARISM factors

( ) = -----------------------------------------------------------------------------------------------------

( ) = -----------------------------------------------------------------------------------------------------

( ) = -----------------------------------------------------------------------------------------------------

( ) = -----------------------------------------------------------------------------------------------------

( ) = ------------------------------------------------------------------------------------------------------

( ) = ----------------------------------------------------------------------------------------------------

* Interpret the full (with interactions) 2 way anova model to compute the mean predicted values for CONFORMITY according joint contribution of STATUS, AUTHORITARISM factors

( ) = -----------------------------------------------------------------------------------------------------

( ) = -----------------------------------------------------------------------------------------------------

( ) = -----------------------------------------------------------------------------------------------------

( ) = -----------------------------------------------------------------------------------------------------

( ) = ------------------------------------------------------------------------------------------------------

( ) = ----------------------------------------------------------------------------------------------------

* Use model building by standard multiple regression for all possible ANCOVA models: lm(.) in R
* Test significant gross/net effects STATUS, Numeric AUTHORITARISM in ANCOVA models CONFORMITY. Hint***: Compute models lm() and test multiple hypothesis with Fisher test implemented in R in anova()..***
* Interpret the additive ancova model to compute the predicted values for CONFORMITY according joint contribution of STATUS, Numeric\_AUTHORITARISM covariate.

( ) = -----------------------------------------------------------------------------------------------------

( ) = -----------------------------------------------------------------------------------------------------

* Interpret the full (with interactions) ancova model to compute the predicted values for CONFORMITY according joint contribution of STATUS, Numeric\_AUTHORITARISM covariate.

( ) = -----------------------------------------------------------------------------------------------------

( ) = -----------------------------------------------------------------------------------------------------

**Comands in R … suggestions …**

**moore <- read.table("Moore.txt",header=TRUE, sep="\t", na.strings="?")**

**save.image("W:/seccio\_fme/teaching/mlgz/dades/Moore.RData")**

**# Sessio 7: No oblideu la Descriptiva**

**summary(moore)**

**options()**

**options(contrasts=c("contr.treatment","contr.treatment"))**

**par(mfrow=c(2,2))**

**attach(moore)**

**# Podria semblar adequada una transformació logarítmica: s'obvia.**

**plot.design(conformity ~ status\*authoritarism, data=moore)**

**par(mfrow=c(2,2))**

**plot(conformity ~ status\*authoritarism, data=moore) # son 2 grafics**

**par(mfrow=c(2,2))**

**interaction.plot(status, authoritarism, conformity)**

**interaction.plot(authoritarism, status, conformity)**

**# One-way i two-way directament amb el mètode aov()**

**oneway.test(conformity ~ status, data=moore, var.equal=FALSE)**

**oneway.test(conformity ~ authoritarism, data=moore, var.equal=FALSE)**

**# Interaccions són significatives ¿¿?**

**aov(conformity ~ status\*authoritarism, data=moore)**

**# Calculo tots els models ANOVA: fixeu-vos en la reparametrització**

**moore.w1 <- lm(conformity ~ status, data=moore)**

**summary(moore.w1)**

**moore.w1**

**moore.w2 <- lm(conformity ~ authoritarism, data=moore)**

**summary(moore.w2)**

**moore.w2**

**moore.w3 <- lm(conformity ~ status+authoritarism, data=moore)**

**summary(moore.w3)**

**moore.w3**

**plot(moore.w3)**

**moore.w4 <- lm(conformity ~ status\*authoritarism, data=moore)**

**anova(moore.w4)**

**summary(moore.w4)**

**moore.w4**

**# Calculo el model nul i els comparo amb anova()**

**moore.w0 <- lm(conformity~1, data=moore)**

**summary(moore.w0)**

**moore.w0**

**anova(moore.w0, moore.w1)**

**anova(moore.w0, moore.w2)**

**anova(moore.w1, moore.w3)**

**anova(moore.w2, moore.w3)**

**# Interaccions són significatives ¿¿?**

**anova(moore.w3, moore.w4)**

**# Falten els contrastos demanats**

**# Falta l'anàlisi dels residus i diagnosi del model …**

**# Hi han 2 outliers dels residus a 16 i 22 són status low i author high amb conformity alta**

**df<- moore[-c(16,19),]**

**m3 <- lm(conformity ~ status+authoritarism, data=df)**

**m4 <- lm(conformity ~ status\*authoritarism, data=df)**

**m0 <- lm(conformity~ 1, data=df)**

**# Prediccions:**

**pdf <- data.frame(status=rep(levels(df$status),3),authoritarism=rep(levels(df$authoritarism),2))**

**pdf**

**pdf$pm4 <- predict(m4,new=pdf)**

**pdf$pm3 <- predict(m3,new=pdf)**

**pdf**

**# Models ANCOVA**

**# Calculo tots els models ANOVA i aplico contrastos**

**# o bé faig us del procediment step**

**moore.w1 <- lm(conformity ~ status, data=moore)**

**summary(moore.w1)**

**moore.w1**

**moore.w2 <- lm(conformity ~ numeric\_authoritarism, data=moore)**

**summary(moore.w2)**

**moore.w2**

**moore.w3 <- lm(conformity ~ status+numeric\_authoritarism, data=moore)**

**summary(moore.w3)**

**moore.w3**

**moore.w4 <- lm(conformity ~ status\*numeric\_authoritarism, data=moore)**

**summary(moore.w4)**

**moore.w4**

**# Calculo el model nul i els comparo automàticament**

**moore.w0 <- lm(conformity~1, data=moore)**

**summary(moore.w0)**

**moore.w0**

**anova(moore.w0, moore.w1)**

**anova(moore.w0, moore.w2)**

**anova(moore.w0, moore.w3)**

**anova(moore.w2, moore.w3)**

**anova(moore.w1, moore.w3)**

**anova(moore.w3, moore.w4)**

## Exercise 4: Accident Rates in Minnesota (Fox and Weisberg)

The data comes from a unpublished master's paper by Carl Hoffstedt. They relate the automobile accident rate, in accidents per million vehicle miles to several potential terms. The data include 39 sections of large highways in the state of Minnesota in 1973. The goal of this analysis was to understand the impact of design variables, acpt, slim, sigs1, and shld that are under the control of the highway department, on accidents. This data frame contains the following columns:

**rate** 1973 accident rate per million vehicle miles

**len**  length of the Highway1 segment in miles

**ADT** average daily traffic count in thousands

**trks** truck volume as a percent of the total volume

**sigs1** (number of signalized interchanges per mile times len + 1)/len, the number of signals per mile of roadway, adjusted to have no zero values.

**slim** speed limit in 1973

**shld** width in feet of outer shoulder on the roadway

**lane** total number of lanes of traffic

**acpt** number of access points per mile

**itg** number of freeway-type interchanges per mile

**lwid** lane width, in feet

**hwy** An indicator of the type of roadway or the source of funding for the road, either MC, FAI, PA, or MA

### References

Fox, J. and Weisberg, S. (2011) An R Companion to Applied Regression, Second Edition, Sage.

Weisberg, S. (2005) Applied Linear Regression, Third Edition. Wiley, Section 7.2.

> summary(high)

rate len ADT trks sigs1 slim

Min. :1.610 Min. : 2.960 Min. : 1.00 Min. : 6.000 Min. :0.04545 Min. :40

1st Qu.:2.630 1st Qu.: 7.995 1st Qu.: 5.00 1st Qu.: 8.000 1st Qu.:0.08738 1st Qu.:50

Median :3.050 Median :11.390 Median :13.00 Median : 9.000 Median :0.17666 Median :55

Mean :3.933 Mean :12.884 Mean :19.62 Mean : 9.333 Mean :0.51072 Mean :55

3rd Qu.:4.595 3rd Qu.:17.800 3rd Qu.:24.00 3rd Qu.:11.000 3rd Qu.:0.71515 3rd Qu.:60

Max. :9.230 Max. :40.090 Max. :73.00 Max. :15.000 Max. :2.78933 Max. :70

shld lane acpt itg lwid hwy

Min. : 1.000 Min. :2.000 Min. : 2.20 Min. :0.0000 Min. :10.00 FAI: 5

1st Qu.: 4.000 1st Qu.:2.000 1st Qu.: 6.95 1st Qu.:0.0000 1st Qu.:12.00 MA :13

Median : 8.000 Median :2.000 Median :10.30 Median :0.1300 Median :12.00 MC : 2

Mean : 6.872 Mean :3.128 Mean :12.16 Mean :0.2964 Mean :11.95 PA :19

3rd Qu.:10.000 3rd Qu.:4.000 3rd Qu.:14.60 3rd Qu.:0.3600 3rd Qu.:12.00

Max. :10.000 Max. :8.000 Max. :53.00 Max. :1.5400 Max. :13.00

>

**Response variable is *rate*.**

1. Determinar si la variable de respuesta (accidentes) tiene una distribución aceptablemente normal.
2. Valorad la estadística descriptiva de las variables explicativas numéricas presente en el espacio de trabajo.
3. Indicad por exploración de los datos cuales son aparentemente las variables más asociadas con la variable de respuesta (emplead únicamente las variables indicadas).
4. Indicad a nivel marginal, una a una, cuáles de las covariables disponibles tienen una correlación lineal significativamente distinta de cero.
5. Calculad e interpretar la explicabilidad del modelo anova de una vía que explica la TASA DE ACCIDENTES POR MILLA y ADT según el tipo de carretera (HWY).
6. Pensáis que la variabilidad de la respuesta POR MILLA y ADT depende del tipo de carretera?
7. Pensáis que la tasa media de accidentes POR MILLA y ADT depende del tipo de carretera?
8. Calculad el modelo de regresión lineal que explica la respuesta a partir de la longitud en millas del segmento, el aforo diario medio, el porcentaje de camiones y la tasa de señalización por milla. Interpretad el modelo de regresión y valorad la aportación de cada una de sus variables.
9. Cuál es el porcentaje de la variabilidad del número de accidentes que viene explicada por la longitud del segmento una vez que el aforo medio diario, el porcentaje de camiones y la tasa de señalización por milla ya están incluídos en el modelo?
10. Valorad a grandes rasgos la calidad del modelo en función de la primera impresión de la diagnosis de sus residuos.
11. Aplicad alguna transformación a la variable de respuesta que mejore las propiedades del modelo de regresión lineal.
12. Estudiad alguna transformación de las variables longitud del segmento, intensidad media diaria y porcentaje de camiones que mejore la calidad del modelo de regresión lineal para explicar la respuesta a partir de las variables explicativas len, ADT, trks y sigs1.

**Se construye un nuevo modelo en escala logarítmica de la respuesta con los términos logarítmicos de la longitud, la intensidad media diaria, el porcentaje de camiones y la tasa de señalización por milla y el resto de las variables numéricas disponibles (sin transformarlas).**

1. Determinar qué variables son estadísticamente significativas en el modelo descrito. Valorad la calidad del modelo.
2. Determinar si en el modelo resultante del punto anterior, el efecto aditivo del tipo de carretera merece la pena incluirlo en el modelo.
3. Determinar si en el modelo resultante del **punto 17**, la relación entre la tasa de accidentes (respuesta) y el límite de velocidad (slim) depende del tipo de carretera.
4. Valorad la presencia de outliers en los residuos estudentizados a un nivel de confianza del 99%. Indicad cuáles son esas observaciones **en el modelo resultante del Punto 19**.
5. Estudiad la presencia de valores influyentes a priori, indicando su número según el criterio estudiado en clase **en el modelo resultante del Punto 19**.
6. Estudiad la presencia de valores influyentes a posteriori, indicando el criterio estudiado en clase y las observaciones realmente atípicas **en el modelo resultante del Punto 19**.
7. Dado un segmento con valores de las variables explicativas en la mediana, cuál sería la tasa de accidentes esperada con un intervalo de confianza del 95% para cada tipo de carretera ?
8. Valorad lo que habéis aprendido trabajando con este interesante juego de datos reales.

## Exercise 5: Performance of sorting algorithms

We have collected performance data related to Sorting Algorithm. Aggregated data on:

* Size in thousands of list length
* CPUtime
* f.algo: 3 sorting algorithm are compared Quick sort, Heap sort and Intro sort. Introsort was invented by David Musser in Musser (1997), in which he also introduced introselect, a hybrid selection algorithm based on quickselect (a variant of quicksort), which falls back to median of medians and thus provides worst-case linear complexity, which is optimal. It begins with quicksort and switches to heapsort when the recursion depth exceeds a level based on (the logarithm of) the number of elements being sorted.

Analyze the performance of the algorithms.