

Práctica 3

Alejandro Cáceres
UPC - Statistics 2019/2020

Objetivo

Regresión lineal

- ▶ Ajuste por mínimos cuadrados
- ▶ Predicción

Regresión lineal

El ozono es tóxico para la salud, complica casos de asma y de obstrucción pulmonar crónica

Si trabajas en la AEMET y tienes que predecir los días que pueden haber niveles altos de ozono para que la agencia de salud pública envíe una alarma de contaminación alta; que predictor usaras, la temperatura o la radiación solar?

A qué temperatura activarías la alarma de contaminación por ozono (niveles mayores de 60)?

cargar y guardar datos

La función **read.table** lee un un fichero de texto en un data.table

```
a <- read.table(file="data.txt", sep=";",  
header=TRUE, na.string="NA")
```

- ▶ file="data.txt" es el nombre del fichero en comillas
- ▶ sep=";" es la separación de los datos por ;
- ▶ header=TRUE lee la primer fila como los nombres de las variables
- ▶ na.string="NA" codifica los missings como NA

cargar datos

Hemos cargado en la variable **a** los datos de airquality que habíamos almacenado en un fichero de texto

```
> head(a)
```

	Ozone	Solar.R	Wind	Temp	Month	Day
1	41	190	7.4	67	5	1
2	36	118	8.0	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4
7	23	299	8.6	65	5	7
8	19	99	13.8	59	5	8

Ozone

Convirtamos ozono en una variable categórica

dozone

```
> a$Ozone
> n <- min(a$Ozone, na.rm=TRUE)
> x <- max(a$Ozone, na.rm=TRUE)
> s <- seq(n, x , length=10)
> dozone <- cut(a$Ozone,breaks=s,right=FALSE)
> dozone
[1] [38.1,56.7) [19.6,38.1) [1,19.6)      [1,
[6] [19.6,38.1) [19.6,38.1) [1,19.6)      [1,
> head(a$Ozone)
[1] 41 36 12 18 NA 28
```

a cada valor de ozono se le asigna un intervalo.

Recordemos las tablas de frecuencia y apliquemoslas a **dozone**

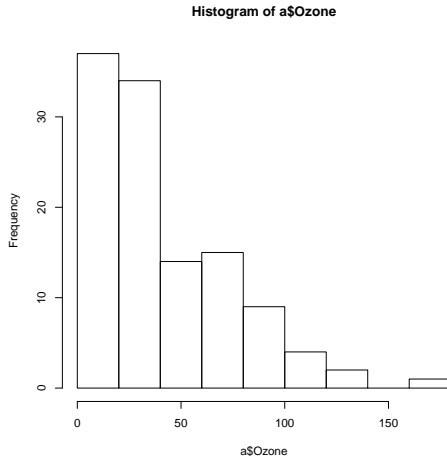
```
> ni <- table(dozone)
> fi <- prop.table(ni)
> Ni <- cumsum(ni)
> Fi <- cumsum(fi)
> tab <- cbind(ni,fi,Ni,Fi)
> tab
```

	ni	fi	Ni	Fi
[1,19.6)	33	0.286956522	33	0.2869565
[19.6,38.1)	35	0.304347826	68	0.5913043
[38.1,56.7)	15	0.130434783	83	0.7217391
[56.7,75.2)	11	0.095652174	94	0.8173913
[75.2,93.8)	12	0.104347826	106	0.9217391
[93.8,112)	5	0.043478261	111	0.9652174
[112,131)	3	0.026086957	114	0.9913043
[131,149)	1	0.008695652	115	1.0000000
[149,168)	0	0.000000000	115	1.0000000

Ozone Vs Solar.R

Recordemos el histograma

```
> hist(a$Ozone)
```



El histograma es un gráfico de ni!

Ozone Vs Solar.R

Vamos a explorar la relación entre Ozone y las otras variables (viento, radiación solar y temperatura)

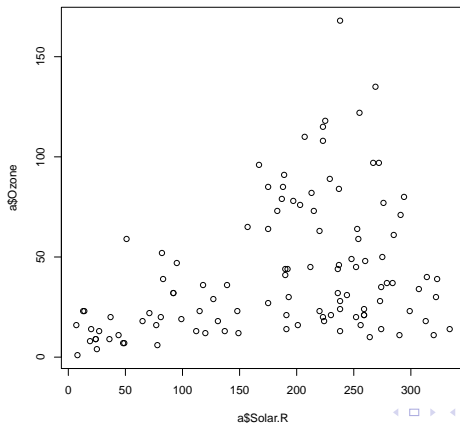
Usa la función **plot** para las variables **a\$Solar.R** y **a\$Ozone** para ver su dependencia funcional

Ozone Vs Solar.R

```
> plot(a$Solar.R, a$Ozone)
```

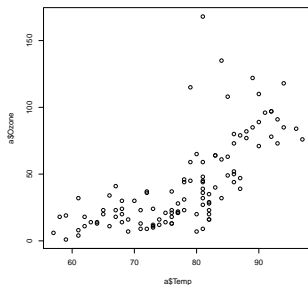
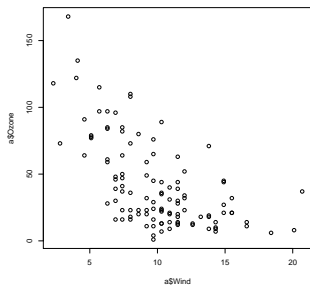
o

```
> plot(a$Ozone~a$Solar.R)
```



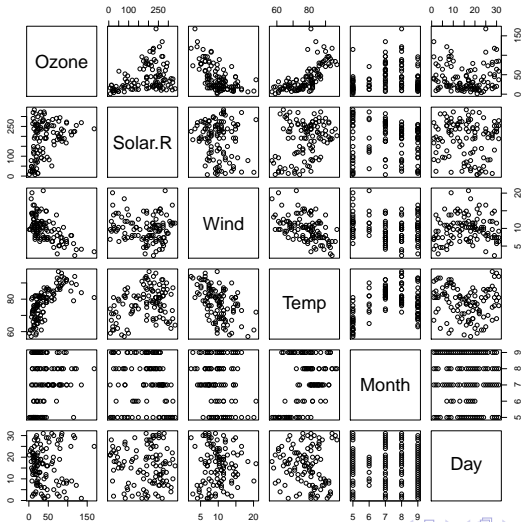
Ozone Vs Solar.R

Cómo son los gráficos para **a\$Ozone** como función de **a\$Wind** y **a\$Temp**?



Ozone Vs Solar.R

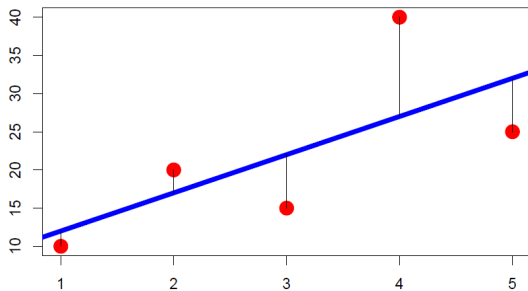
Usa la función **pairs** sobre la variable **a**
que ves?



Regresión lineal

Dados los datos de dos variables, queremos saber si su relación se puede describir por una línea recta.

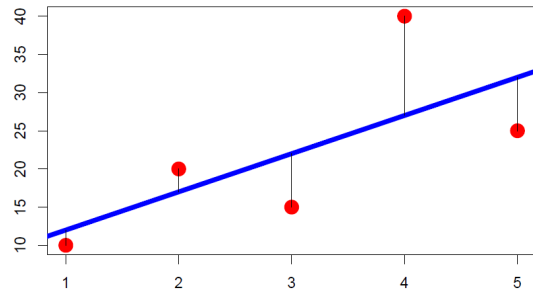
Como hay variación (error) entre las medidas, cuál sería una buena línea recta que describa esa relación?



Regresión líneal

para cada valor de x_i queremos encontrar un valor nuevo \hat{y}_i tal que

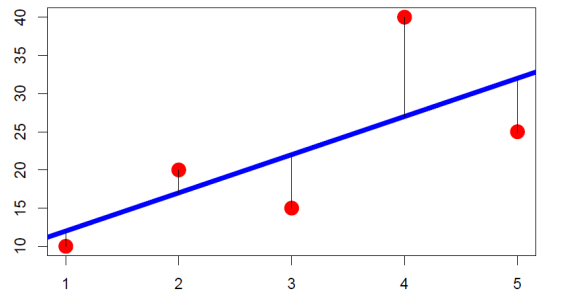
$$\hat{y}_i = mx_i + b$$



o sea queremos determinar m y b

Regresión lineal

Podemos encontrar m y b tal que la suma de las distancias de los datos a los puntos sea mínima



$$Q = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (mx_i + b))^2$$

Regresión líneal

Derivando con respecto a m y b

$$Q = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (mx_i + b))^2$$

e igualando a cero obtenemos

$$m = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

y

$$b = \bar{y} - m\bar{x}$$

Regresión lineal

calculemos m y b para la relación entre **a\$Solar.R**
y **a\$Ozone**

Regresión lineal

Para facilitar los cálculos vamos quitar todos los "NA" de **a**, usando la función **complete.cases**

```
> select <- complete.cases(a)
> head(select)
[1] TRUE TRUE TRUE TRUE TRUE TRUE
> a <- a[select,]
> head(a)
```

	Ozone	Solar.R	Wind	Temp	Month	Day
1	41	190	7.4	67	5	1
2	36	118	8.0	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4
7	23	299	8.6	65	5	7

Regresión lineal

calculemos m y b para la relación entre **a\$Solar.R** y **a\$Ozone**

- ▶ asignemos **x** a **a\$Solar.R** y **y** a **a\$Ozone**
- ▶ asignemos **xbar** a la media de **x** y **ybar** a la media de **y**
- ▶ usando **sum** asignemos a **m** la cantidad

$$\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- ▶ asignemos **b** a

$$\bar{y} - m\bar{x}$$

Regresión lineal

```
x <- a$Solar.R
```

```
y <- a$Ozone
```

```
xbar <- mean(x)
```

```
ybar <- mean(y)
```

```
m <- sum((x-xbar)*(y-ybar))/sum((x-xbar)^2)
```

```
b <- ybar-m*xbar
```

```
> m
```

```
[1] 0.1271653
```

```
> b
```

```
[1] 18.59873
```

Regresión lineal

Vamos a definir una función que prediga el valor de **y** en $x=20$ o en cada uno de los valores de **x** (**ypred**)

```
> ypred <- function (x) {m*x+b}
```

```
> ypred(0)
```

```
[1] 18.59873
```

```
> ypred(20)
```

```
[1] 21.14203
```

```
> ypred(x)
```

```
[1] 42.76013 33.60423 37.54635 58.40146 56.6
```

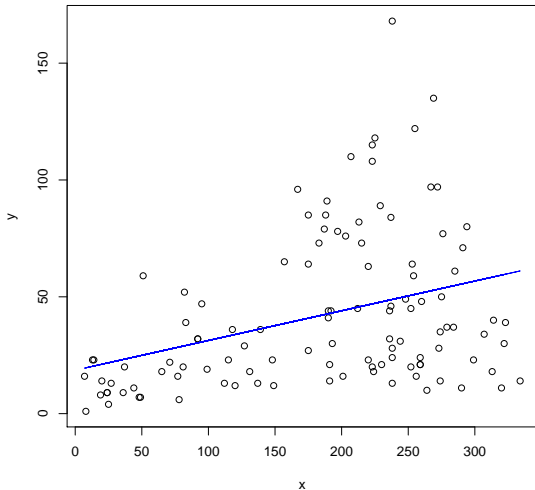
la función es vectorial!

Regresión lineal

Queremos comparar nuestros datos con la recta $mx+b$

```
> plot(y ~ x)
> lines(x, ypred(x), col="blue")
```

Regresión líneal



Regresión lineal

Qué tanto se asemeja la relación entre x e y a una recta?

Queremos una medida $R^2 > 0$

- ▶ $R^2 = 1$ es una recta perfecta!
- ▶ $R^2 = 0$ es lo mas lejano a una recta!
(qué es lo mas lejano a una recta?)

Regresión lineal

- ▶ $\sum_{i=1..n}(y_i - \bar{y})^2$ mide cuanto varían los datos y_i con respecto a la media
- ▶ $\sum_{i=1..n}(ypred_i - \bar{y})^2$ mide cuanto varían las **predicciones** de la línea respecto a la media

cuanto porcentaje de variación de los datos es explicado por variación de la recta?

Regresión líneal

Coefficiente de variación

$$R^2 = \frac{\sum_{i=1..n}(ypred_i - \bar{y})^2}{\sum_{i=1..n}(y_i - \bar{y})^2}$$

- ▶ $R^2 = 1$ cuando los datos caen todos en una línea
- ▶ $R^2 = 0$ cuando los datos se distribuyen en un disco (y ninguna recta es mejor que otra)
- ▶ R es el coeficiente de correlación de Pearson

Regresión líneal

Calculemos

$$R^2 = \frac{\sum_{i=1..n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1..n}(y_i - \bar{y})^2}$$

para nuestros datos

donde $\hat{y}_i = ypred_i = ypred(x)$

Regresión lineal

$$R^2 = \frac{\sum_{i=1..n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1..n}(y_i - \bar{y})^2}$$

```
R2 <- sum((ypred(x)-ybar)^2)/sum((y-ybar)^2)
> R2
[1] 0.1213419
```

El 12% de la variación de los datos es explicado por una dependencia lineal (subyacente) entre x e y

Regresión lineal

Todo esto (y algo más) se hace con la función **lm** de R

```
#usando el data.frame a (parametro data)  
mod <- lm(Ozone ~ Solar.R, data=a)
```

```
#usando las variables x e y  
mod <- lm(y ~ x)
```

Regresión lineal

Todo esto (y algo mas) se hace con la función **lm** de R. Hay que usar la función **summary** sobre **mod**

```
> summary(mod)
```

Call:

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

Residuals:

	Min	1Q	Median	3Q	Max
	-48.292	-21.361	-8.864	16.373	119.136

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	18.59873	6.74790	2.756	0.006856	**
x	0.12717	0.03278	3.880	0.000179	***

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 31.33 on 109 degrees of freedom

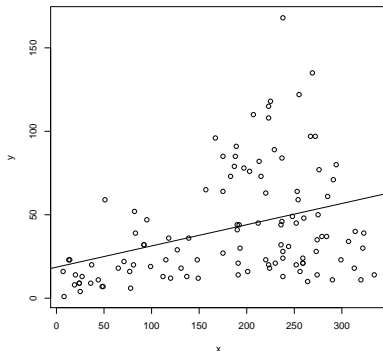
Multiple R-squared: 0.1213, Adjusted R-squared: 0.1133

F-statistic: 15.05 on 1 and 109 DF, p-value: 0.0001793

Regresión líneal

El gráfico de la línea se recupera con la función **abline** sobre el resulatdo de **lm**

```
plot(x, y)  
abline(mod)
```



Regresión lineal

Y los valores de la producción sobre la línea recta se obtienen con la función **predict**.

La predicción del ozono para una radiación solar de 20 es:

```
> predict(mod, data.frame(x=20))  
      1  
21.14203
```

que clase tiene mod?

Regresión lineal

nota: la clase de mod es una nueva estructura `lm`, generada por la función **`lm`**

```
> class(mod)
[1] "lm"
```

`predict` es una función sobre la estructura `lm`. Qué genera `predict`?

Observación:

`vector` \rightarrow `data.frame(vector)` \rightarrow `lm(data.frame)` \rightarrow
`predict(lm)` \rightarrow `vector`

Regresión exponencial

Estudiamos ahora la dependencia de **a\$Ozone** con **a\$Temp**

- ▶ hagamos un gráfico de Ozone Vs Temp y otro de $\log(\text{Ozone})$ Vs temp
- ▶ usa la función **lm** para Ozone Vs Temp y otra para $\log(\text{Ozone})$ Vs Temp
- ▶ pinta el gráfico para cada una con su recta de regresión
- ▶ Cuál regresión explica mas variabilidad?
- ▶ Cuál es mejor predictor del ozono, la radiación solar o la temperatura?

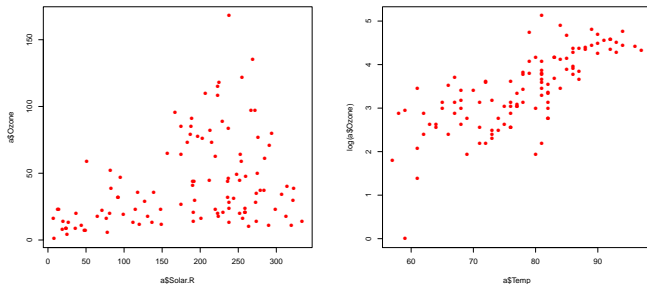
Regresión exponencial

El ozono es tóxico para la salud, complica casos de asma y de obstrucción pulmonar crónica

Si trabajas en la AEMET y tienes que predecir los días que pueden haber niveles altos de ozono para que la agencia de salud pública envíe una alarma de contaminación alta; que predictor usaras, la temperatura o la radiación solar?

A qué temperatura activarías la alarma de contaminación por ozono (niveles mayores de 60)?

Regresión líneal



A 80F (26C) la mayoría de días supera los límites de ozono ($Oz=60, \log(Oz)=4.09$).

Por qué hay días de bajo ozono con alta radiación solar?