PRÁCTICA 6: Regresión y Correlación.

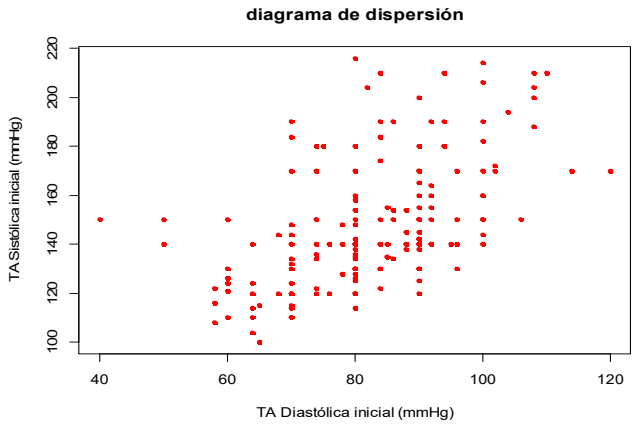
DESARROLLO DE LA PRÁCTICA

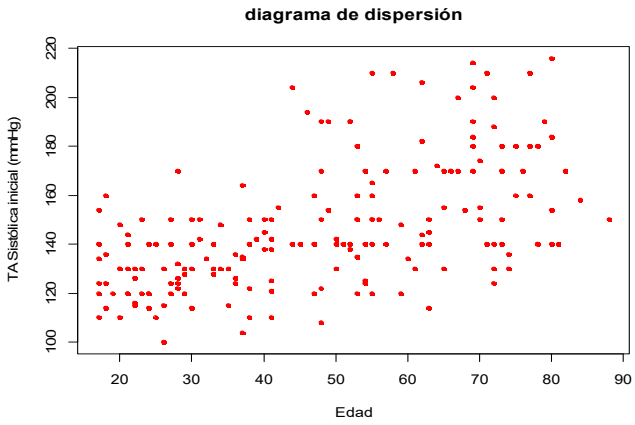
Diagrama de dispersión

Cuando se dispone de dos variables estadísticas cuantitativas, y no realizamos agrupamiento en intervalos de clase, prácticamente cada par de valores diferentes (modalidades) tienen una frecuencia unitaria. Por ello, sería deseable representar dichos pares de valores como una primera aproximación de la posible relación que existe entre ambas variables. Los siguientes comandos

plot(TAdias0,TAsist0, pch=20, col="red", xlab="TA Diastólica inicial (mmHg)", ylab="TA Sistólica inicial (mmHg)",main="diagrama de dispersión")
plot(edad,TAsist0, pch=20, col="red", xlab="Edad", ylab="TA Sistólica inicial (mmHg)")

nos proporcionan los consiguientes diagramas de dispersión





Si dicho diagrama de dispersión nos indica una marcada tendencia de los datos podemos, con ciertas garantías, modelizar dicha relación a través de diferentes modelos de regresión. Denotamos a la variable independiente por X y a la variable dependiente por Y. Podemos plantear, según lo estudiado en clase, a los siguientes modelos:

1- lineal $y = a + bx$
2- cuadrático $y = a + bx + cx^2$
3- cúbico $y = a + bx + cx^2 + dx^3$
4- hipérbola $y = a + b\frac{1}{x}$
5- logaritmo $y = a + b\log(x)$

Para ello, podemos utilizar el comando **lm** del paquete básico "**stats**", que nos ajusta los modelos lineales que le indiquemos. Así, para los datos del ejercicio 1, los comandos:

```
x<-c(2,3,4,4,5,5,6,7,7,9,9,10,11,11,12)
y<-c(11,12,10,13,11,9,10,7,12,8,7,3,6,5,5)
reg1<-lm(y~x)
summary(reg1)
reg2<-lm(y~x+I(x^2))
summary(reg2)
reg3<-lm(y~x+I(x^2)+I(x^3))
summary(reg3)
reg41<-lm(y~I(1/x))
summary(reg41)
reg5<-lm(y~log(x))
summary(reg5)
```

nos permiten entrar los datos de X e Y en forma de vector, y sucesivamente realizar los ajustes lineales (en los parámetros) propuestos con los cinco primeros modelos planteados.
Si nos detenemos un poco, en el primer ajuste (recta), los resultados resumidos del ajuste son:

```
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    14.2197      -0.8028
```

Si se desea una salida pormenorizada del ajuste realizado, tenemos los siguientes resultados

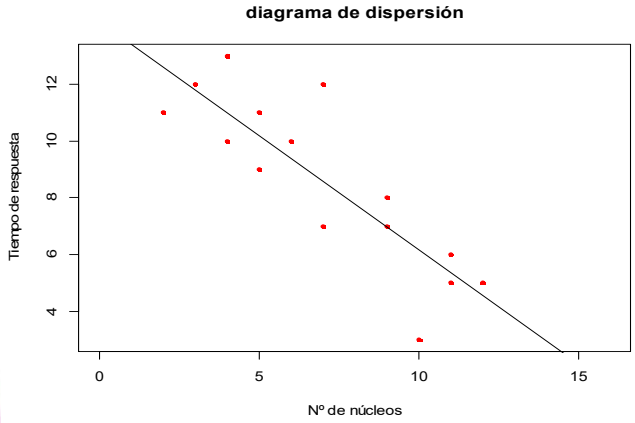
```
Call:
lm(formula = y ~ x)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.2197     1.0690  13.302 6.02e-09 ***
x            -0.8028     0.1398  -5.743 6.79e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.666 on 13 degrees of freedom
Multiple R-squared:  0.7173,   Adjusted R-squared:  0.6955
F-statistic: 32.98 on 1 and 13 DF,  p-value: 6.793e-05
```


De ambos cuadros de resultados observamos que la recta de regresión ajustada a los datos que se proporcionan (X=número de núcleos e Y=tiempos de respuesta) es

$$\hat{y} = 14.2197 - 0.8028 * x$$

que su medida de la bondad del ajuste es $R^2 = 0.7173$ ($\Rightarrow r = -0.8469$) y una varianza residual (varianza de los errores) de $S_e^2 = 1.666^2 = 2.4053$

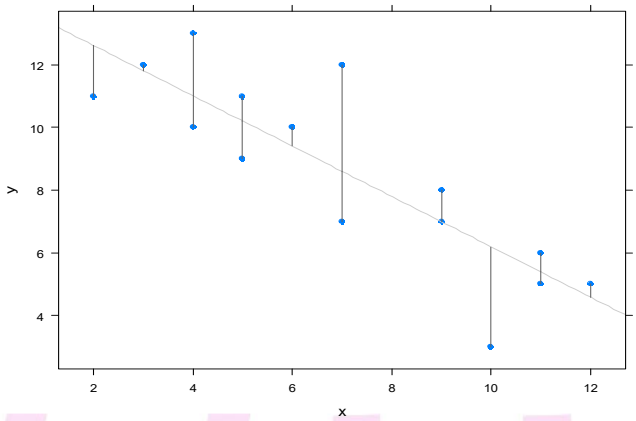
El diagrama de dispersión de ambas variables y la recta de regresión encontrada se pueden ver el en siguiente grafico



Si utilizamos dicha recta en todos los valores de X observados, tenemos los valores predichos(pronósticos) $\hat{y} = 14.2197 - 0.8028x$ para los valores de X observados y si restamos dichos valores a cada valor de Y tenemos los residuales $e = y - \hat{y}$. De manera conjunta se pueden alojar en la siguiente tabla de datos

núcleos	tiempo	predicho	residual
2	11	12.6141	-1.6141
3	12	11.8113	0.1887
4	10	11.0085	-1.0085
4	13	11.0085	1.9915
5	11	10.2056	0.7944
5	9	10.2056	-1.2056
6	10	9.40285	0.5972
7	7	8.6000	-1.6000
7	12	8.6000	3.4000
9	8	6.9944	1.0056
9	7	6.9944	0.0056
10	3	6.1915	-3.1915
11	6	5.3887	0.6113
11	5	5.3887	-0.3887
12	5	4.5859	0.4141

Los residuales son las distancias verticales (respecto al eje Y) para cada valor de (X,Y) observado a la recta de regresión obtenida

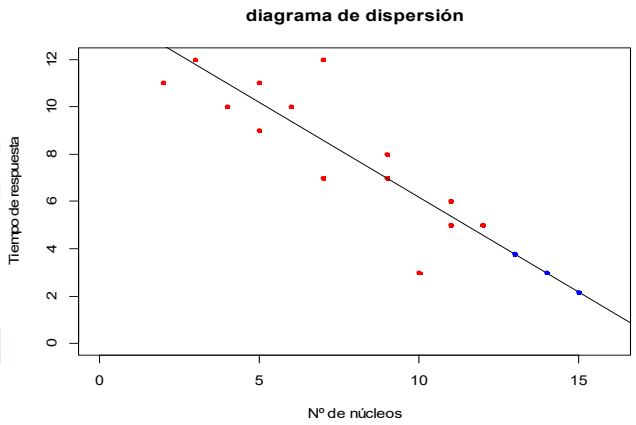


Si se desea realizar una predicción de Y en ciertos valores de X, podemos usar el siguiente comando

predict (reg1 ,data.frame(x=c(13,14,15)),interval ="prediction")

	fit	lwr	upr
1	3.783099	-0.3521396	7.918337
2	2.980282	-1.2959338	7.256497
3	2.177465	-2.2558498	6.610779

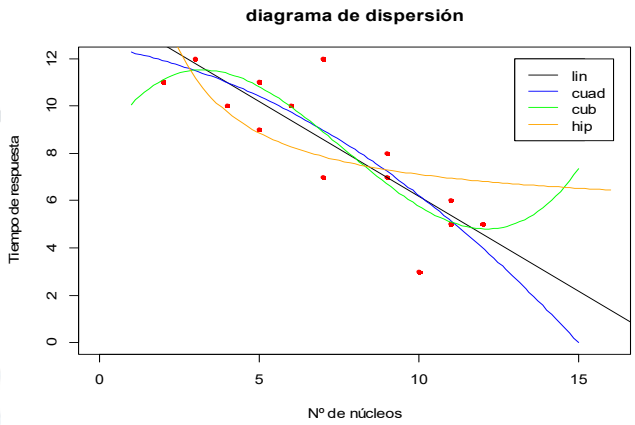
La representación gráfica de los valores de X e Y, su recta de regresión y pronósticos en tres valores de X, pueden verse en el siguiente grafico



Para los ajustes que restan, anotaremos de forma esquemática sus coeficientes de regresión y bondad de ajuste

orden	modelo	coeficientes	R^2	Var. Res.
2	cuadrático	a=12.50818 b=-0.21224 c=-0.04143	0.7303	1.693
3	cúbico	a=8.24984 b=-2.2250 c=-0.4370 d=0.01898	0.7523	1.695
4	hipérbola	a=5.352 b=17.568	0.4358	2.353
5	logaritmo	a=16.8205 b=-4.4930	0.6192	1.933

aunque puede realizarse en la práctica el mismo estudio que para el primer modelo.
Para finalizar representamos el diagrama de dispersión de nuestros datos con las gráficas de los cuatro primeros modelos ajustados



EJERCICIOS:

1.- En una multinacional, dedicada a la fabricación de ordenadores, se están probando varios prototipos de arquitectura multinúcleo. Tras tomar los tiempos de respuesta (Y), en millonésimas de segundo, ante un código de prueba, el equipo de desarrollo sospecha que a medida que aumenta el número de procesadores (X) se reducen los tiempos de cómputo:

X	2	3	4	4	5	5	6	7	7	9	9	10	11	11	12
Y	11	12	10	13	11	9	10	7	12	8	7	3	6	5	5

Asesora al equipo de desarrollo confirmándoles que:
a) la sospecha es cierta, y
b) prediciendo el tiempo de cómputo para un prototipo con 13 núcleos.

2.- Se miden dos características (X,Y) a una muestra de 25 individuos. Los datos obtenidos son:

X	3	6	7	4	5	4	6	8	9	9	7	6	5	6	6	4	3	2	7	8	9	3	5	7	8
Y	9	5	6	7	5	8	4	3	4	2	3	3	4	5	4	4	6	8	4	3	2	8	6	2	2

a) Hallar la media aritmética para X y para Y.
b) Hallar el coeficiente de correlación e interpretarlo.
c) Determinar las rectas de regresión de Y sobre X y de X sobre Y.

3.- Se miden pesos (en Kgr.) y alturas (en cm.) a una muestra de 10 alumnos. Los datos obtenidos son:

pesos	56,5	58,25	60,5	67	70	74,5	68	65	62	75
alturas	162	164	168	172	172	178	174	177	173	184

a) Hallar el coeficiente de correlación lineal e interpretarlo.
b) ¿Cuánto pesará un alumno que mida 185 cm.?
c) ¿Cuál será la altura de un alumno cuyo peso sea igual a 65 kgr??

4. Para el conjunto de datos GaltonFamilies (librería Histdata), creado sobre 1886, donde se recogen las alturas de los padres y de los hijos, y Galton F. aplicó por primera vez el concepto de regresión, que denominó "regresión hacia la media"; se solicitan las siguientes cuestiones:
a) Hallar la recta de regresión de la altura del hijo respecto a la altura del padre.
b) Hallar la recta de regresión de la altura del hijo respecto a la altura de la madre.
c) Hallar la recta de regresión de la altura del hijo respecto a la altura promedio de los padres.
d) Realizar diagramas de dispersión para estas cuatro variables.
e) Hallar la recta de regresión de la altura del hijo respecto a la altura promedio de los padres, para cada género (masculino, femenino) por separado. ¿se obtienen mejores resultados que en el apartado c)?

5. Para el conjunto de datos **iris** (librería Datasets, don se recogen las anchuras y longitudes de los pétalos de tres tipos de lirios
a) Realizar en un solo gráfico, los diagramas de dispersión de las primeras cuatro variables.
b) Realizar un ajuste lineal entre las dos variables más correladas.
c) Este ajuste lineal entre las variables se mantiene si consideramos cada grupo por separado.