laSalle
UNIVERSITAT RAMON LLULL

Matriz de Covarianza
ANOVA y sus amigos

Míriam Calvo
Xavier Vilasís
Ángel Berián
Ricard Sierra
Máster Universitario en Data Science

Resumen de lo Publicado

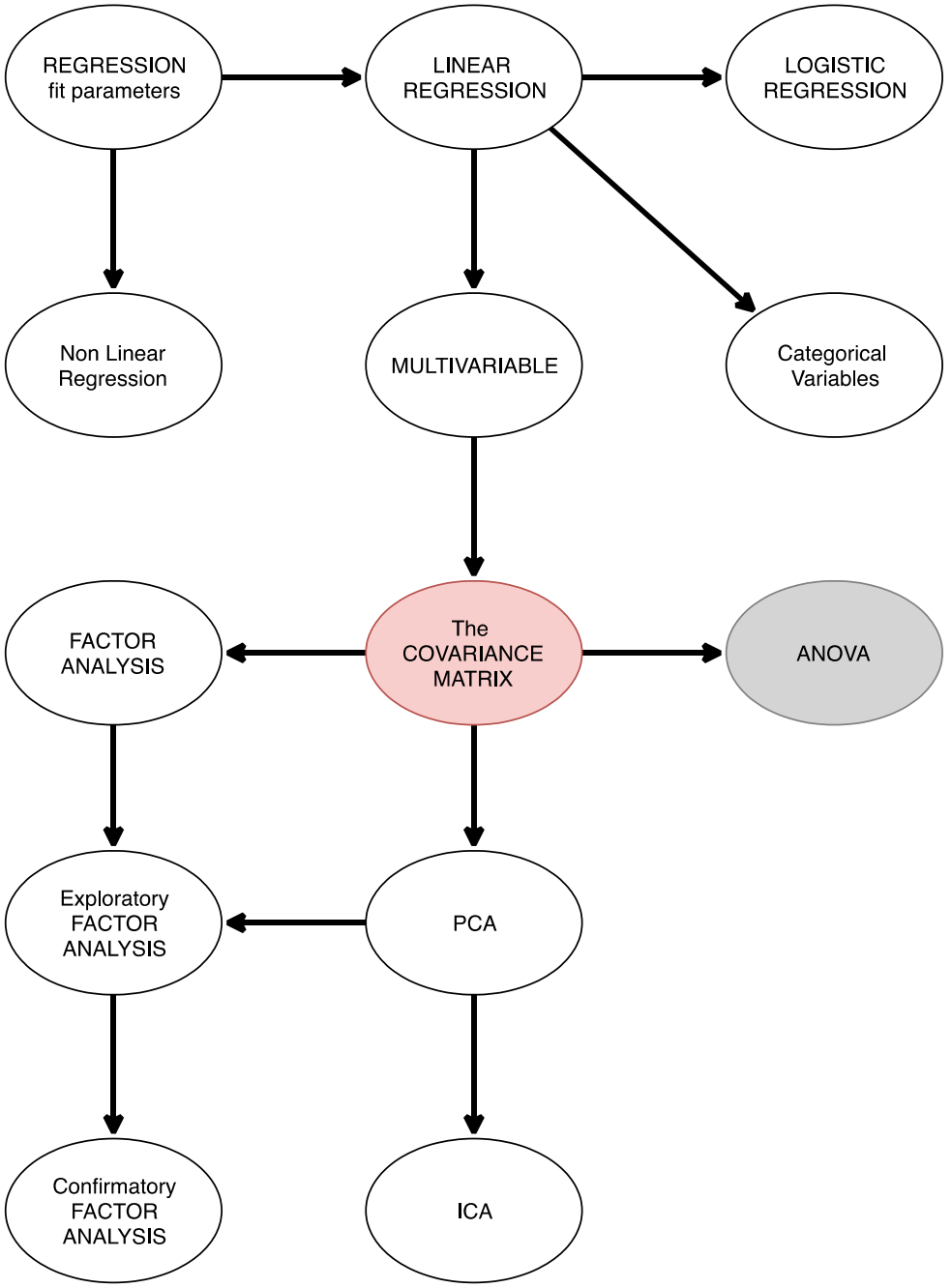
Probabilidad

Estadística Descriptiva

Test de hipótesis

Regresión lineal

The map


REGRESSION
fit parameters
LINEAR
REGRESSION
LOGISTIC
REGRESSION
Non Linear
Regression
MULTIVARIABLE
Categorical
Variables
FACTOR
ANALYSIS
The
COVARIANCE
MATRIX
ANOVA
Exploratory
FACTOR
ANALYSIS
PCA
Confirmatory
FACTOR
ANALYSIS
ICA

Regresión lineal simple

- La variable dependiente es combinación lineal de los parámetros (no necesariamente de las variables independientes).
- El caso más simple es

$$Y = \beta_0 + \beta_1 X$$

- De la estimación de los parámetros por mínimos cuadrados (otra manera alternativa de expresarlos en función de los valores medios):

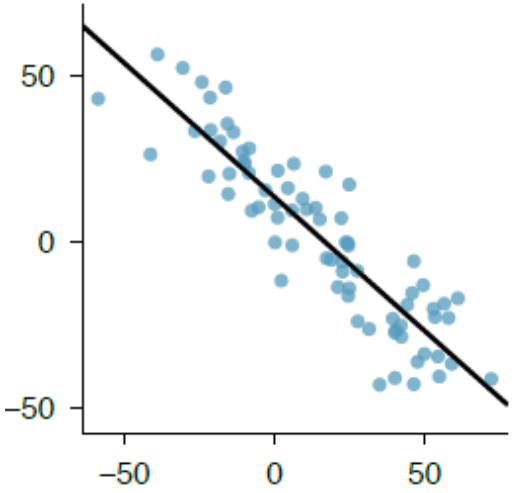
$$\hat{y} = \beta_0 + \beta_1 x \quad \begin{cases} \beta_1 = \dfrac{\sum (x_i - \bar{x})\,(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \\ \beta_0 = \bar{y} - \beta_1 \bar{x} \end{cases}$$

$$\hat{y} = \bar{y} + \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$$

covarianza

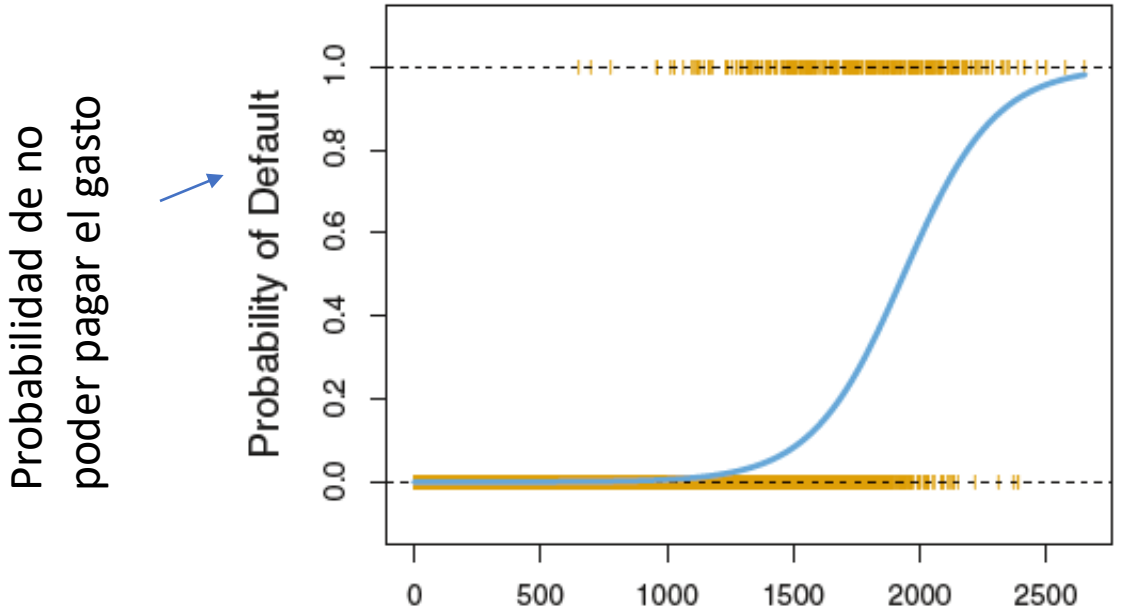
varianza

la recta de regresión pasa por el punto $(\bar{x}, \bar{y})$



Regresión logística

- Cuando el output o predicción es cualitativo o una categoría (sí/no, 0/1, etc), hablamos de clasificación.
- En el caso de 2 categorías posibles, podemos usar una regresión logística, que dará la probabilidad de que Y pertenezca a una categoría determinada.
- Da un valor entre 0 y 1.
- (Para el ajuste se usa el método de *maximum likelihood)*.



- ¿Cómo pasar de una valor continuo obtenido en una regresión lineal a una probabilidad?

Measure	Min	Max	Name
Pr(Y = 1)	0	1	"probability"
$\frac{\Pr(Y=1)}{1-\Pr(Y=1)}$	0	∞	"odds"
$\log\left(\frac{\Pr(Y=1)}{1-\Pr(Y=1)}\right)$	-∞	∞	"log-odds" or "logit"

$$\log\left(\frac{\Pr(Y=1)}{\Pr(Y=0)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_r X_r$$

- Es decir, podemos obtener la probabilidad $\Pr(Y = 1)$ en función de una variable X como:

$$\Pr(Y = 1) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

- En el ejemplo de *Default* vs *Balance*:

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.6513	0.3612	−29.5	<0.0001
balance	0.0055	0.0002	24.9	<0.0001

Podemos usar los coeficientes obtenidos para predecir la probabilidad de no poder hacer frente al pago. Para alguien con un balance de 1000$:

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1{,}000}}{1 + e^{-10.6513+0.0055\times 1{,}000}} = 0.00576$$

La probabilidad es menor al 1%. Para alguien con balance de 2000$, será el 58.6%.

Ejemplo de regresión logística

Usamos el paquete

ISLR: Data for an Introduction to Statistical Learning with Applications in R

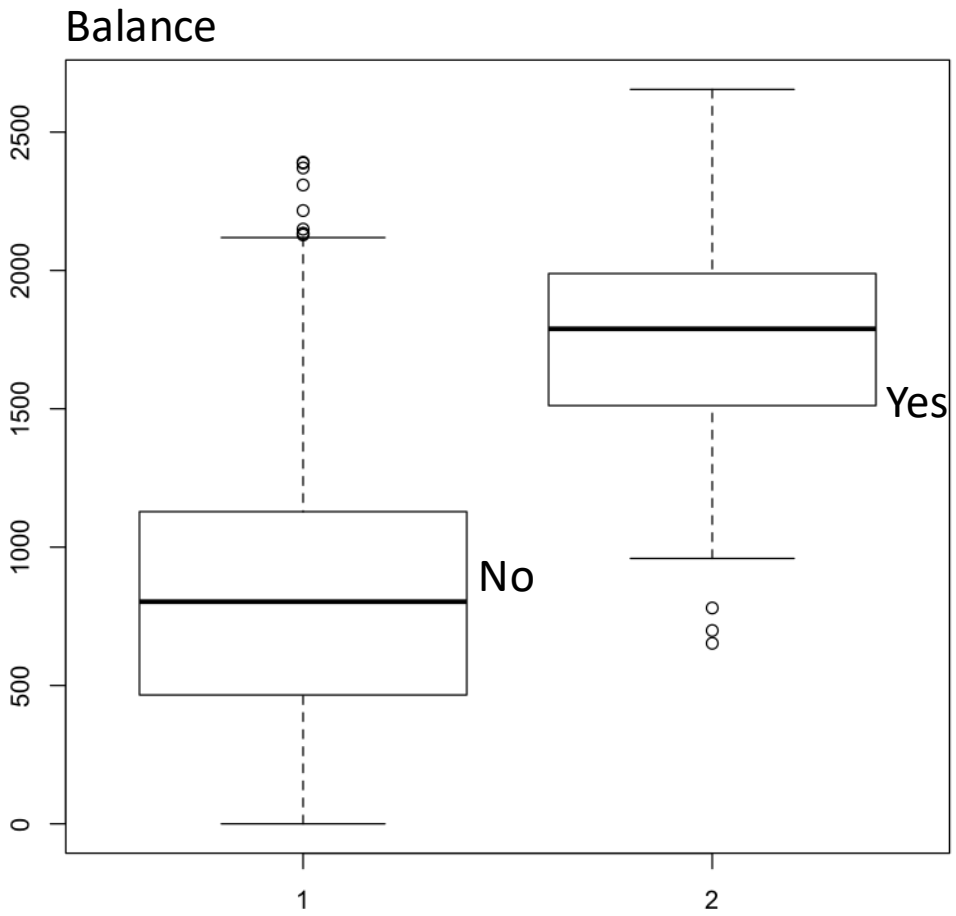
The **default** data set contains selected variables and data for 10,000 credit card users. Some of the variables are:

- **student** - A binary factor containing whether or not a given credit card holder is a student.
- **income** - The gross annual income for a given credit card holder.
- **balance** - The total credit card balance for a given credit card holder.
- **default** - A binary factor containing whether or not a given user has defaulted on his/her credit card.

The goal of our investigation is to fit a model such that the relavent predictors of credit card default are illucidated given these variables.

default	student	balance	income
No	No	729.5265	44361.625
No	Yes	817.1804	12106.13

Ejemplo de regresión logística



$$P(\text{default}) = \frac{\exp(\beta_0 + \beta_1 b)}{1 + \exp(\beta_0 + \beta_1 b)}$$

b is for balance

glm(formula = default ~ balance, family = "binomial", data = Default)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.2697	-0.1465	-0.0589	-0.0221	3.7589

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.065e+01	3.612e-01	-29.49	<2e-16 ***
balance	5.499e-03	2.204e-04	24.95	<2e-16 ***

Otros tipos de regresiones

- Simple quiere decir con una sola variable independiente X.
- Lineal en los parámetros.

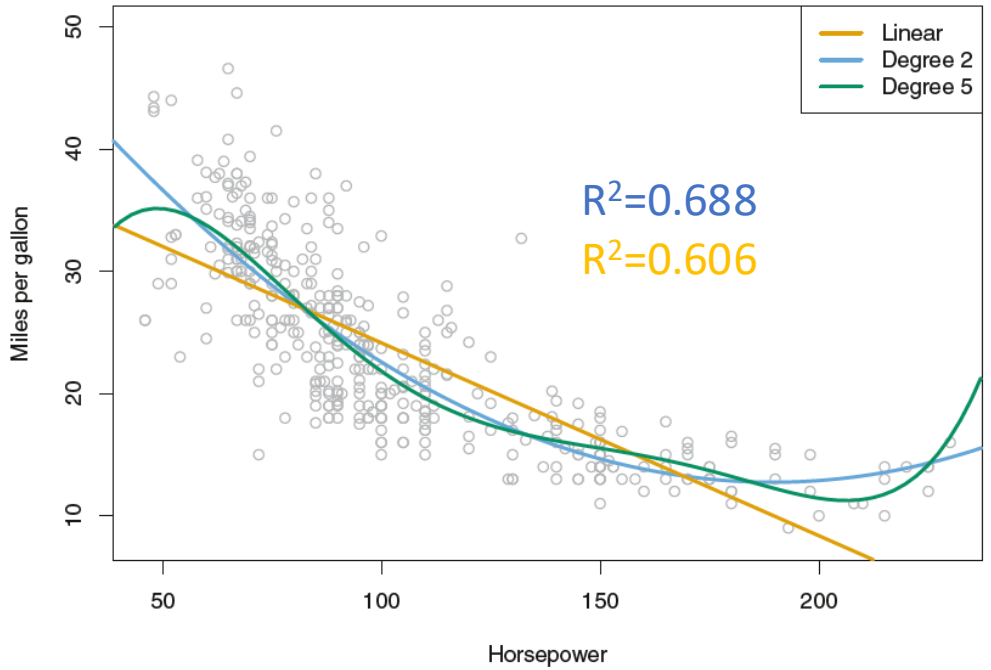
- Regresión polinomial:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \cdots + \beta_d X_1^d + \epsilon$$

- Regresión con *splines*: permite cambiar la curva de regresión según el intervalo.
  - El punto de cambio es arbitrario, se suele decidir tras una inspección visual.

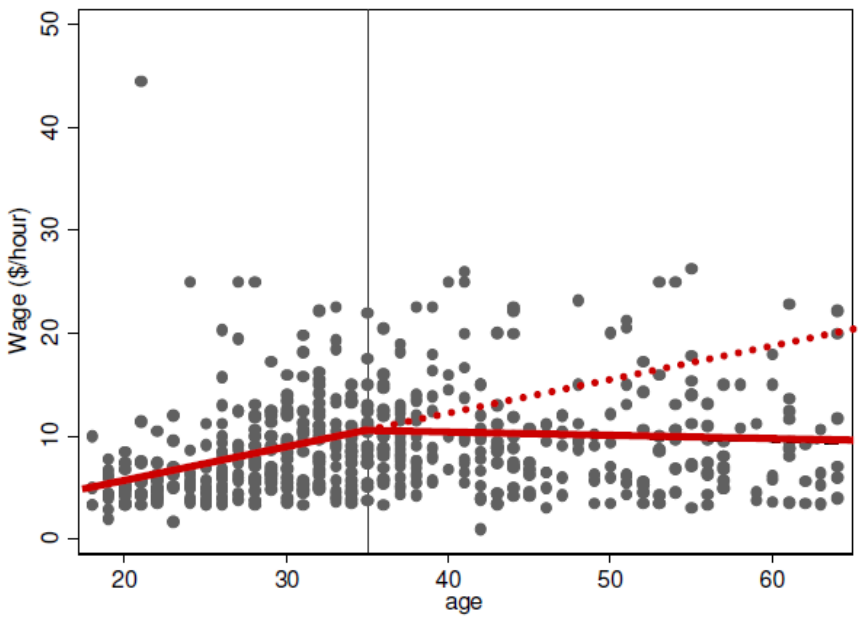
- **Regresión polinomial**
  - Sigue siendo lineal (en los parámetros), aunque algún X_i sea una potencia.
  - No es usual ir más alla de 3º o 4º orden, o el polinomio puede tomar formas raras.

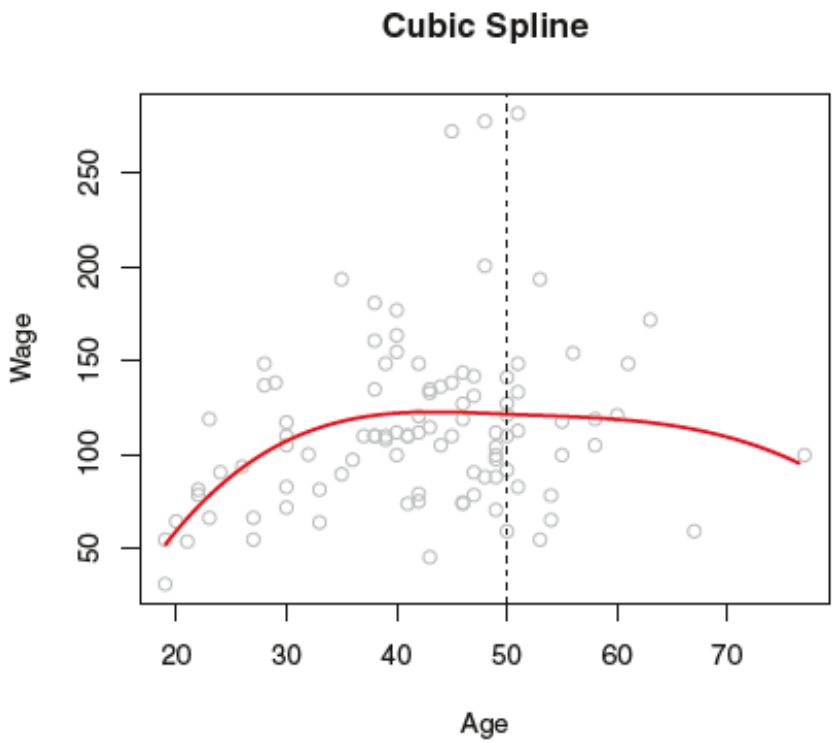
$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$



	Coefficient	Std. error	t-statistic	p-value
Intercept	56.9001	1.8004	31.6	< 0.0001
horsepower	-0.4662	0.0311	-15.0	< 0.0001
horsepower^2	0.0012	0.0001	10.1	< 0.0001

- **Regresión con splines**
- En vez de ajustar un polinomio en todo el rango de una variable, se trata de ajustar polinomios de menor orden en diferentes regiones de X.
- Los puntos donde los coeficientes cambian se conocen como *knots*.
- Suele imponerse que en los *knots* la función sea continua (así como la d-1 derivadas para polinomios de grado d).





Regresión multivariable (o múltiple)

- En general, si tenemos p variables independientes.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- Si al añadir una nueva variable R^2 aumenta, quiere decir que el modelo es mejor.

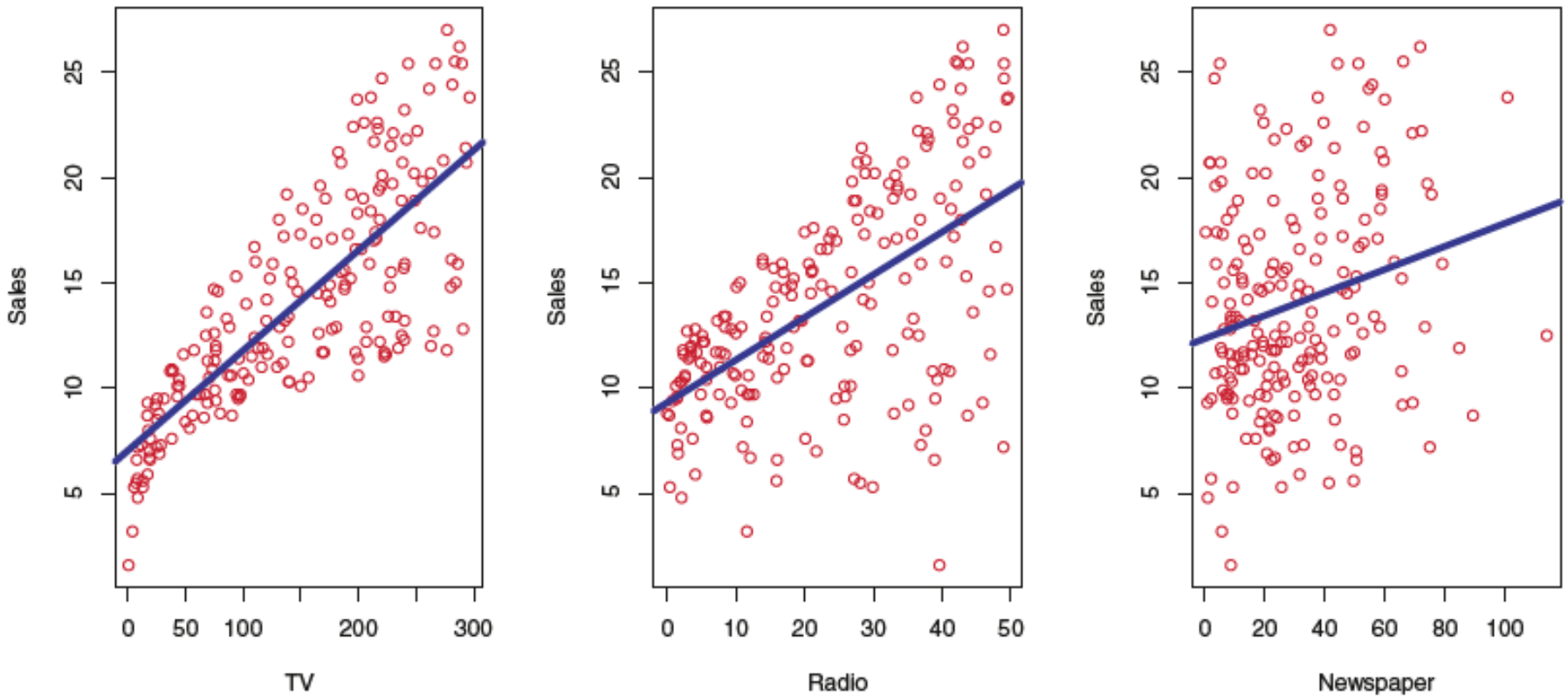
$$\boldsymbol{R^2} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

- Siguiendo el ejemplo de ventas vs anuncios:

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon.$$

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	−0.001	0.0059	−0.18	0.8599

Quantity	Value
Residual standard error	1.69
R^2	0.897
F-statistic	570


Sales
5
10
15
20
25
0
50
100
200
300
TV
Sales
5
10
15
20
25
0
10
20
30
40
50
Radio
Sales
5
10
15
20
25
0
20
40
60
80
100
Newspaper

Regresiones lineales y variables discretas

Para introducir el efecto de una variable categórica, tenemos dos opciones:

- Hacer dos regresiones lineales

$$d = 0, \quad y = \beta_{00} + \beta_{01} x,$$
$$d = 1, \quad y = \beta_{10} + \beta_{11} x.$$

Regresiones lineales y variables discretas

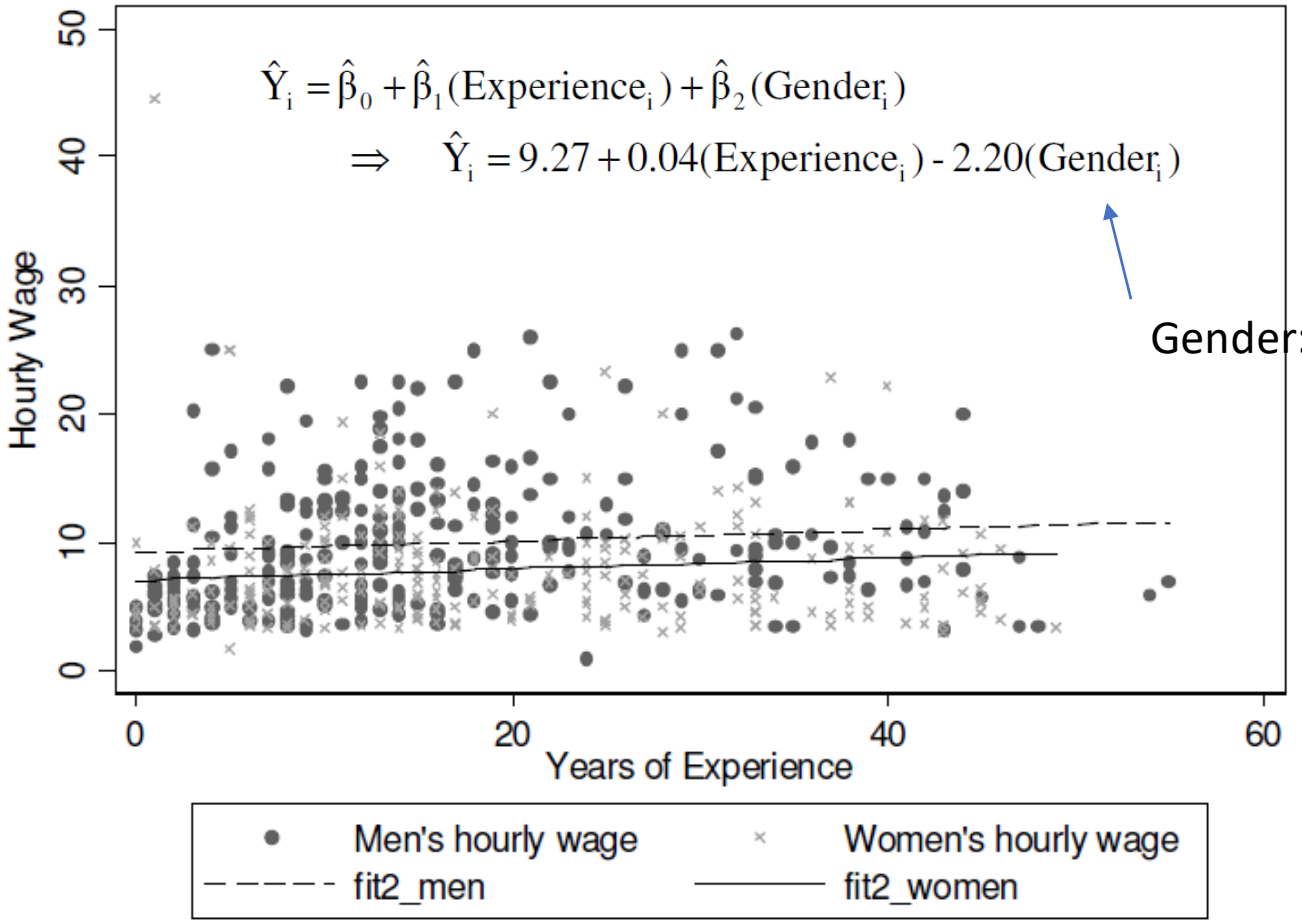
- Usar un término de interacción

$$y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 x d.$$

- La relación entre los coeficientes es

$$\beta_0 = \beta_{00}, \quad \beta_2 = \beta_{10} - \beta_{00}, \quad \beta_1 = \beta_{01}, \quad \beta_3 = \beta_{11} - \beta_{01}.$$

- Hemos visto el caso en que X es una variable continua, pero también podemos aplicar una regresión para una <u>variable X discreta</u> (suele referirse como *dummy variable*).
  - P. ej. para distinguir diferentes grupos o categorías.



Gender: 0 = Hombre
1 = Mujer

- Cuando una variable independiente modifica el efecto asociado a otra, hará falta añadir términos de interacción.

$$E[\text{Wage}_i] = \hat{\beta}_0 + \hat{\beta}_1(\text{Experience}_i) + \hat{\beta}_2(\text{Gender}_i) + \hat{\beta}_3(\text{Gender}_i \times \text{Experience}_i)$$

Permite que la diferencia entre hombres y mujeres cambie según la experiencia (las rectas de regresión ya no son paralelas).
Ecuación para hombres:

$$E[Wage_i] = \hat{\beta}_0 + \hat{\beta}_1(\text{Experience}_i)$$

Ecuación para mujeres:

$$E[Wage_i] = (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_3)(\text{Experience}_i)$$

Variables correlacionadas

- ¿Qué ocurre cuando tenemos una regresión con múltiples variables que no son independientes ?
- (o al menos no están de-correlacionadas)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_0$$

$$x_2 = \alpha_0 + \alpha_1 x_1 + \varepsilon_1$$

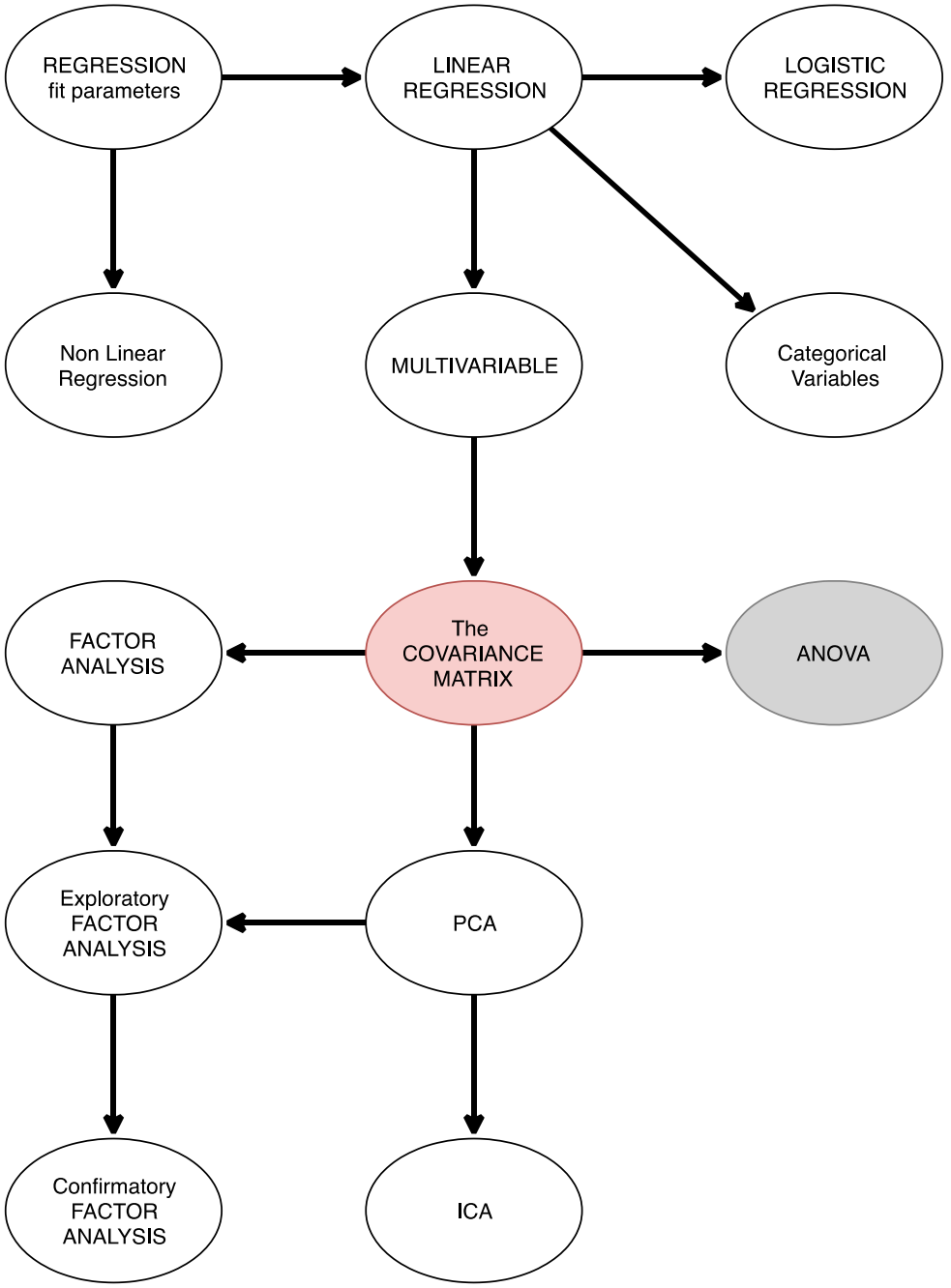
- Sustituyendo, nos queda

$$y = \beta_0 + \beta_1 x_1 + \beta_2 (\alpha_0 + \alpha_1 x_1 + \varepsilon_1) + \varepsilon_0$$

$$y = (\beta_0 + \beta_2 \alpha_0) + (\beta_1 + \beta_2 \alpha_1) x_1 + (\varepsilon_0 + \beta_2 \varepsilon_1)$$

$$y = \gamma_0 + \gamma_1 x_1 + \varepsilon_2$$

The Map


REGRESSION
fit parameters
LINEAR
REGRESSION
LOGISTIC
REGRESSION
Non Linear
Regression
MULTIVARIABLE
Categorical
Variables
FACTOR
ANALYSIS
The
COVARIANCE
MATRIX
ANOVA
Exploratory
FACTOR
ANALYSIS
PCA
Confirmatory
FACTOR
ANALYSIS
ICA

La Matriz de covarianza

- De la regresión lineal hemos aprendido que la **covarianza** (correlación) nos da una medida de la **relación lineal** que hay entre dos variables.
- Del **teorema del límite central** sabemos que muchos fenómenos son **gausianos**
- De la función generatriz de los cumulantes (tranquilo todo el mundo, que ya se que esto no lo hemos visto, pero si que lo he comentado), sabemos que las **gausianas 'solo' tienen media y varianza (covarianza)** diferente de cero.
- **Toda la relación en entre gausianas es forzosamente lineal**.

La Matriz de Covarianza

- Recordamos la covarianza,

$$\mathrm{Cov}(x_i, x_j) =< (x_i - \mu_i)(x_j - \mu_j) >= \sigma_{ij}$$
$$\mathrm{Cov}(x_i, x_i) =< (x_i - \mu_i)(x_i - \mu_i) >= \sigma_{ii} = \sigma_i^2$$

- La covarianza incorpora toda la información del modelo, para fenómenos gausianos (y para los no gausianos, seguro que lleva algo importante.
- Construimos una matriz con las covarianzas. La matriz de covarianza.

$$\Sigma_{ij} = \sigma_{ij}$$

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12}^{(2)} & \sigma_{13}^{(2)} & \dots \\ \sigma_{12}^{(2)} & \sigma_2^2 & \sigma_{23}^{(2)} & \dots \\ \sigma_{13}^{(2)} & \sigma_{23}^{(2)} & \sigma_3^2 & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix}$$

La Matriz de Covarianza y de Correlación

- La varianza nos aporta información sobre el grado de aleatoriedad de una variable.
- Entender qué elementos contribuyen a la varianza nos ayuda a entender el fenómeno que estamos estudiando.

- Podemos construir una matriz con los coeficientes de correlación en vez de las covarianzas : la matriz de correlación.

$$R = \begin{pmatrix} \rho_1^2 & \rho_{12}^{(2)} & \rho_{13}^{(2)} & \dots \\ \rho_{12}^{(2)} & \rho_2^2 & \rho_{23}^{(2)} & \dots \\ \rho_{13}^{(2)} & \rho_{23}^{(2)} & \rho_3^2 & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix}$$

Ejemplo : babies

Datos sobre peso de los niños al nacer
Extraído del libro
D.Nolan, T.Speed, Stat Labs, Mathematical Statistics Through Applications, Springer Verlag (2000)
https://www.stat.berkeley.edu/users/statlabs/

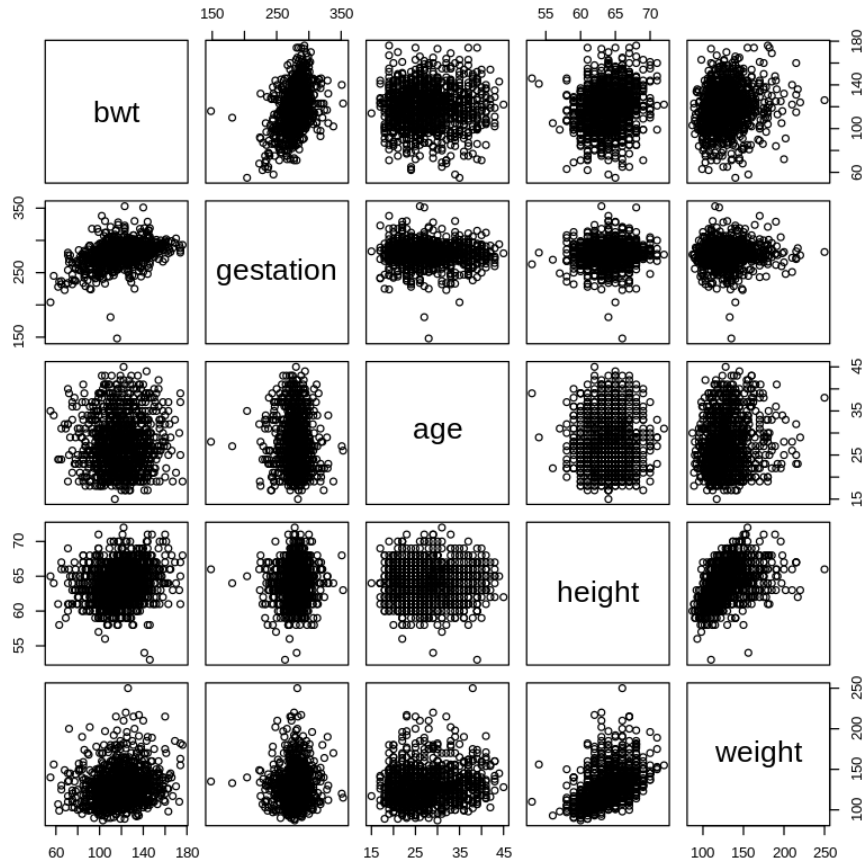
Variable	Description
bwt	Birth weight in ounces (999 unknown)
gestation	Length of pregnancy in days (999 unknown)
parity	0= first born, 9=unknown
age	mother's age in years
height	mother's height in inches (99 unknown)
weight	Mother's prepregnancy weight in pounds (999 unknown)
smoke	Smoking status of mother 0=not now, 1=yes now, 9=unknown

	bwt <int>	gestation <int>	parity <int>	age <int>	height <int>	weight <int>	smoke <int>
1	120	284	0	27	62	100	0
2	113	282	0	33	64	135	0
3	128	279	0	28	64	115	1
4	123	999	0	36	69	190	0
5	108	282	0	23	67	125	1
6	136	286	0	25	62	93	0

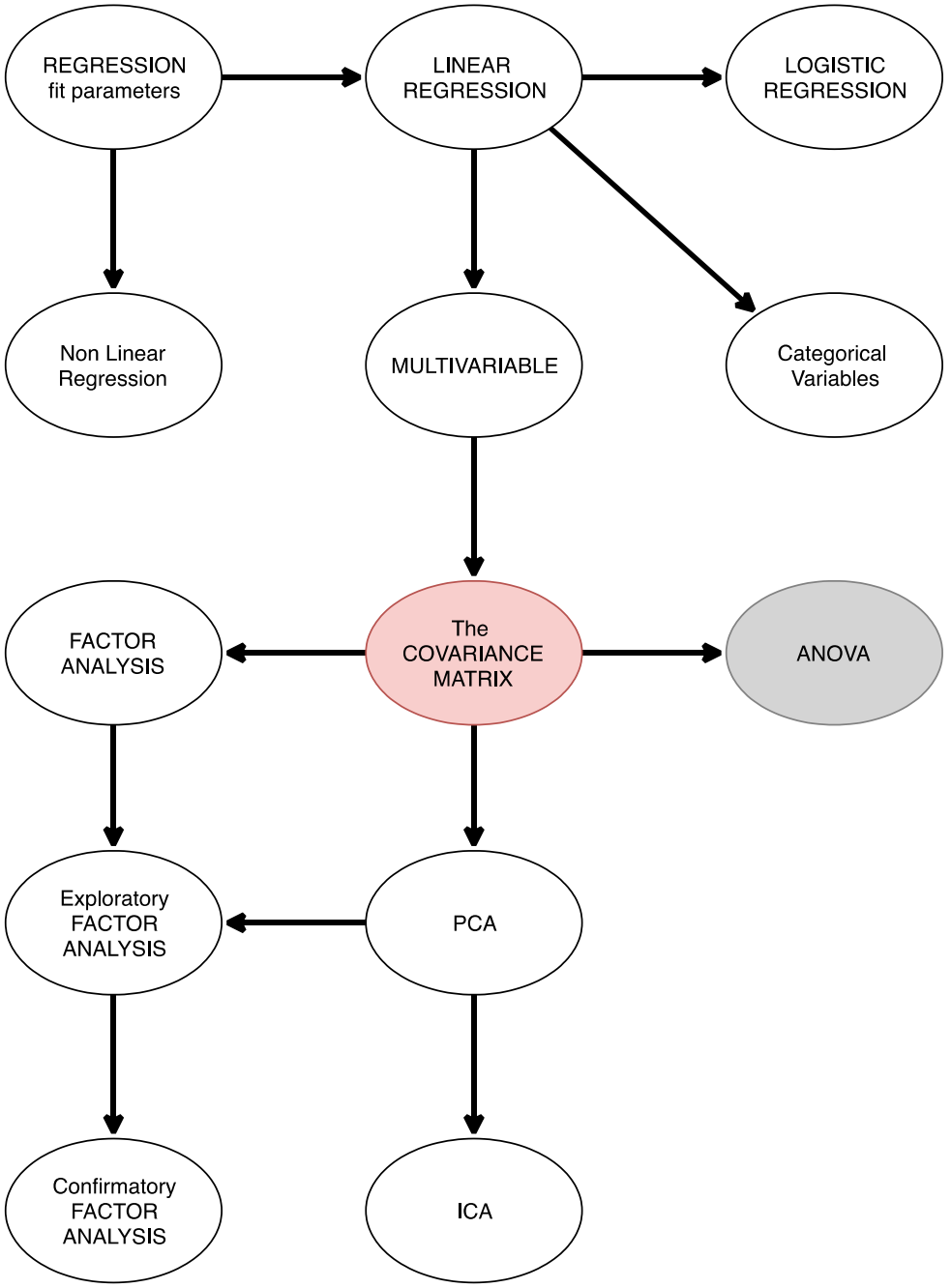
Ejemplo : babies

	bwt	gestation	age	height	weight
bwt	337.159204	120.379444	3.32411585	9.32802826	59.401573
gestation	120.379444	255.260549	-4.71144351	2.76848470	7.397214
age	3.324116	-4.711444	33.74935959	-0.07472913	17.830763
height	9.328028	2.768485	-0.07472913	6.38364942	22.924155
weight	59.401573	7.397214	17.83076266	22.92415498	432.436451

	bwt	gestation	age	height	weight
bwt	1.00000000	0.41033917	0.031162039	0.201065626	0.15556765
gestation	0.41033917	1.00000000	-0.050760907	0.068582887	0.02226461
age	0.03116204	-0.05076091	1.000000000	-0.005091229	0.14759648
height	0.20106563	0.06858289	-0.005091229	1.000000000	0.43631252
weight	0.15556765	0.02226461	0.147596481	0.436312519	1.00000000



The Map


REGRESSION
fit parameters
LINEAR
REGRESSION
LOGISTIC
REGRESSION
Non Linear
Regression
MULTIVARIABLE
Categorical
Variables
FACTOR
ANALYSIS
The
COVARIANCE
MATRIX
ANOVA
Exploratory
FACTOR
ANALYSIS
PCA
Confirmatory
FACTOR
ANALYSIS
ICA

ANOVA

ANOVA stands for ANalysis Of Variance

What is this about ?

Model - Factor – Group effects

Example

Birth-weight of children from smoking mothers

bwt	gestation	parity	age	height	weight	smoke
120	284	0	27	62	100	0
113	282	0	33	64	135	0
128	279	0	28	64	115	1
123	999	0	36	69	190	0
108	282	0	23	67	125	1
136	286	0	25	62	93	0

All

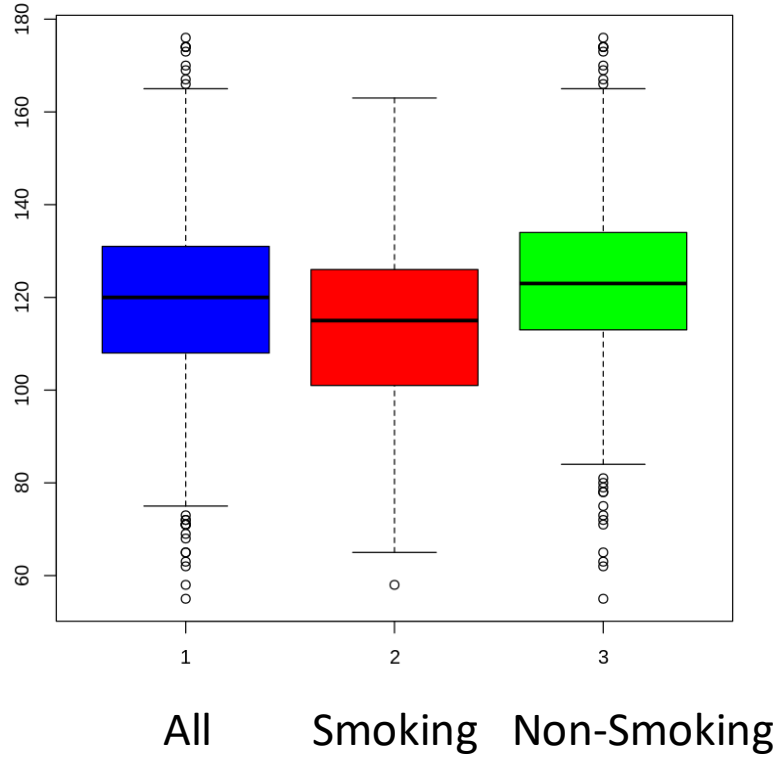
```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
55.0   108.0   120.0   119.5   131.0   176.0
```

Smoking

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
58.0   101.0   115.0   113.8   126.0   163.0
```

Non-Smoking

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
55.0   113.0   123.0   123.1   134.0   176.0
```



Example

- Our goal is to model bwt according to

$$bwt_{ij} = \mu + \alpha_i + u_{ij}$$

- Mu is a general average
- Alpha is the class average
  - Smoking
  - Nonsmoking
- U is the inherent difference of the individual inside his group.

Example

- Mu takes into account all global effects.
- This means that alpha's are understood as relative differences

$$\sum_i \alpha_i = 0$$

- U_{ij} should be zero mean gaussian random variable.

The problem

Our data x_{ij}

- i is the class index
- j is the index inside the class

$$x_{ij} = \mu + \alpha_i + u_{ij}$$

Determine whether this is a good model or not.

For this we need to compute mu and alphas.

Formula ball

- The global average
  - N number of clases
  - n number of elements in each class

$$\overline{x} = \frac{1}{Nn} \sum_{i,j=1}^{N,n} x_{ij}$$

- The average inside class i

$$\overline{x}_i = \frac{1}{n} \sum_{j=1}^{n} x_{ij}$$

- Sums of squares
  - Total

$$\mathrm{SST} = \sum_{i,j=1}^{N,n} (x_{ij} - \overline{x})^2$$

Formula ball, extended

- SSE (Sum of Squared Errors)
  - Differences inside groups – **Residual differences**

$$\mathrm{SSE} = \sum_{i,j=1}^{N,n} (x_{ij} - \overline{x}_i)^2$$

- SSA (Sum of Squares for a factor A)
  - **Differences from one group to the other**

$$\mathrm{SSA} = \sum_{i=1}^{N} (\overline{x}_i - \overline{x})^2$$

- And SST (Total Sum of Squares)

$$\mathrm{SST} = \mathrm{SSA} + \mathrm{SSE}$$

Key idea

There is a difference between groups if

SSA

is larger than the residual differences

SSE

Back to the model

$$x_{ij} = \mu + \alpha_i + u_{ij}$$

- If we plug our model, we have

$$\overline{x}_i = \mu + \alpha_i + \frac{1}{n}\sum_{j=1}^{n} u_{ij}$$

$$\overline{x} = \mu + \frac{1}{N}\sum_{i=1}^{N} \alpha_i + \frac{1}{Nn}\sum_{i,j=1}^{N,n} u_{ij}$$

SSA and SSE

- So

$$\mathrm{SSE} = \sum_{i,j=1}^{N,n} u_{ij}^2$$

$$\mathrm{SSA} = n \sum_{i=1}^{N} \alpha_i^2$$

To compare adequately

- The mean squares

$$\mathrm{MSA} = \frac{1}{N-1}\mathrm{SSA}$$

$$\mathrm{MSE} = \frac{1}{n(N-1)}\mathrm{SSE}$$

- Should follow a χ^2 distribution of N-1 and n(N-1) degrees of freedom

The F-Test

- The hypothesis is tested using an the F-distribution

$$F = \frac{\mathrm{MSA}}{\mathrm{MSE}}$$

The F-Test

- We compute the value of F and we get the probability of getting a value of F larger than the one obtained.

- Large values of F show difference between groups.

$$H_0: \quad \beta_{k-q+1} = \beta_{k-q+2} = \cdots = \beta_k = 0$$

$$H_1: \quad \text{At least one of } \beta_{k-q+1}, \ldots, \beta_k \text{ is } \neq 0$$

Assumptions

- This analysis assumes
  - That variables follow a Gaussian distribution and that residues are independent of alphas and between themselves.
  - That variances are homogeneous : variance of data in groups is the same (homoscedasticity)

Back to babies

[6]:
```
fit<-aov(bwt ~ smoke, data=babies)
summary(fit)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
smoke         1   3835    3835   11.48 0.000728 ***
Residuals  1182 395024     334
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Mus and alphas

[13]:
```
avg<-mean(babies$bwt)
mus<-mean(smoking$bwt)
muns<-mean(nonsmoking$bwt)
mu<-(mus+muns)/2
alphas<-(mus-muns)/2
alphans<-(muns-mus)/2
cat("avg = " ,avg, "\n")
cat("mus = " ,mus, "\n")
cat("muns = ", muns, "\n")
cat("mu = ", mu, "\n")
cat("alphas = ", alphas, "\n")
cat("alphans = ", alphans, "\n")
```

```
avg =  119.5236
mus =  113.8192
muns =  123.0853
mu =  118.4522
alphas =  -4.633071
alphans =  4.633071
```

ANOVA Variations

- Deal with unhomogeneous groups
- One way ANOVA
  - One factor
- Two way ANOVA
  - Several factors
- MANOVA
  - Several variables
- ANCOVA
  - General linear model with factors

Some hint on MANOVA

We have several data from subjects following a treatment

		Treatment			
		1	2	$\cdots$	g
Subject	*Treatment* 1	$\mathbf{Y_{11}} = \begin{pmatrix} Y_{111} \\ Y_{112} \\ \vdots \\ Y_{11p} \end{pmatrix}$	$\mathbf{Y_{21}} = \begin{pmatrix} Y_{211} \\ Y_{212} \\ \vdots \\ Y_{21p} \end{pmatrix}$	$\cdots$	$\mathbf{Y_{g1}} = \begin{pmatrix} Y_{g11} \\ Y_{g12} \\ \vdots \\ Y_{g1p} \end{pmatrix}$
	Treatment 2	$\mathbf{Y_{21}} = \begin{pmatrix} Y_{121} \\ Y_{122} \\ \vdots \\ Y_{12p} \end{pmatrix}$	$\mathbf{Y_{22}} = \begin{pmatrix} Y_{221} \\ Y_{222} \\ \vdots \\ Y_{22p} \end{pmatrix}$	$\cdots$	$\mathbf{Y_{g2}} = \begin{pmatrix} Y_{g21} \\ Y_{g22} \\ \vdots \\ Y_{g2p} \end{pmatrix}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$
	n_i	$\mathbf{Y_{1n_1}} = \begin{pmatrix} Y_{1n_11} \\ Y_{1n_12} \\ \vdots \\ Y_{1n_1p} \end{pmatrix}$	$\mathbf{Y_{2n_2}} = \begin{pmatrix} Y_{2n_21} \\ Y_{2n_22} \\ \vdots \\ Y_{2n_2p} \end{pmatrix}$	$\cdots$	$\mathbf{Y_{gn_g}} = \begin{pmatrix} Y_{gn_g1} \\ Y_{gn_g2} \\ \vdots \\ Y_{gn_2p} \end{pmatrix}$

https://online.stat.psu.edu/stat505/lesson/8/8.2

Some hints on MANOVA

- we are interested in testing the null hypothesis that the group mean vectors are all equal to one another. Mathematically this is expressed as:

$$H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_g$$

- The alternative hypothesis being:

$$H_a: \mu_{ik} \neq \mu_{jk}$$

for at least one i differenet from j and at least one variable k.

- This says that the null hypothesis is false if at least one pair of treatments is different on at least one variable.