La recta de regresión

Análisis de la varianza
El coeficiente de correlación lineal de Pearson

Grado en biología sanitaria

M. Marvá

e-mail: marcos.marva@uah.es

Unidad docente de Matemáticas, Universidad de Alcalá

15 de septiembre de 2017

El problema

Disponemos de una muestra de (pares de) valores

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

calculamos la recta

$$y(x) = b_1 x + b_0$$

que minimiza el ECM para:

- Medir el **grado de asociación** entre las dos variables
- **Estimar** valores de Y no observados (NO extrapolar)
- Si los x_i son diferentes entre sí **existe la recta de regresión**:

¿es siempre útil?

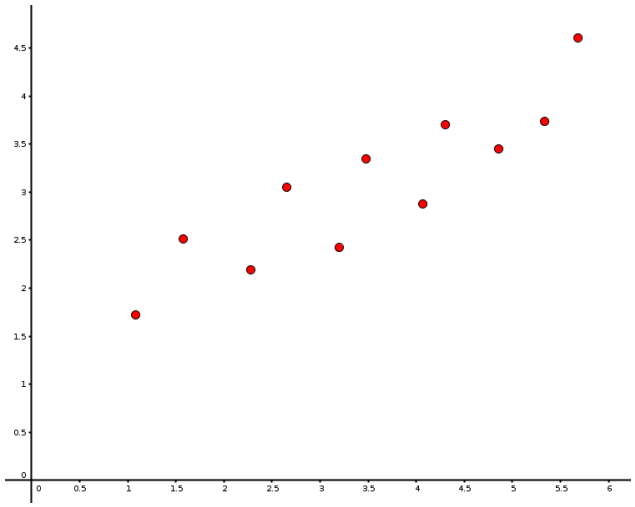
Si $Y \sim X$,

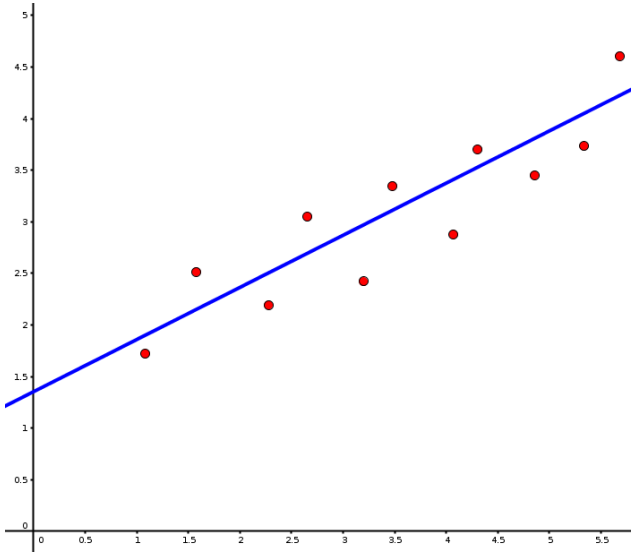
- variaciones en X deben traducirse en variaciones en Y
- explicar las variaciones en Y a partir de las de X

El problema 1

El diagrama de dispersión marca una tendencia (cuando X crece, y crece)

- ¿Se debe al azar?
- ¿Hay un modelo $y \sim b_0 + b_1 x$ que describe cómo varía Y con X?



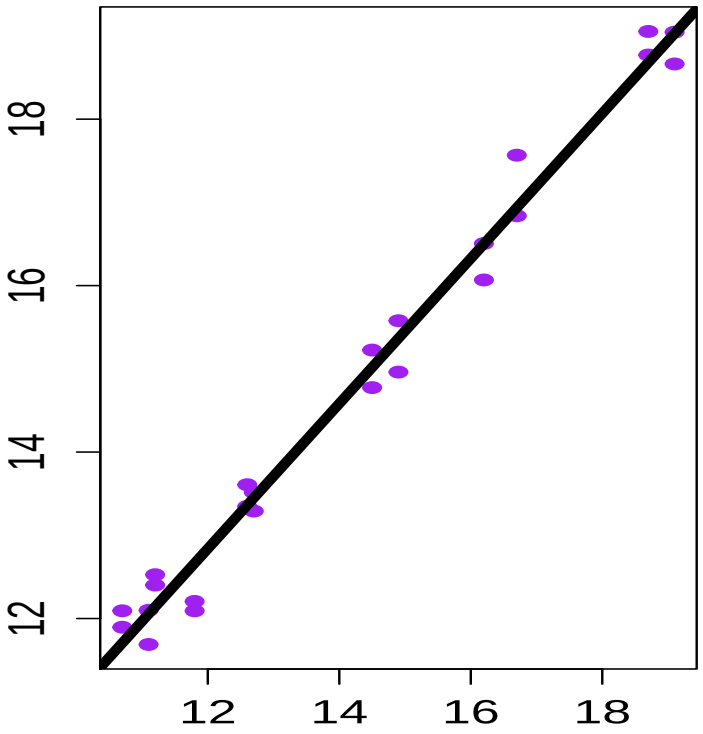


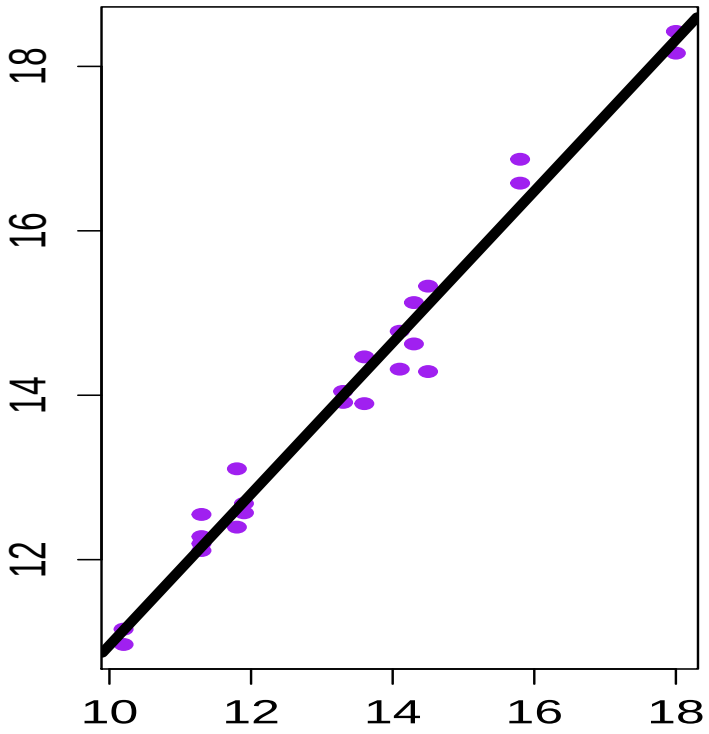
Sección 10.2 del libro

El problema 2

Dados dos conjuntos de datos

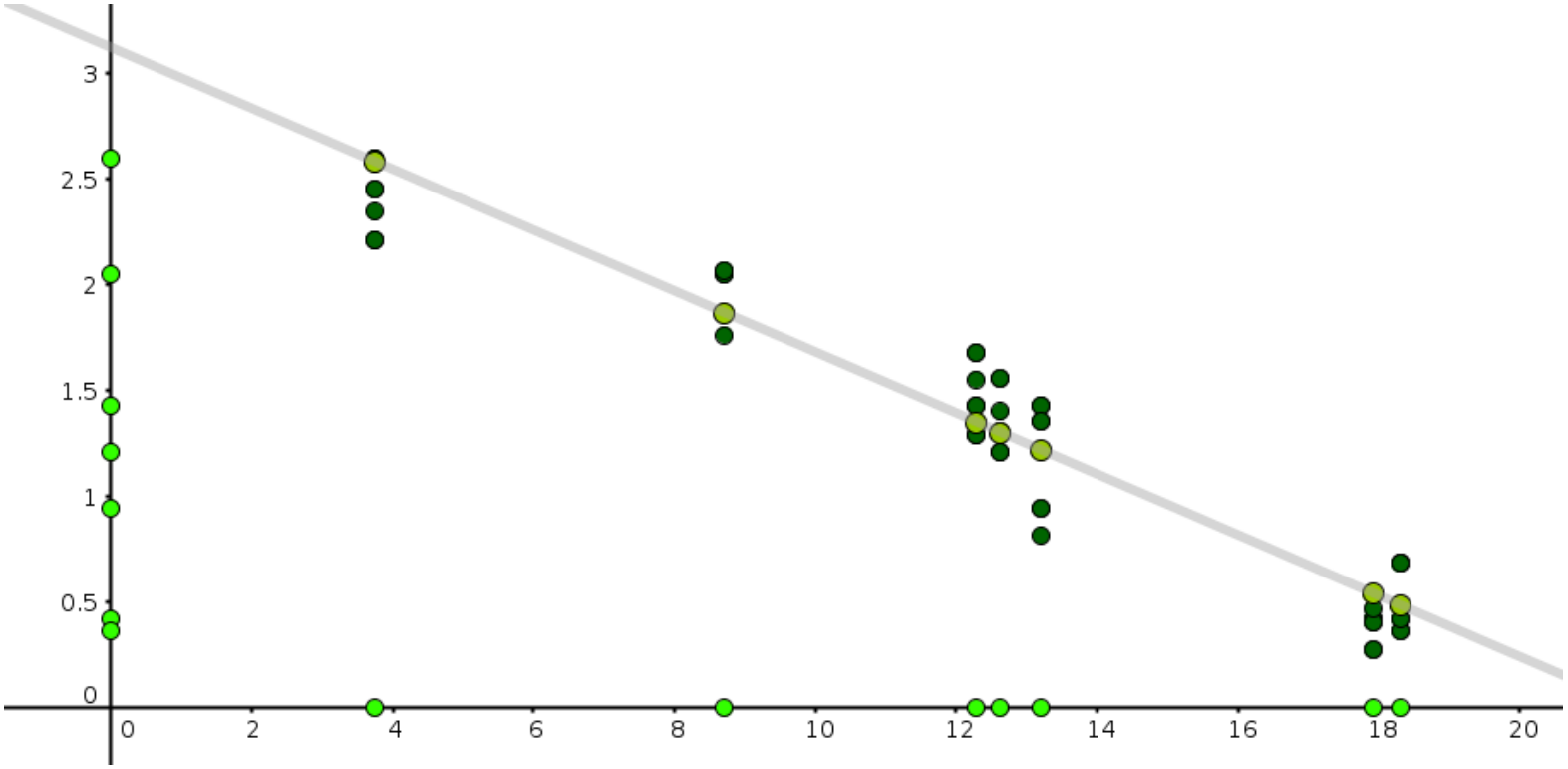
¿qué recta aproxima mejor su nube de puntos?





ANOVA: variabilidad debida al **azar**

Cada individuo es diferente: por eso para valores fijos de X se pueden observar distintas medidas de Y.



variabilidad de Y: usar el EC:

$$\sum(\hat{y}_j - y_j)^2$$

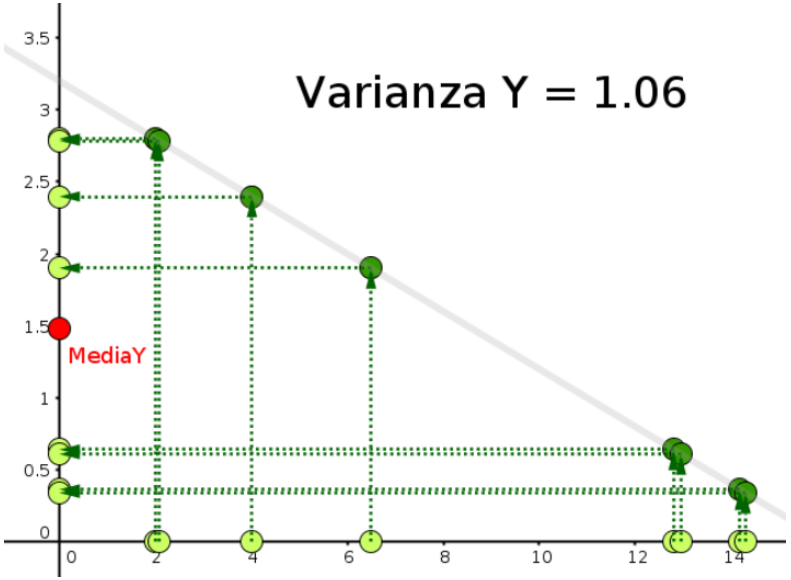
La variable x está presenta a través de $\hat{y}$

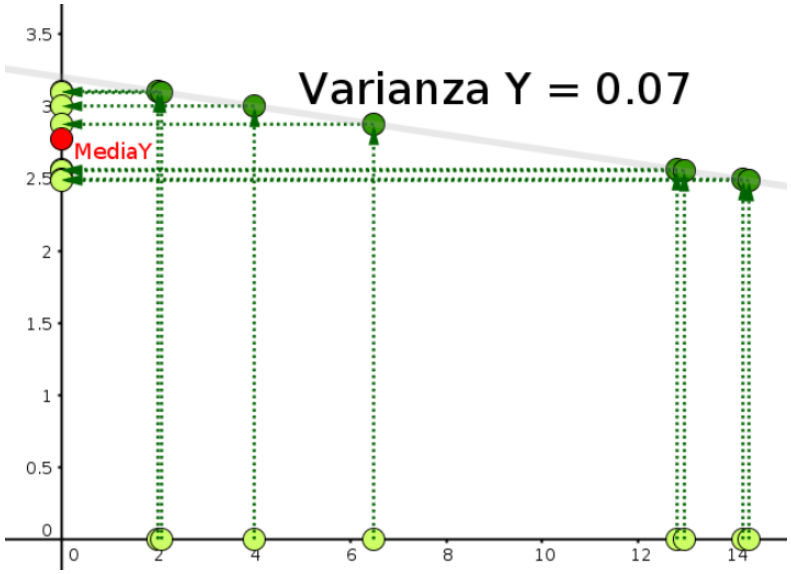
ANOVA: variabilidad debida al **modelo**

Si existe relación lineal entre dos variables

$$y(x) = b_1 x + b_0$$

la recta traslada la variabilidad de X a la variable Y





Variabilidad en Y: diferencia entre la media de lo "predicho por la recta" y cada "predicción":

$$\sum (\hat{y}_j - \bar{y})^2$$

La variable x está presenta a través de $\hat{y}$

La identidad ANOVA

Reagrupando términos, se verifica la identidad Anova (**An**alysis **o**f **va**riance)

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{Error total}} = \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{\text{EC}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{\text{Error modelo (recta)}}$$

Notación

- Dispersión total de Y o error total (SS_{total}): $\sum_{j=1}^{n}(y_j-\bar{y})^2$
- Dispersión debida al azar: EC (SS_{residual}): $\sum_{j=1}^{n}(y_j-\hat{y}_j)^2$
  - El EC$=0$ si los puntos están exáctamente alineados
  - El EC aumenta conforme los puntos se “desalinean”
- Dispersión explicada por la regresión (SS_{modelo}): $\sum_{j=1}^{n}(\hat{y}_j-\bar{y})^2$

La identidad Anova

La identidad Anova reparte la variabilidad total (SStotal) entre

- la parte debida al modelo (SSmodelo)
- y la parte debida al azar (SSresidual).

Sin embargo

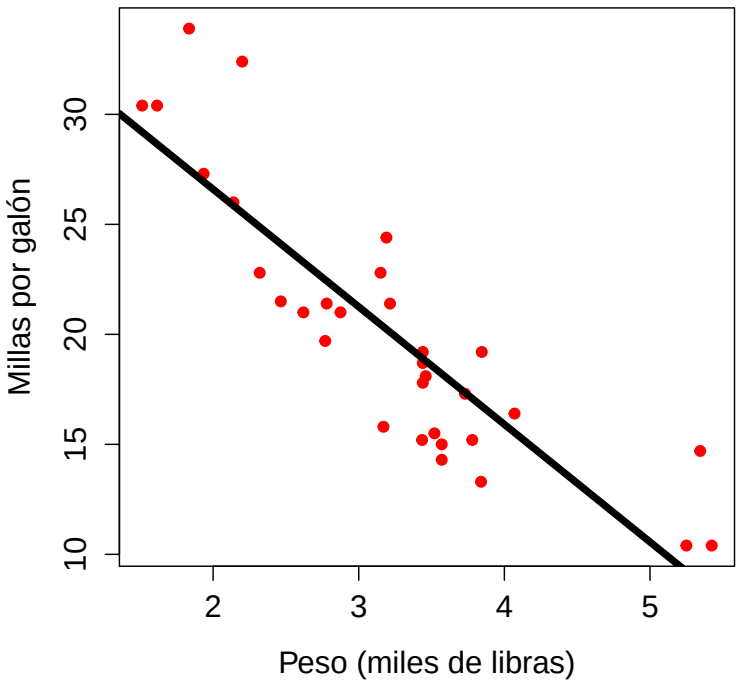
La identidad Anova es sensible a las unidades y a cambios de escala

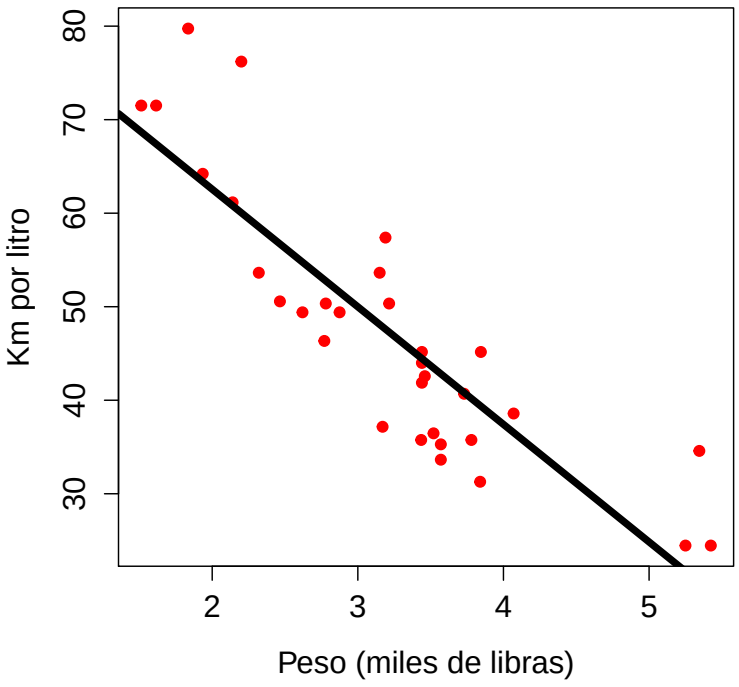
Considera los datos `mtcars` de la librería `car` de R:

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb	kpl
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4	49.40
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4	49.40
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1	53.63
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1	50.34
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2	43.99
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1	42.57
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4	33.64
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2	57.39
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2	53.63
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4	45.16

¿Relación entre el peso del vehículo y la distancia recorrida con cierta cantidad de combustible?

Identidad Anova y cambios de escala





Recta regresión $mpg(peso) = -5.3 * peso + 37.3,$ datos

Recta regresión $kpl(peso) = -12.6 * peso + 87.7,$ datos

Identidad Anova y cambios de escala

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	peso (miles libras)	Millas por galón	mpg esp	SS_total	SS_res	SS_mod		peso (miles libras)	Km por litro	Kpl esp	SS_total	SS_res	SS_mod
2	2,62	21	23,28	0,83	5,21	10,19		2,62	49,4	54,77	859,04	28,79	1202,35
3	2,875	21	21,92	0,83	0,85	3,35		2,875	49,4	51,56	859,04	4,66	990,30
4	2,32	22,8	24,89	7,34	4,35	23,00		2,32	53,63	58,54	1124,89	24,08	1478,13
5	3,215	21,4	20,10	1,71	1,68	0,00		3,215	50,34	47,29	915,02	9,33	739,54
6	3,44	18,7	18,90	1,93	0,04	1,42		3,44	43,99	44,46	571,18	0,22	593,69
7	3,46	18,1	18,79	3,96	0,48	1,68		3,46	42,57	44,21	505,32	2,67	581,50
8	3,57	14,3	18,21	33,53	15,25	3,55		3,57	33,64	42,82	183,59	84,31	516,72
9	3,19	24,4	20,24	18,57	17,34	0,02		3,19	57,39	47,60	1391,24	95,85	756,74
10	3,15	22,8	20,45	7,34	5,52	0,13		3,15	53,63	48,10	1124,89	30,55	784,66
11	3,44	19,2	18,90	0,79	0,09	1,42		3,44	45,16	44,46	628,47	0,49	593,69
12	3,44	17,8	18,90	5,25	1,21	1,42		3,44	41,87	44,46	474,34	6,69	593,69
13	4,07	16,4	15,53	13,62	0,75	20,77		4,07	38,58	36,54	341,86	4,18	270,45
14	3,73	17,3	17,35	7,79	0,00	7,51		3,73	40,69	40,81	424,33	0,01	429,32
15	3,78	15,2	17,08	23,92	3,55	9,05		3,78	35,75	40,18	245,22	19,64	403,66
16	5,25	10,4	9,23	93,91	1,38	118,03		5,25	24,46	21,70	19,09	7,61	2,59
17	5,424	10,4	8,30	93,91	4,42	139,10		5,424	24,46	19,51	19,09	24,47	0,33
18	5,345	14,7	8,72	29,06	35,77	129,32		5,345	34,58	20,51	209,94	198,05	0,17
19	2,2	32,4	25,53	151,52	47,23	29,56		2,2	76,21	60,05	3149,38	261,28	1596,41
20	1,615	30,4	28,65	106,28	3,05	73,33		1,615	71,51	67,40	2643,95	16,89	2238,21
21	1,835	33,9	27,48	190,70	41,24	54,57		1,835	79,74	64,63	3558,05	228,18	1984,16
22	2,465	21,5	24,11	1,99	6,82	16,16		2,465	50,57	56,71	928,99	37,75	1341,28
23	3,52	15,5	18,47	21,07	8,84	2,62		3,52	36,46	43,45	267,96	48,87	545,69
24	3,435	15,2	18,93	23,92	13,89	1,35		3,435	35,75	44,52	245,22	76,90	596,76
25	3,84	13,3	16,76	46,11	11,99	11,08		3,84	31,28	39,43	125,20	66,38	373,92
26	3,845	19,2	16,74	0,79	6,07	11,26		3,845	45,16	39,36	628,47	33,58	371,49
27	1,935	27,3	26,94	51,98	0,13	46,96		1,935	64,21	63,38	1946,52	0,69	1873,74
28	2,14	26	25,85	34,92	0,02	33,15		2,14	61,16	60,80	1686,69	0,13	1657,26
29	1,513	30,4	29,20	106,28	1,44	82,96		1,513	71,51	68,68	2643,95	7,99	2361,19
30	3,17	15,8	20,34	18,41	20,64	0,06		3,17	37,16	47,85	291,36	114,30	770,63
31	2,77	19,7	22,48	0,15	7,73	5,71		2,77	46,34	52,88	689,03	42,77	1075,12
32	3,57	15	18,21	25,91	10,27	3,55		3,57	35,28	42,82	230,72	56,88	516,72
33	2,78	21,4	22,43	1,71	1,06	5,46		2,78	50,34	52,75	915,02	5,83	1066,89
34	Media =	20,09		1126,05	278,32	847,73		Media =	47,26		29847,09	1540,06	28307,03
35					1126,05							29847,09	

Coeficiente de correlación

La identidad Anova depende de las unidades: no podríamos comparar

La identidad Anova adimensionalizada

$$1 = \frac{\text{EC}}{\text{Variabilidad total}} + \frac{\text{Variabilidad modelo}}{\text{Variabilidad total}}$$

manipulando términos

$$1 = \frac{ECM}{s^2(y)} + \left(\frac{\text{Cov}(x,y)}{s(x)\cdot s(y)}\right)^2$$

donde $\text{Cov}(x,y)$, $s(x)$ y $s(y)$ son la covarianza y las varianza **muestrales** y

$$r = \frac{\text{Cov}(x,y)}{s(x)\cdot s(y)}$$

es el **coeficiente de correlación lineal de Pearson**

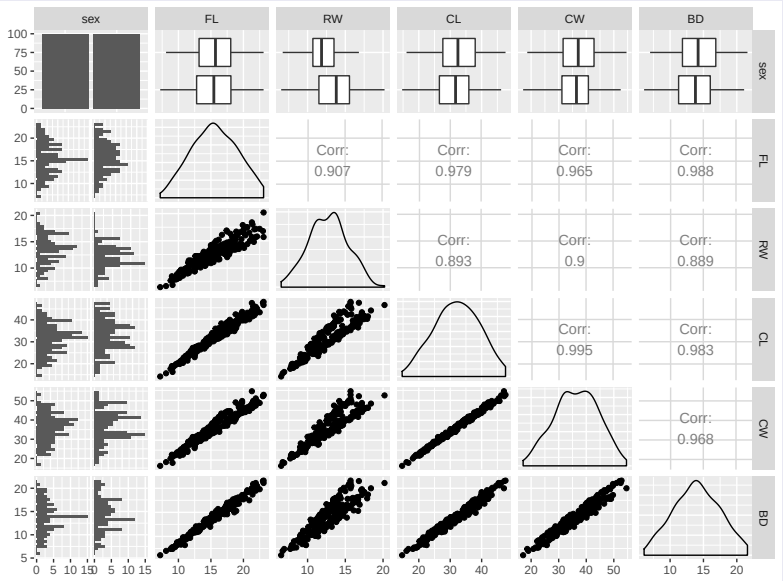
Coeficiente de correlación

Ejercicio: descarga los datos adjuntos a la presentación el consumo en millas por galón frente a kilómetros por litro. Calcula el coeficiente de correlación de Pearson; en ambos casos vale $r = -0.8677$

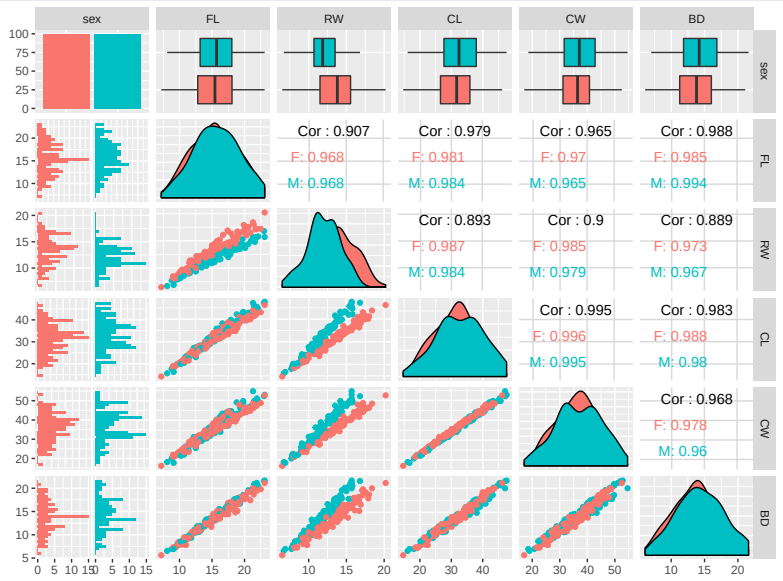
Pista Si los datos están en `vector1` y `vector2` , la fórmula en `R` es

$$\frac{\text{cov}(\texttt{vector1},\texttt{vector2})}{\text{sd}(\texttt{vector1})*\text{sd}(\texttt{vector2})}$$

Ejemplo: datos crabs de la librería MASS:



ggpairs(data = crabs[, c(2, 4:8)])



ggpairs(data = crabs[, c(2, 4:8)], mapping = aes(color = sex))

script completo

Coeficiente de correlación

$$1 = \frac{\text{ECM}}{s^2(y)} + r^2$$

- $-1 < r < 1$
- Cuanto más cerca de 0 está r^2 peor es el ajuste (tiene menos peso la parte de modelo y más el azar)
- En el caso extremo de $r^2 = 1$, entonces EC $= 0$ y los valores observados están sobre la recta de regresión
- Si el ajuste es bueno, el valor de r^2 está cerca de 1 (ojo orden r^2 y r)
- La interpretación contraria puede no ser cierta (Anscombe!)
- El signo de r (es el signo de Cov(x,y)) es el de b_1
- Se dice que dos variables están fuertemente correlacionadas si el correspondiente $r^2 \approx 1$

La correlación no implica causalidad