UNIVERSIDAD NACIONAL DE COLOMBIA

Sede-Medellin

Facultad de Ciencias

Estadistica No Paramétrica

Estadistica

Tarea 2

Juan David Garcia Zapata
Juan Diego Espinosa Hernandez
Katerin Gomez Castrillon
Andrés Camilo Usuga Montoya

2024

Punto 1

Se desea ver si la temperatura en la ciudad 1 es superior a la temperatura en la ciudad 2, las temperaturas tomadas en las dos ciudades, en el verano, son las siguientes:

Ciudad 1	83	89	89	90	91	91	92	94	96
Ciudad 2	77	78	79	80	81	81	81	82	

use un $\alpha = 0{,}05$

Respuesta Para este caso se denotara "x" como la ciudad 1 y "y" como la ciudad 2.

x	y	Rango
	77	1
	78	2
	79	3
	80	4
	81	6
	81	6
	81	6
	82	8
83		9
89		10.5
89		10.5
90		12
91		13.5
91		13.5
92		15
94		16
96		17

Recordemos que para definir el rango cuando hay valores que se repiten se suman las posiciones correspondientes y se divide entre la cantidad de veces que se repita el valor, luego este resultado se le pone en su rango a todos los datos implicados.

Retomando tenemos que: n=9, m=8, N= 9+8=17

De los 17 datos en total hay tres grupos de temperaturas empatadas lo cual reune un total de 7 observaciones, dada la cantidad de datos repetidos se usa el siguiente estadístico de prueba para el siguiente juego de hipótesis.

H_0 : La temperatura en la ciudad 1 no es superior a la temperatura en la ciudad 2.
H_1 : La temperatura en la ciudad 1 es supertior a la temperatura en la ciudad 2.

Es decir:

$$H_0 : E[X] \leq E[Y] \; vs \; H_1 : E[X] > E[Y]$$

Estadístico de prueba:

$$\sum_{i=1}^{n} R_i(x) = \sum_{i=1}^{9} R_i(x) = 117$$

$$\sum_{i=1}^{17} R_i^2 = 1782$$

Sin factor de corrección

$$T_1 = \frac{T - n\frac{N+1}{2}}{\sqrt{\frac{nm}{N(N-1)}\sum_{i=1}^{N} R_i^2 - \frac{nm(N+1)^2}{4(N-1)}}}$$

$$= \frac{117 - 9\frac{17+1}{2}}{\sqrt{\frac{9*8}{17(17-1)}1782 - \frac{9*8(17+1)^2}{4(17-1)}}} = \frac{36}{10{,}35402} = 3{,}47690$$

Valor P: P(Z >3.47690) = 0.000253

Con factor de correción

$$T_1 = \frac{T - n\frac{N+1}{2} - \frac{1}{2}}{\sqrt{\frac{nm}{N(N-1)}\sum_{i=1}^{N} R_i^2 - \frac{nm(N+1)^2}{4(N-1)}}}$$

$$= \frac{117 - 9\frac{17+1}{2} - \frac{1}{2}}{\sqrt{\frac{9*8}{17(17-1)}1782 - \frac{9*8(17+1)^2}{4(17-1)}}} = \frac{35{,}5}{10{,}35402} = 3{,}42861$$

Valor P: P(Z >3.42861) = 0.000303

Como el valor P es inferior a 0.05 rechazamos la hipótesis nula y decimos que con un nivel de significancia del 5 % la temperatura en la ciudad 1 es superior a la temperatura en la ciudad 2.

En R

```
x <- c(83,89,89,90,91,91,92,94,96)
y <- c(77,78,79,80,81,81,81,82)
wilcox.test(x,y,alternative = "greater", correct = T)
```

```
## Warning in wilcox.test.default(x, y, alternative = "greater", correct = T):
## cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  x and y
## W = 72, p-value = 0.0003033
## alternative hypothesis: true location shift is greater than 0
```

Segundo Punto

Siete estudiantes aprendieron álgebra usando el método presente y seis estudiantes aprendieron álgebra usando un método nuevo, halle un intervalo de confianza del 90% para la diferencia entre los puntajes promedio de los 2 métodos, los resultados fueron:

Método presente	68	72	79	69	84	80	78
Método nuevo	64	60	68	73	72	70	

En este caso los datos consisten de dos muestras independientes, $X_1, .., X_7$ y $Y_1, ..., Y_6$; y el objetivo es calcular un intervalo de confianza para $E[X] - E[Y]$, donde X es el puntaje de los estudiantes que aprendieron con el método presente, mientras que Y es el puntaje de los alumnos que utilizaron el método nuevo.

De manera que $n = 7$, pues hay siete estudiantes que aprendieron con el método presente, y $m = 6$, ya que hay seis estudiantes que utilizaron el método nuevo.

Para la construcción del intervalo de interés, y debido a que ambas muestras tienen pocos datos, se utilizarán los valores que aparecen en la tabla A7, en la página 536 del texto "Practical nonparametric statistics", de William Jay Conover.Dicha tabla es de especial interés porque permite determinar el cuantil $\frac{\alpha}{2}, w_{\frac{\alpha}{2}}$, en donde se obtiene que , para $n = 7$, $m = 6$ y $\alpha = 0.1$,$w_{0.05} = 37$, y por lo tanto $k = 37 - \frac{7(7+1)}{2} = 9$

El siguiente paso es obtener la matriz de las $r = m * n = 7 * 6 = 42$ diferencias:

```
diferencias
```

Table 2: Tabla de diferencias entre puntajes de estudiantes

	60	64	68	70	72	73
68	8	4	0	-2	-4	-5
69	9	5	1	-1	-3	-4
72	12	8	4	2	0	-1
78	18	14	10	8	6	5
79	19	15	11	9	7	6
80	20	16	12	10	8	7
84	24	20	16	14	12	11

El siguiente paso es ordenar las diferencias de menor a mayor y asignarle un rango a cada posición.

```
tabla2_parte1
```

Table 3: Rangos de los valores asociados a las diferencias

Rango	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Diferencia	-5	-4	-4	-3	-2	-1	-1	0	0	1	2	4	4	5	5	6	6	7	7	8	8

```
tabla2_parte2
```

Table 4: Rangos de los valores asociados a las diferencias

Rango	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42
Diferencia	8	8	9	9	10	10	11	11	12	12	12	14	14	15	16	16	18	19	20	20	24

Por lo tanto, un intervalo de confianza del 90 para $E[X] - E[Y]$ está dado por:

$$(L, U) = (d^{(k)}, d^{r-k+1}) = (d^9, d^{42-9+1}) = (d^9, d^{34}) = (0, 14)$$

Ahora bien, haciéndolo en R:

```r
options(scipen = 100000)
x <- c(68, 72, 79, 69, 84, 80, 78)
y <- c(64, 60, 68, 73, 72, 70)

suppressWarnings( wilcox.test(x, y ,
alternative = c("two.sided"),
mu = 0, paired = FALSE, exact = NULL,

conf.int = TRUE, conf.level = 0.90))
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  x and y
## W = 34, p-value = 0.07335
## alternative hypothesis: true location shift is not equal to 0
## 90 percent confidence interval:
##   0.0000001449613 14.0000366351815
## sample estimates:
## difference in location
##               7.999948
```

En ambos casos, y debido a que el respectivo intervalo no contiene el cero, es posible afirmar que, con una confianza del 90%, hay evidencia muestral suficiente para sugerir que sí hay diferencias en el puntaje promedio de la asignatura de álgebra entre aquellos estudiantes que utilizaron el método de estudio presente y quienes aprendieron con el método nuevo.

Pregunta 3

Una sicóloga investiga la hipótesis de que el orden de nacimiento en una familia afecta la asertividad. Los sujetos del experimento son 20 jóvenes adultos entre 20 y 25 años de edad. Hay 7 primogénitos, 6 individuos que nacieron en segundo lugar y 7 que en tercero. Cada sujeto presenta una prueba de asertividad con los siguientes resultados. Las calicaciones altas indican una mayor asertividad. Suponga que los datos no se distribuyen de manera normal, por lo que la prueba F no puede utilizarse, pero los datos presentan al menos una escala ordinal.

Cuadro 1: Datos de los primogénitos y nacidos en segundo y tercer lugar.

Primogénito	**Nacido en segundo lugar**	**Nacido en tercer lugar**
18	18	7
8	12	19
4	3	2
21	24	30
28	22	18
32	1	5
10		14

Para abordar este problema utilizando métodos de estadística no paramétrica, una de las pruebas más comunes es la prueba de Kruskal-Wallis. Esta prueba se usa para comparar las medianas de dos o más grupos independientes cuando no se puede asumir una distribución normal de los datos, pero se sabe que los datos tienen al menos una escala ordinal. Es un equivalente no paramétrico de la ANOVA de un solo sentido

El Test de Kruskal-Wallis es una extensión para K muestras del test de Mann-Whitney. Se consideran K muestras aleatorias X_{ij} de tamaños n_i, para $i = 1, \ldots, K$ y $j = 1, \ldots, n_i$. Se asignan rangos a cada observación $R(X_{ij})$ y se calculan las sumas de rangos para cada muestra $R_i = \sum_{j=1}^{n_i} R(X_{ij})$. Bajo los supuestos de aleatoriedad, independencia y escala ordinal.

El estadístico de prueba es: dado que se evidencia empates.

$$T = \left(\frac{12}{N(N+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} \right) - 3(N+1),$$

donde $N = \sum_{i=1}^{K} n_i$ y, en ausencia de empates, $S^2 = \frac{N(N+1)}{12}$. La distribución nula de T sigue una χ^2 con $K-1$ grados de libertad si N es grande. Las hipótesis son:

$$\begin{cases} H_0: & \text{Las distribuciones de todas las } K \text{ poblaciones son idénticas.} \\ H_1: & \text{Al menos una población tiende a tener valores mayores que otra.} \end{cases}$$

La región crítica es $T > \chi^2_{1-\alpha,(K-1)}$. Para nuestro caso las hipotesis son de la forma:

$$\begin{cases} H_0: & \text{No existe diferencia en los niveles de asertividad entre los diferentes órdenes de nacimiento.} \\ H_1: & \text{Existe al menos una diferencia en los niveles de asertividad entre los diferentes órdenes de nacimiento.} \end{cases}$$

Primogénito		Nacido en segundo lugar		Nacido en tercer lugar	
Obs.	Rango	Obs.	Rango	Obs.	Rango
18	12	18	12	7	6
8	7	12	9	19	14
4	4	3	3	2	2
21	15	24	17	30	19
28	18	22	16	18	12
32	20	1	1	5	5
10	8			14	10
R_i	84		58		68
n_i	7		6		7

Cuadro 2: Rangos de las observaciones por muestra.

N=20

$$\sum_{i=1}^{K}\sum_{j=1}^{n_i} R^2(X_{ij}) = 12^2+7^2+4^2+15^2+18^2+20^2+8^2+12^2+9^2+3^2+17^2+16^2+1^2+6^214^2+2^2+19^2+12^2+5^2+10^2 = 2868$$

$$S^2 = \frac{1}{19} * [2868 - \frac{20(21)^2}{4}] = 34{,}89474$$

$$T_{cal} = \frac{1}{34{,}89474} * [\frac{84^2}{7} + \frac{58^2}{6} + \frac{68^2}{7} - \frac{20(21)^2}{4}] = 0{,}69461$$

Región de rechazo:

$$P(\chi^2(2) \geq T_{cal}) < \alpha$$

Con $\alpha = 0{,}05$

$$P(\chi^2(2) \geq 0{,}69461) < 0{,}05$$

Con ayuda de r obtenemos:

Dado que el valor p de 0.7065898 supera ampliamente el umbral de significancia de 0.05, no se encuentra evidencia estadística para rechazar la hipótesis Por lo tanto, con un alto grado de confianza, podemos concluir que las diferencias observadas en los niveles de asertividad entre los primogénitos, los nacidos en segundo lugar y los nacidos en tercer lugar no son estadísticamente significativas. Esto sugiere que el orden de nacimiento no tiene un impacto discernible en la asertividad de los individuos dentro de la muestra estudiada.

En R se puede hacer con el siguiente codigo.

```
x1<-c(18,8,4,21,28,32,10,18,12,3,24,22,1,7,19,2,30,18,5,14)
g1<-c(rep(1,7),rep(2,6),rep(3,7))
kruskal.test(x1,g1)

        Kruskal-Wallis rank sum test

data:  x1 and g1
Kruskal-Wallis chi-squared = 0.69461, df = 2, p-value = 0.7066
```

Pregunta 4

Una pareja de esposos salieron a jugar bolos y guardaron sus resultados para ver si existia una relación entre dichos resultados:

Esposo	147	158	131	142	183	151	196	129	155
Esposa	122	128	125	123	115	120	108	143	124

Cuadro 3: Puntajes de bolos de la pareja

Use ρ de Pearson, el τ de kendall y el ρ de spearman para realizar una prueba de independencia entre los puntajes.

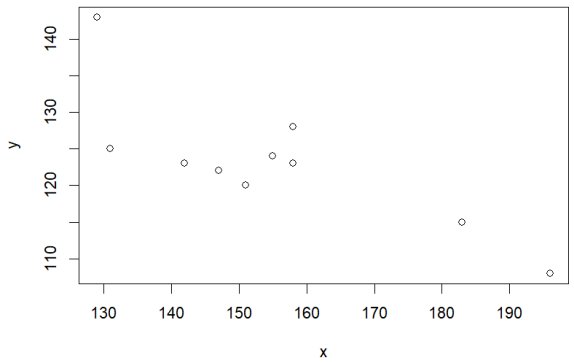
Coeficiente de correlación de ρ de peason.

Esta medida de correlacón es la mas utilizada que se puede calcular de la siguiente forma:

$$r = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2]^{1/2}[\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2]^{1/2}} \quad (1)$$

r es una medida de la fuerza de asociacion lineal entre X e Y, si en un grafico de dispersio (X,Y) todos muy próximos a una linea recta, entonces r sera muy cercano a 1 si la recta tiene pendiente ascendente, y sera muy cercano a -1 si la recta tiene pendiente descendente.

Además, teniendo en cuenta que el coeficiente de Pearson es útil para evaluar relaciones lineales, resulta beneficioso representar los datos mediante un gráfico de dispersión para observar mejor la tendencia.



La observación inicial fromula una aparente tendencia negativa en la correlación de Pearson basandonos en la imagen proporcionada. Sin embargo, no se distingue claramente una relación lineal. A pesar de ello, se procederá a calcular la correlación.

Para calcular necesitamos la siguiente información:

- X_i: Puntaje del esposo.
- Y_i: Puntaje de la esposa.
- n: numero total de pares.
- $\bar{X}$: la media del puntaje del esposo.
- $\bar{Y}$: la media del puntaje de la esposa.
- $\bar{X}$ $\bar{Y}$: El producto de las medias.

X_i **y** Y_i

De los cuales ya tenemos los valores en el cuadro anterior.

n

como son muestras en pares tenemos un total de 10

$\bar{X}$

$$\frac{147+158+131+142+183+151+196+129+155+158}{10} = \frac{1550}{10} = 155$$

$\bar{Y}$

$$\frac{122+128+125+123+115+120+108+143+124+123}{10} = \frac{1231}{10} = 123{,}1$$

$\bar{X}$ * $\bar{Y}$

$$155 * 123{,}1 = 19080{,}5$$

Dado que la ecuación (1) es extensa, procederemos a desglosarla en partes por los colores rojo, verde y azul. Comenzaremos con la sección resaltada en rojo.

$$\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y} = [147*122-10*19080{,}5]+[158*128-10*19080{,}5]+[131*125-10*19080{,}5]+[142*123-10*19080{,}5]....$$

$$+[183*115-10*19080{,}5]+[151*120-10*19080{,}5]+[196*108-10*19080{,}5]+[129*143-10*19080{,}5].....$$
$$+[155*124-10*19080{,}5]+[158*123-10*19080{,}5] = -1372$$

$$[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2] = [147^2-10*155^2]+[158^2-10*155^2]+[131^2-10*155^2]+[142^2-10*155^2]+[183^2-10*155^2]....$$

$$+[151^2-10*155^2]+[196^2-10*155^2]+[129^2-10*155^2]+[155^2-10*155^2]+[158^2-10*155^2] = 63{,}11894$$

$$[\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2] = [122^2-10*123{,}1^2]+[128^2-10*123{,}1^2]+[125^2-10*123{,}1^2]+[123^2-10*123{,}1^2]+[115^2-10*123{,}1^2]+[120^2-10*123{,}1^2]....$$

$$+[108^2-10*155^2]+[143^2-10*155^2]+[124^2-10*155^2]+[123^2-10*155^2] = 26{,}99815$$

La ecuación (1) queda de la forma:

$$\frac{-1372}{[63{,}11894]*[26{,}99815]} = -0{,}8051196$$

El coeficiente de correlación de Pearson nos arroja un valor de -0.8051196, el cual también podemos calcular en R utilizando el siguiente comando:

```
pearson <- cor(x, y, method = "pearson")
pearson
[1] -0.8051197
```

En el juego bolos, note que el coeficiente de correlación de Pearson es -0.805, lo cual indica una fuerte relación inversa entre los puntajes del esposo y los de la esposa. Esto significa que, en los días de juego, generalmente cuando el esposo logro un puntaje alto, la esposa tiende a tener un puntaje más bajo, y viceversa.

Tambie sabiendo que el coeficiente de pearson es para relaciones de pearson es bueno ver un grafico de dispersión como es la tendencia

Correlación de spearman

Los datos consisten en una muestra aleatoria bivariada de tamaño n, donde cada observación se representa como (X_i, Y_i). Se define $R(X_i)$ como el rango de X_i en comparación con los otros valores de X, para $i = 1, 2, ..., n$; y $R(Y_i)$ como el rango de Y_i en comparación con los otros valores de Y, para $i = 1, 2, ..., n$. Además, si las observaciones son no numéricas pero pueden ser clasificadas según su calidad o preferencia, se pueden asignar rangos basados en estas categorías. En caso de empate, se asigna el promedio de los rangos a cada valor empatado, similar a como se hace en el test de Mann-Whitney.

$$\frac{\sum_{i=1}^{n} R(x_i)R(y_i) - n\left(\frac{n+1}{2}\right)^2}{\sqrt{\left(\sum_{i=1}^{n} R^2(x_i) - n\left(\frac{n+1}{2}\right)^2\right)\left(\sum_{i=1}^{n} R^2(y_i) - n\left(\frac{n+1}{2}\right)^2\right)}} \quad (2)$$

Como se observa en la tabla de datos, los registros del esposo presentan empates, ya que el valor "158.ªparece dos veces.

Elaboramos una tabla para organizar los rankins: Para abordar la repetición del valor "158.en las posiciones 7 y 8,

x	y	R(x)	R(y)	R(x)R(y)
147	122	4.0	4.0	16.00
158	128	7.5	9.0	67.50
131	125	2.0	8.0	16.00
142	123	3.0	5.5	16.50
183	115	9.0	2.0	18.00
151	120	5.0	3.0	15.00
196	108	10.0	1.0	10.00
129	143	1.0	10.0	10.00
155	124	6.0	7.0	42.00
158	123	7.5	5.5	41.25

Cuadro 4: Datos de ejemplo

podemos reemplazar ambas instancias por el promedio de las posiciones, es decir, $\frac{7+8}{2} = 7{,}5$. En donde se sustituye "158"por '7.5´ en las posiciónes correspondiente.

$$\sum_{i=0}^{n} R^2(X_i) = 4{,}0^2 + 7{,}5^2 + 2{,}0^2 + 3{,}0^2 + 9{,}0^2 + 5{,}0^2 + 10{,}0^2 + 1{,}0^2 + 6{,}0^2 + 7{,}5^2 = 384{,}5$$

para:

$$\sum_{i=0}^{n} R^2(y_i) = 4{,}0^2 + 9{,}0^2 + 8{,}0^2 + 5{,}5^2 + 2{,}0^2 + 3{,}0^2 + 1{,}0^2 + 10{,}0^2 + 7{,}0^2 + 5{,}5^2 = 384{,}5$$

Para:

$$\sum_{0}^{n} R(X_i)R(Y_i = 16{,}00 + 67{,}50 + 16{,}00 + 16{,}50 + 18{,}00 + 15{,}00 + 10{,}00 + 10{,}00 + 42{,}00 + 41{,}25 = 252{,}25$$

Reemplazando en (2) tenemos

$$\frac{252{,}25 - 10(\frac{10+1}{2})^2}{\sqrt{(384{,}5 - 10(\frac{10+1}{2})) * (384{,}5 - 10(\frac{10+1}{2}))}} = -0{,}6128049$$

El coeficiente de correlación de spearman nos arroja un valor de -0.6128049, el cual también podemos calcular en R utilizando el siguiente comando:

```
F_c <- ((10 + 1) / 2)
numerador <- 252.25 -10*(F_c^2)
denomidador_term <- 384.5 - 10 * F_c^2
denomidador <- sqrt(denomidador_term * denomidador_term)
result <- numerador / denomidador
result
[1] -0.6128049
```

El coeficiente de Spearman () de -0.613 revela una correlación negativa que oscila entre moderada y fuerte. Este tipo de correlación es análoga a la de Pearson, aunque se fundamenta en rangos en vez de valores precisos.

Coeficiente τ de Kendall

Empates: Si $X_1 = X_2$, no hay comparación. Si $Y_1 = Y_2$ y $X_1 \neq X_2$, el par se considera mitad concordante y mitad discordante.

$$\tau = \frac{N_c - N_d}{N_c + N_d},$$

donde N_c es el número de pares concordantes y N_d es el número de pares discordantes. En resumen,

Si $\frac{Y_j - Y_i}{X_j - X_i} > 0$, hay concordancia.

Si $\frac{Y_j - Y_i}{X_j - X_i} < 0$, hay discordancia.

Si $\frac{Y_j - Y_i}{X_j - X_i} = 0$, hay $\frac{1}{2}$ concordancia y $\frac{1}{2}$ discordancia.

Si $X_i = X_j$ no se hace comparación.

El cálculo de N_c y N_d se hace más sencillo si las observaciones (X_i, Y_i) se arreglan de acuerdo con valores crecientes de X, y luego con respecto a valores crecientes de Y, así cada Y se compara con los de abajo.

Ahora definimos lo siguiente:

- $\frac{Y_j - Y_i}{X_j - X_i} > 0$=+
- $\frac{Y_j - Y_i}{X_j - X_i} < 0$=-
- $\frac{Y_j - Y_i}{X_j - X_i} = 0$=0
    - se suma 0.5 cuando es concordancia o discordancia

Esto ya se hace para faciltar los calculos:

Primer Par (129, 143)

$\frac{129-131}{143-125} = \frac{-}{+} = -$ $\quad \frac{129-151}{143-120} = \frac{-}{+} = -$ $\quad \frac{129-158}{143-128} = \frac{-}{+} = -$

$\frac{129-142}{143-123} = \frac{-}{+} = -$ $\quad \frac{129-155}{143-124} = \frac{-}{+} = -$ $\quad \frac{129-183}{143-115} = \frac{-}{+} = -$

$\frac{129-147}{143-122} = \frac{-}{+} = -$ $\quad \frac{129-158}{143-123} = \frac{-}{+} = -$ $\quad \frac{129-196}{143-108} = \frac{-}{+} = -$

concordancias
0
discordancias
9

Segundo par (131, 125)

$\frac{131-142}{125-123} = \frac{-}{+} = -$ $\quad \frac{131-155}{125-124} = \frac{-}{+} = -$ $\quad \frac{131-183}{125-115} = \frac{-}{+} = -$

$\frac{131-147}{125-122} = \frac{-}{+} = -$ $\quad \frac{131-158}{125-123} = \frac{-}{+} = -$ $\quad \frac{131-196}{125-108} = \frac{-}{+} = -$

$\frac{131-151}{125-120} = \frac{-}{+} = -$ $\quad \frac{131-158}{125-128} = \frac{-}{-} = +$

concordancias
1
discordancias
7

tercer par (142, 123)

$\frac{142-147}{123-122} = \frac{-}{+} = -$ $\quad \frac{142-158}{123-123} = \frac{-}{0} = 0$

$\frac{142-151}{123-120} = \frac{-}{+} = -$ $\quad \frac{142-158}{123-128} = \frac{-}{-} = +$ $\quad \frac{142-196}{123-108} = \frac{-}{+} = -$

$\frac{142-165}{123-124} = \frac{-}{-} = +$ $\quad \frac{142-183}{123-115} = \frac{-}{+} = -$

concordancias
2 + 0.5 = 2.5
discordancias
4 + 0.5 = 4.5

cuarto par (147, 122)

$\frac{147-151}{122-120} = \frac{-}{+} = -$ $\quad \frac{147-158}{122-128} = \frac{-}{-} = +$

$\frac{147-155}{122-124} = \frac{-}{-} = +$ $\quad \frac{147-183}{122-115} = \frac{-}{+} = -$

$\frac{147-158}{122-123} = \frac{-}{-} = +$ $\quad \frac{147-196}{122-108} = \frac{-}{+} = -$

concordancias
3
discordancias
3

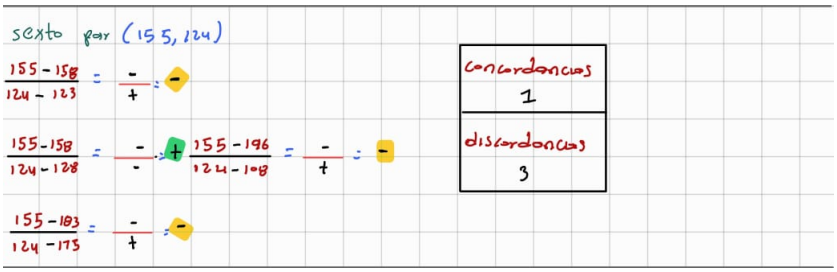
quinto par (151, 120)

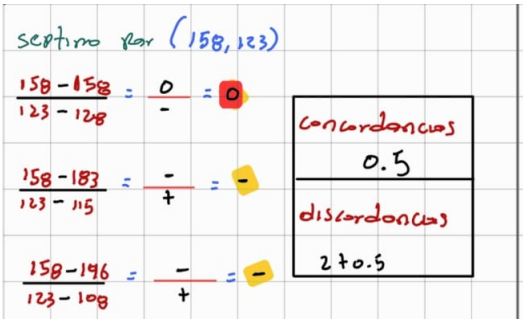
$\frac{151-155}{120-124} = \frac{-}{-} = +$ $\quad \frac{151-183}{120-115} = \frac{-}{+} = -$

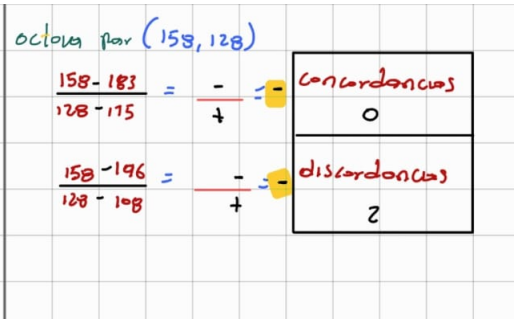
$\frac{151-158}{120-123} = \frac{-}{-} = +$ $\quad \frac{151-196}{120-108} = \frac{-}{+} = -$

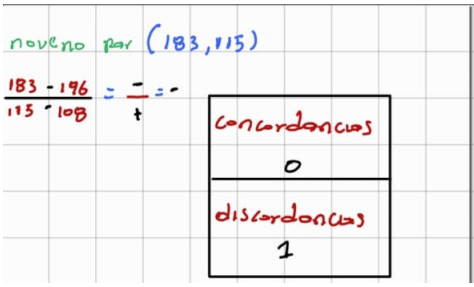
$\frac{151-158}{120-128} = \frac{-}{-} = +$

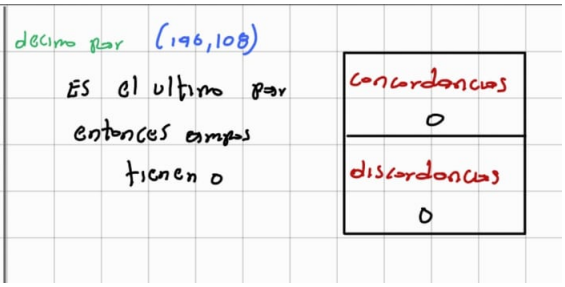
concordancias
3
discordancias
2


sexto par (155, 124)
155 - 158 / 124 - 123 = - / + = -
155 - 158 / 124 - 128 = - / - + 155 - 196 / 124 - 108 = - / + = -
155 - 183 / 124 - 175 = - / + = -
concordancias
1
discordancias
3


septimo par (158, 123)
158 - 158 / 123 - 128 = 0 / - = 0
158 - 183 / 123 - 115 = - / + = -
158 - 196 / 123 - 108 = - / + = -
concordancias
0.5
discordancias
2 + 0.5


octava par (158, 128)
158 - 183 / 128 - 175 = - / + = -
158 - 196 / 128 - 108 = - / + = -
concordancias
0
discordancias
2


noveno par (183, 115)
183 - 196 / 115 - 108 = - / + = -
concordancias
0
discordancias
1


decimo par (196, 108)
Es el ultimo par
entonces ambos
tienen 0
concordancias
0
discordancias
0

En resumen tenemos:

(X_i, Y_i)	**Pares Concordantes**	Pares Discordantes
(129,143)	0	9
(131,125)	1	7
(142,123)	2.5	4.5
(147,122)	3	3
(151,120)	3	2
(155,124)	1	3
(158,123)	0.5	2.5
(158,128)	0	2
(183,115)	0	1
(196,108)	0	0
	$N_c 11$	$N_d = 34$

$$\tau = \frac{N_c - N_d}{N_c + N_d} = \frac{11 - 34}{11 + 34} = -0{,}5111111$$

El coeficiente de correlación de Kendall nos arroja un valor de -0.5111111, el cual también podemos calcular en R utilizando el siguiente comando:

```
kendall <- cor(esposo, esposa, method = "kendall")
kendall
[1] -0.5227273
```

El coeficiente de Tau de Kendall, calculado manualmente, presenta una ligera diferencia respecto al obtenido mediante R, pero llegamos a la misma interpretación cualitativa: hay una correlación negativa. Esta correlación se sitúa en un rango moderado, menos intensa que la reflejada por Spearman, y se distingue aún más de la correlación de Pearson. Este resultado refuerza que la elección del método de correlación es clave para una interpretación adecuada de la relación entre variables.

Conclusiones

Las distintas métricas de correlación apuntan consistentemente a una relación negativa entre las variables. El análisis gráfico de dispersión sugiere que la relación no es lineal, lo que justifica que la correlación de Pearson, que presupone linealidad, no sea el indicador más apropiado en este contexto. En contraste, las correlaciones de Spearman y Kendall, que no asumen tal linealidad y son más robustas ante relaciones monotónicas, proporcionan estimaciones más coherentes entre sí y, por lo tanto, se perfilan como alternativas más fiables para evaluar la asociación entre los resultados de bolos.

Prueba de independencia entre estos puntajes.

Se realizan pruebas de hipótesis para determinar la independencia entre los puntajes de bolos de una pareja de esposos, utilizando las correlaciones de Pearson, Spearman y Kendall.

Prueba de Independencia de Pearson:

$$H_0 : \rho = 0 \quad \text{(Los puntajes son independientes)},$$

$$H_a : \rho \neq 0 \quad \text{(Los puntajes no son independientes)},$$

$$p\text{-valor} = 0{,}004951 < 0{,}05 \quad \Rightarrow \quad \text{Rechazamos } H_0,$$

Prueba de Independencia de Spearman:

$$H_0 : \rho_s = 0 \quad \text{(Los puntajes son independientes)},$$

$$H_a : \rho_s \neq 0 \quad \text{(Los puntajes no son independientes)},$$

$$p\text{-valor} = 0{,}05961 > 0{,}05 \quad \Rightarrow \quad \text{No rechazamos } H_0,$$

Prueba de Independencia de Kendall:

$$H_0 : \tau = 0 \quad \text{(Los puntajes son independientes)},$$

$$H_a : \tau \neq 0 \quad \text{(Los puntajes no son independientes)},$$

$$p\text{-valor} = 0{,}03811 < 0{,}05 \quad \Rightarrow \quad \text{Rechazamos } H_0,$$

Después de realizar las pruebas de independencia utilizando las correlaciones de Pearson, Spearman y Kendall, podemos decir las siguientes conclusiones:

- La **Prueba de Independencia de Pearson** muestra un p-valor de 0.004951, lo cual es menor que el nivel de significancia de 0.05. Esto nos lleva a rechazar la hipótesis nula de independencia, indicando que existe una correlación lineal negativa significativa entre los puntajes de la pareja.
- En la **Prueba de Independencia de Spearman**, obtenemos un p-valor de 0.05961. Dado que es mayor que 0.05, no podemos rechazar la hipótesis nula de independencia al nivel de confianza del 95 %. Esto implica que no hay evidencia suficiente para afirmar que existe una correlación monotónica significativa entre los puntajes.
- Finalmente, la **Prueba de Independencia de Kendall** arroja un p-valor de 0.03811, que es menor que 0.05, lo que sugiere rechazar la hipótesis nula de independencia. Por lo tanto, se concluye que hay una correlación de rango negativa significativa entre los puntajes.

Con base en los resultados de Pearson y Kendall, hay evidencias de que los puntajes de la pareja no son independientes, lo que sugiere una relación negativa. Sin embargo, la prueba de Spearman no proporcionó suficiente evidencia para confirmar esta dependencia al nivel de confianza del 95 %. Esto puede deberse a la naturaleza de los datos o a la sensibilidad de cada prueba a diferentes tipos de relaciones.

```r
peason_test <- cor.test(x,y,method=c("pearson"))
spearman_test <- cor.test(x, y, method = "spearman")
kendall_test <- cor.test(x, y, method = "kendall")
```

```
        Pearson's product-moment correlation

data:  x and y
t = -3.8394, df = 8, p-value = 0.004951
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9521021 -0.3559160
sample estimates:
       cor
-0.8051197

Warning: Cannot compute exact p-value with ties
        Spearman's rank correlation rho

data:  x and y
S = 266.11, p-value = 0.05961
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.6128049

Warning: Cannot compute exact p-value with ties
        Kendall's rank correlation tau

data:  x and y
z = -2.0737, p-value = 0.03811
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau
-0.5227273
```

Punto 5

Literal A

Como la pendiente es un cambio en Y divido por un cambio en x, esto es equivalente a decir, que la pendiente de la regresion de millas sobre galones es de al menos $\frac{18}{1} = 18$, por lo tanto nuestra juego de hipotesis seria el siguiente:

$$\begin{cases} H_0 : \beta = 18 \\ H_1 : \beta \neq 18 \end{cases}$$

Usando la adaptacion de el ρ de Spearman, se hace lo siguiente:

Se toman los valores de nuestro variable X, se le encuentra los rangos a la misma, se calcula el valor de cada u_i, el cual es igual a: $Y_i - \beta_0 X_i$, y hallamos el rango de estos u_i, al tener todos estos datos, procederemos a calcular el ρ de spearman.

```
millas = c(142, 116, 194, 250, 88, 157, 255, 154, 43, 208)

galones = c(11.1, 5.7, 14.2, 15.8, 7.5, 12.5, 17.4, 8.8, 3.4, 15.2)
```

RANGOS

```
ui = millas - 18*galones

matriz = data.frame(
  "xi" = galones,
  "ui" = ui,
  "R(xi)" = sapply(galones, function(x) which(sort(galones) == x)),
  "R(ui)" = sapply(ui, function(x) which(sort(ui) == x))
)

knitr::kable(matriz, format = "markdown", caption = "Tabla con rangos literal A")
```

Table 1: Tabla con rangos literal A

xi	ui	R.xi.	R.ui.
11.1	-57.8	5	5
5.7	13.4	2	10
14.2	-61.6	7	3
15.8	-34.4	9	7
7.5	-47.0	3	6
12.5	-68.0	6	1
17.4	-58.2	10	4
8.8	-4.4	4	9
3.4	-18.2	1	8
15.2	-65.6	8	2

Ya que tenemos la tabla completa, procederemos a calcular el ρ de spearman, el cual es el siguiente:

$$\rho = \frac{\sum_{i=1}^{n} R(x_i)R(u_i) - n(\frac{n+1}{2})^2}{[\sum_{i=1}^{n} R^2(x_i) - n(\frac{n+1}{2})^2]^{\frac{1}{2}}[\sum_{i=1}^{n} R^2(u_i) - n(\frac{n+1}{2})^2]^{\frac{1}{2}}}$$

```r
resta = length(galones) * ((length(galones) + 1)/2)^2

numerador = sum(matriz$R.xi.*matriz$R.ui.) - resta

denominador = sqrt(sum(matriz$R.xi.^2) - resta)*sqrt(sum(matriz$R.ui.^2) - resta)

p = numerador/denominador
```

Reemplazando y haciendo los calculos respectivos, el resultado es:

$$\rho = -0.6$$

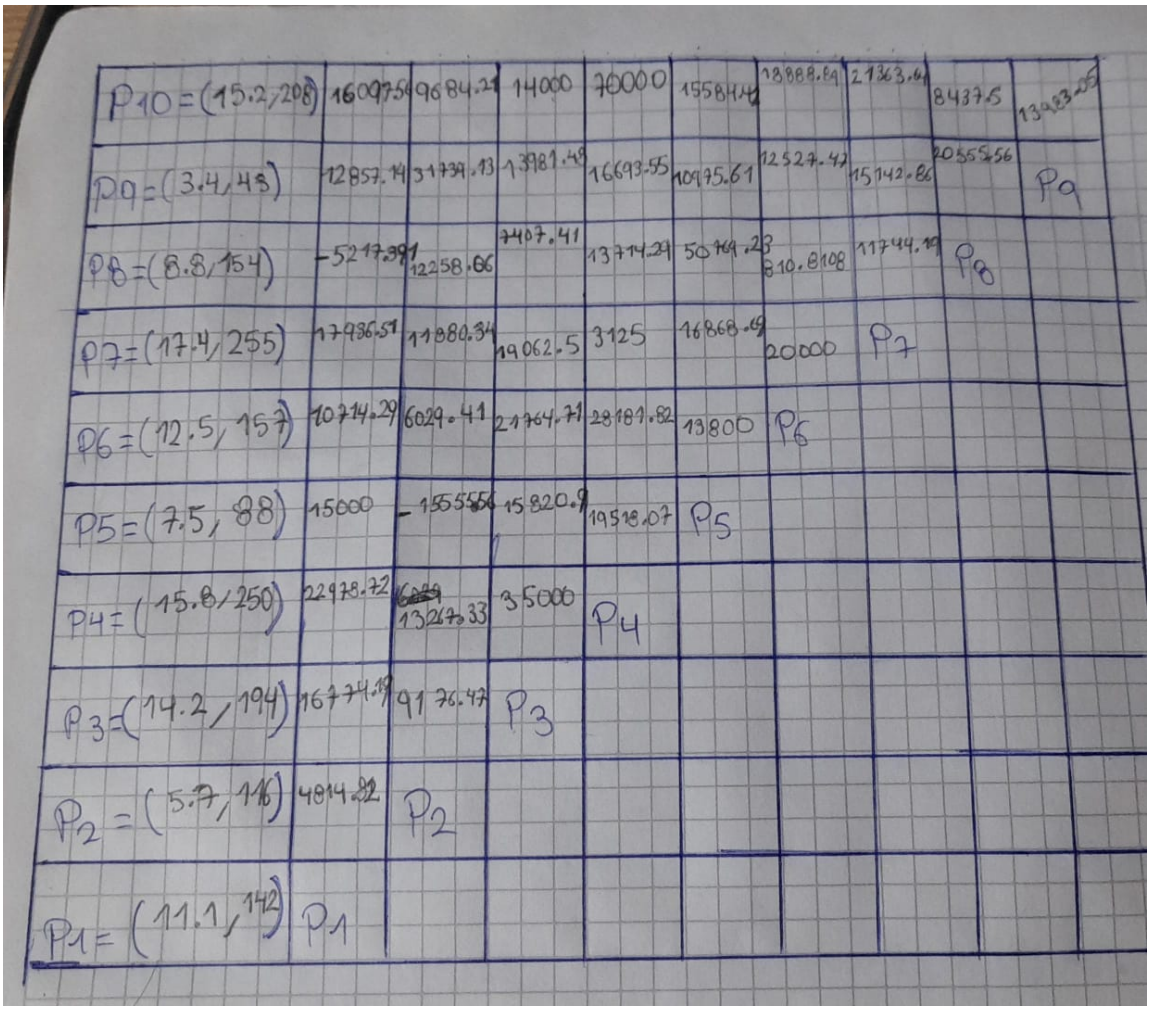
Ahora, nuesto valor $P = 2P(Z >= |-0.6| * \sqrt{9})$

```r
p_value = 2*pnorm(0.6*sqrt(9), lower.tail = F)
```

$$2P(Z >= 1.8) \approx 0.07186064$$

Comparando con un $\alpha = 0.05$, ya que es mayor a nuestro *alpha*, se acepta la hipotesis nula y por lo tanto, esta cifra si se aplica este carro en particular.

Literal B




```
pendientes = c(4814.815, 16774.19, 22978.72, 15000, 10714.29, 17936.51,
               -5217.391, 12857.14, 16097.56, 9176.471, 13267.33, -15555.56,
               6029.412, 11880.34, 12258.06, 31739.13, 9684.211, 35000,
               15820.9, 21764.71, 19062.5, 7407.407, 13981.48, 14000,
               19518.07, 28181.82, 3125, 13714.29, 16693.55, 70000, 13800,
               16868.69, 50769.23, 10975.61, 15584.42, 20000, 810.8108,
               12527.47, 18888.89, 11744.19, 15142.86, 21363.64, 20555.56,
               8437.5, 13983.05)
```

Para $n = 10$, se tiene que el cuantil 0.975 para T es $W_{0.975} = 21$

Como no tenemos valores iguales a 0 en las pendientes, ni valores de $X_i = X_j$, entonces tenemos 2 casos:

- Si $\frac{Y_j - Y_i}{X_j - X_i} > 0$ sumarle 1 a N_c (concordancia)
- Si $\frac{Y_j - Y_i}{X_j - X_i} < 0$ sumarle 1 a N_d (discondancia)

Entonces en nuestro caso:

TABLE A11 Quantiles of the Kendall test statistic $T = N_c - N_d$. Quantiles of Kendall's τ are given in parentheses. Lower quantiles are the negative of the upper quantiles, $w_p = -w_{1-p}$.

n	$p = 0.900$	0.950	0.975	0.990	0.995
4	4 (0.6667)	4 (0.6667)	6 (1.0000)	6 (1.0000)	6 (1.0000)
5	6 (0.6000)	6 (0.6000)	8 (0.8000)	8 (0.8000)	10 (1.0000)
6	7 (0.4667)	9 (0.6000)	11 (0.7333)	11 (0.7333)	13 (0.8667)
7	9 (0.4286)	11 (0.5238)	13 (0.6190)	15 (0.7143)	17 (0.8095)
8	10 (0.3571)	14 (0.5000)	16 (0.5714)	18 (0.6429)	20 (0.7143)
9	12 (0.3333)	16 (0.4444)	18 (0.5000)	22 (0.6111)	24 (0.6667)
10	15 (0.3333)	19 (0.4222)	21 (0.4667)	25 (0.5556)	27 (0.6000)

Figure 1: Tabla A11

$$N_c = 43, N_d = 2$$

$$N = N_c + N_d = 45$$

$$r = \frac{1}{2}(N - W_{1-\frac{\alpha}{2}}) = \frac{1}{2}(45 - 21) = 12$$

$$s = \frac{1}{2}(N + W_{1-\frac{\alpha}{2}}) + 1 = N + 1 - r = 45 + 1 - 12 = 34$$

```
pendientes = sort(pendientes/10^3)

pendientes
```

```
##  [1] -15.5555600  -5.2173910   0.8108108   3.1250000   4.8148150   6.0294120
##  [7]   7.4074070   8.4375000   9.1764710   9.6842110  10.7142900  10.9756100
## [13]  11.7441900  11.8803400  12.2580600  12.5274700  12.8571400  13.2673300
## [19]  13.7142900  13.8000000  13.9814800  13.9830500  14.0000000  15.0000000
## [25]  15.1428600  15.5844200  15.8209000  16.0975600  16.6935500  16.7741900
## [31]  16.8686900  17.9365100  18.8888900  19.0625000  19.5180700  20.0000000
## [37]  20.5555600  21.3636400  21.7647100  22.9787200  28.1818200  31.7391300
## [43]  35.0000000  50.7692300  70.0000000
```

```
paste(pendientes[12],", ", pendientes[34])
```

```
## [1] "10.97561 ,  19.0625"
```

Un I.C del **95%** para β es:

$$[S^{(12)}, S^{(34)}] = [10.97561, 19.0625]$$

Por lo tanto con un 95% de confianza, β se encontrara entre 10.97561 y 19.0625.

Literal C

El metodo de minimos cuadrados tiene la siguiente forma:

$$Y = \alpha + \beta x$$

Los estimadores por minimos cuadrados para α y β son:

$$\beta = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}$$

$$\alpha = \bar{Y} - \beta\bar{X}$$

Reemplazando y haciendo los calculos respectivos, el resultado es el siguiente:

```
b = (sum(galones*millas) - 10*mean(galones)*mean(millas))/(sum(galones^2) - 10*mean(galones)^2)

a = mean(millas) - b*mean(galones)
```

```
## [1] "alpha = 5.87501524204342"
```

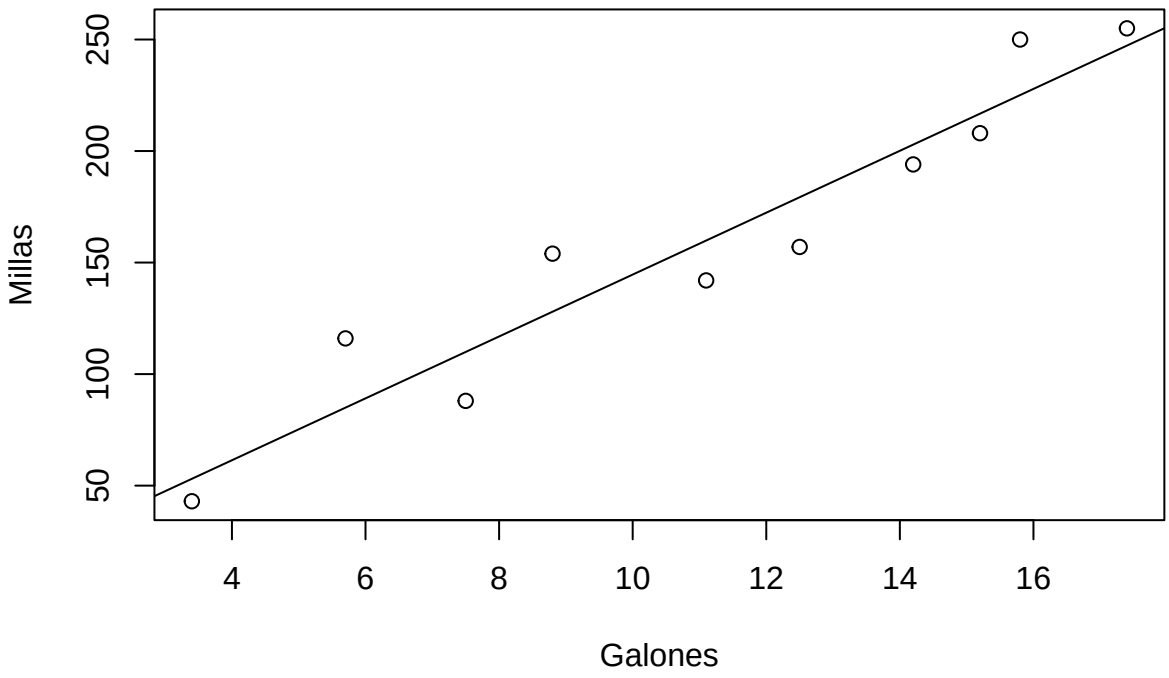
```
## [1] "beta = 13.8732065195302"
```

Y si lo comparamos con la funcion lm de R, la cual hace una estimacion de una curva de regresion por el metodo de minimos cuadrados, podemos ver que nos da exactamente lo mismo:

```
regresion <- lm(millas ~ galones)
summary(regresion)
```

```
##
## Call:
## lm(formula = millas ~ galones)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.290 -15.912  -8.811  20.629  31.048
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.875     18.843   0.312    0.763
## galones       13.873      1.569   8.842 2.11e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.01 on 8 degrees of freedom
## Multiple R-squared:  0.9072, Adjusted R-squared:  0.8956
## F-statistic: 78.18 on 1 and 8 DF,  p-value: 2.11e-05
```

```
plot(galones, millas, xlab = "Galones", ylab = "Millas")

abline(regresion)
```



Literal D

Pasos

Paso 1 Primero debemos hallar los rangos para cada observacion de X y de Y.

```
matriz = matrix(NA, nrow = 10, ncol = 4)

matriz[1:10, 1] = sort(galones)

matriz[1:10, 3] = sort(millas)

matriz[,2] = c(1:10)

matriz[,4] = c(1:10)

matriz = as.data.frame(matriz)

colnames(matriz) = c("X", "Rango X", "Y", "Rango Y")

knitr::kable(matriz, format = "markdown", caption = "Tabla creacion rangos")
```

Table 2: Tabla creacion rangos

X	Rango X	Y	Rango Y
3.4	1	43	1
5.7	2	88	2
7.5	3	116	3
8.8	4	142	4
11.1	5	154	5
12.5	6	157	6
14.2	7	194	7
15.2	8	208	8
15.8	9	250	9
17.4	10	255	10

```
rangos = data.frame(
  "X" = galones,
  "R(x)" = sapply(galones, function(x) which(matriz$X == x)),
  "Y" = millas,
  "R(y)" = sapply(millas, function(x) which(matriz$Y == x))
)

knitr::kable(rangos, format = "markdown", caption = "Tabla con rangos")
```

Table 3: Tabla con rangos

X	R.x.	Y	R.y.
11.1	5	142	4
5.7	2	116	3
14.2	7	194	7
15.8	9	250	9
7.5	3	88	2
12.5	6	157	6
17.4	10	255	10
8.8	4	154	5
3.4	1	43	1
15.2	8	208	8

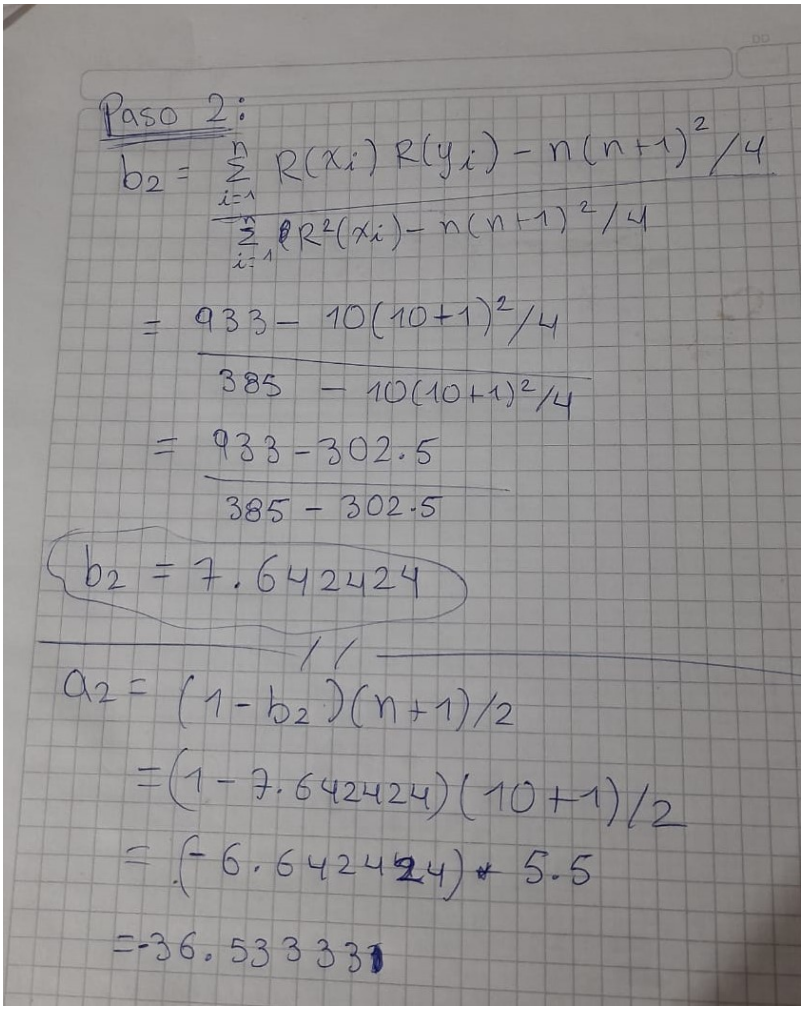
Paso 2 Ahora hallaremos la regresion por minimos cuadrados de los rangos, la cual es igual a:

$$y = a_2 + b_2 x$$

donde:

$$b_2 = \frac{\sum_{i=1}^{n} R(x_i)R(y_i) - \frac{n(n+1)^2}{4}}{\sum_{i=1}^{n} R^2(x_i) - \frac{n(n+1)^2}{4}}$$

$$a_2 = (1 - b_2)\frac{n+1}{2}$$




```
b2 = (sum(rangos$R.x.*rangos$R.y.) - 10*11^2/4)/ (sum(rangos$R.x.^2) - ((10*11^2)/4))

a2 = (1 - b2)*11/2
```

Usando los datos de la tabla, obtenemos el siguiente resultado:

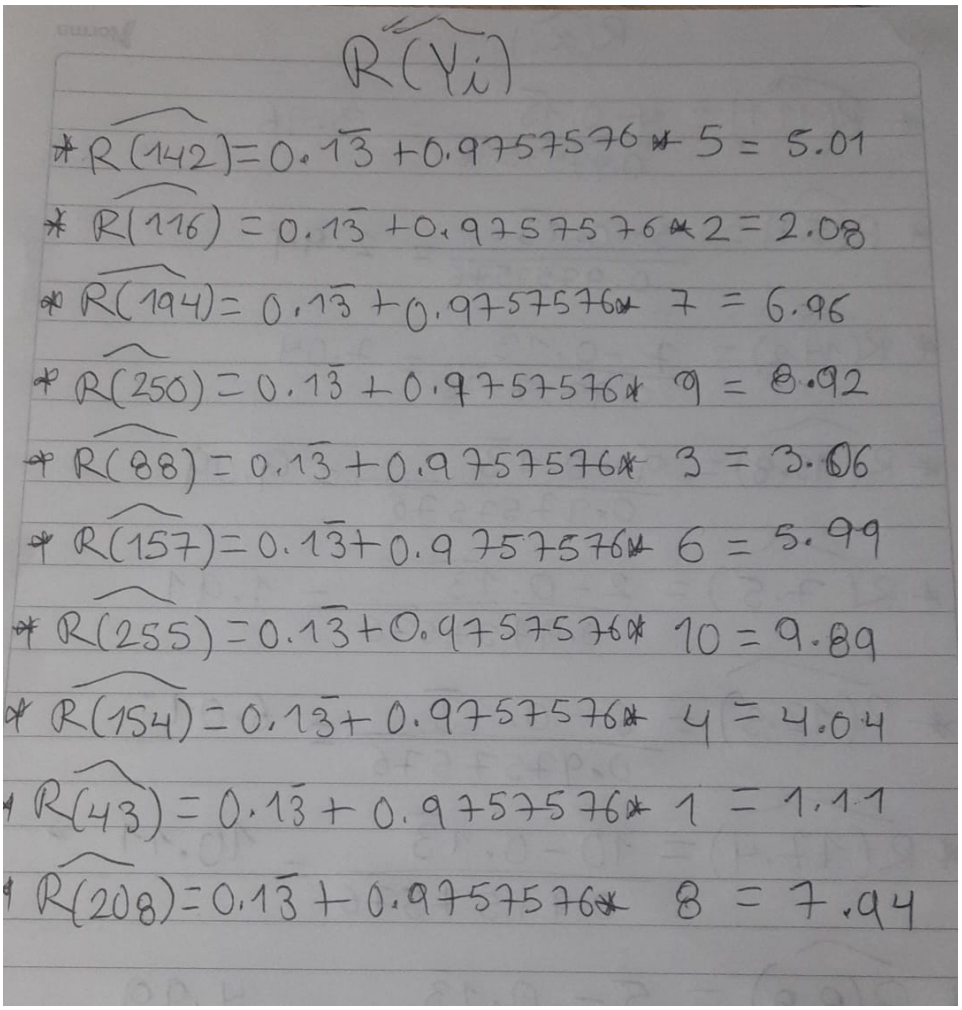
$$y = 0.1333333 + 0.9757576 * x$$

Paso 3 R(yi)

Usando el resultado de la regresion por minimos cuadrados de los rangos, hallaremos los $\widehat{R(y_i)}$.

$$\widehat{R(y_i)} = 0.1333333 + 0.9757576 * R(x_i)$$

Haciendo los calculos a mano obtendremos lo siguiente:

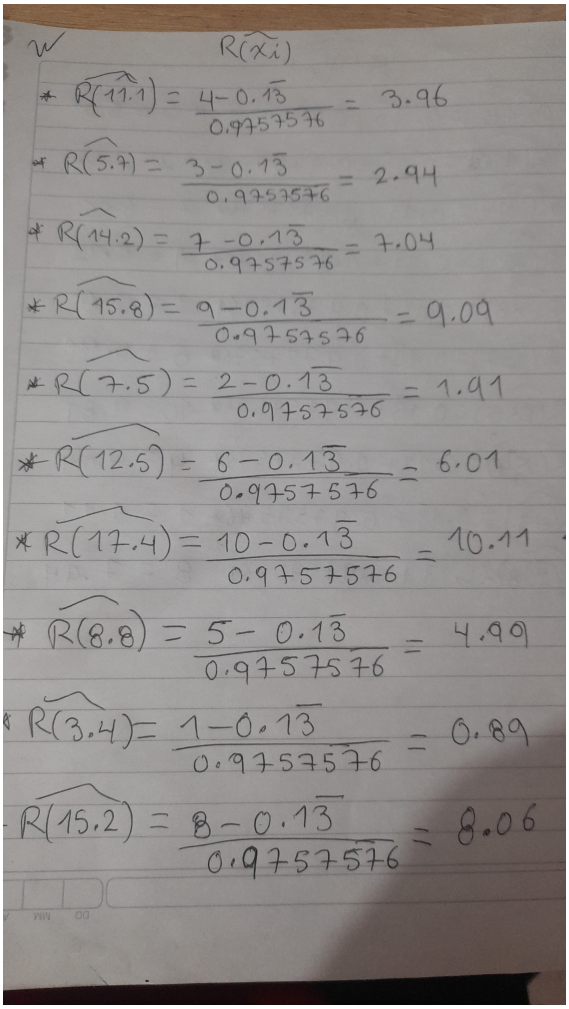


R(xi)

Para hallar cada $\widehat{R(x_i)}$, lo calcularemos de la siguiente forma:

$$\widehat{R(x_i)} = \frac{R(y_i) - 0.1333333}{0.9757576}$$

Haciendo los calculos a mano obtendremos lo siguiente:



Paso 4 De momento nuestra tabla va asi:

Table 4: Tabla con los ajustados

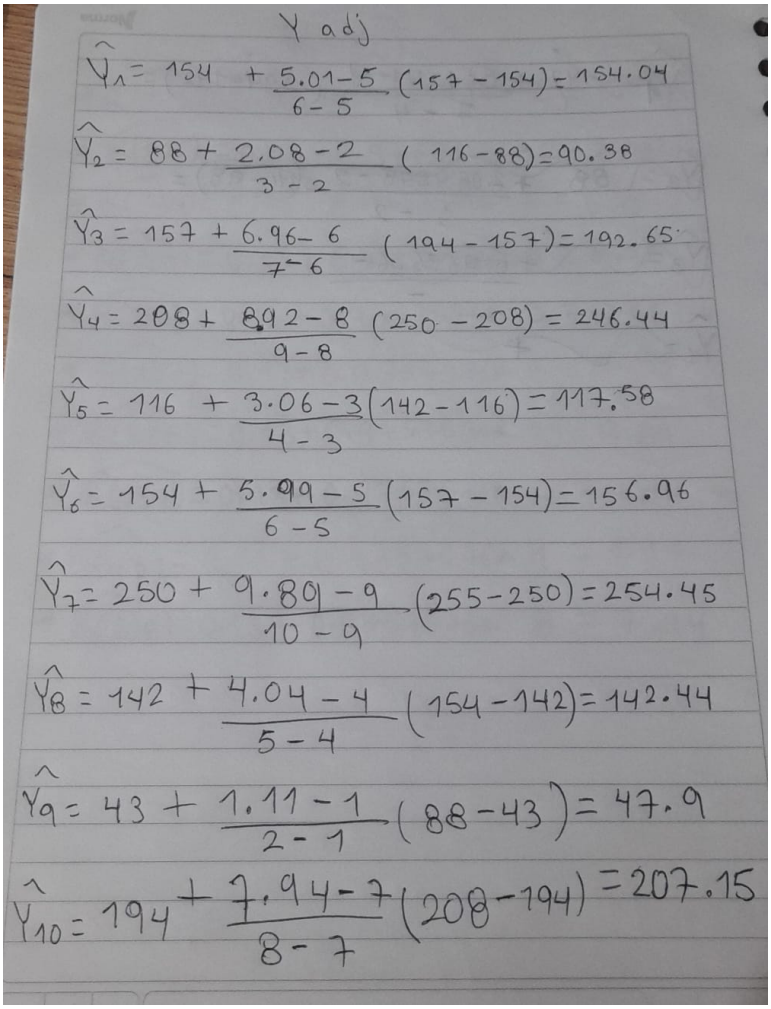
X	R.x.	Y	R.y.	R.y.adj	R.x.adj
11.1	5	142	4	5.012121	3.9627329
5.7	2	116	3	2.084849	2.9378882
14.2	7	194	7	6.963636	7.0372671
15.8	9	250	9	8.915152	9.0869565
7.5	3	88	2	3.060606	1.9130435
12.5	6	157	6	5.987879	6.0124224
17.4	10	255	10	9.890909	10.1118012
8.8	4	154	5	4.036364	4.9875776
3.4	1	43	1	1.109091	0.8881988
15.2	8	208	8	7.939394	8.0621118

$Y_i s$ ajustados

Teniendo en cuenta los valores obtenidos, tenemos lo siguiente:

- Todas las filas corresponden al caso donde $\widehat{R(y_0)}$ esta entre 2 rangos de valores adyancentes, para este caso:

$$\widehat{Y_i} = y_i + \frac{R(y_0) - R(y_i)}{R(y_j) - R(y_i)}(y_j - y_i)$$



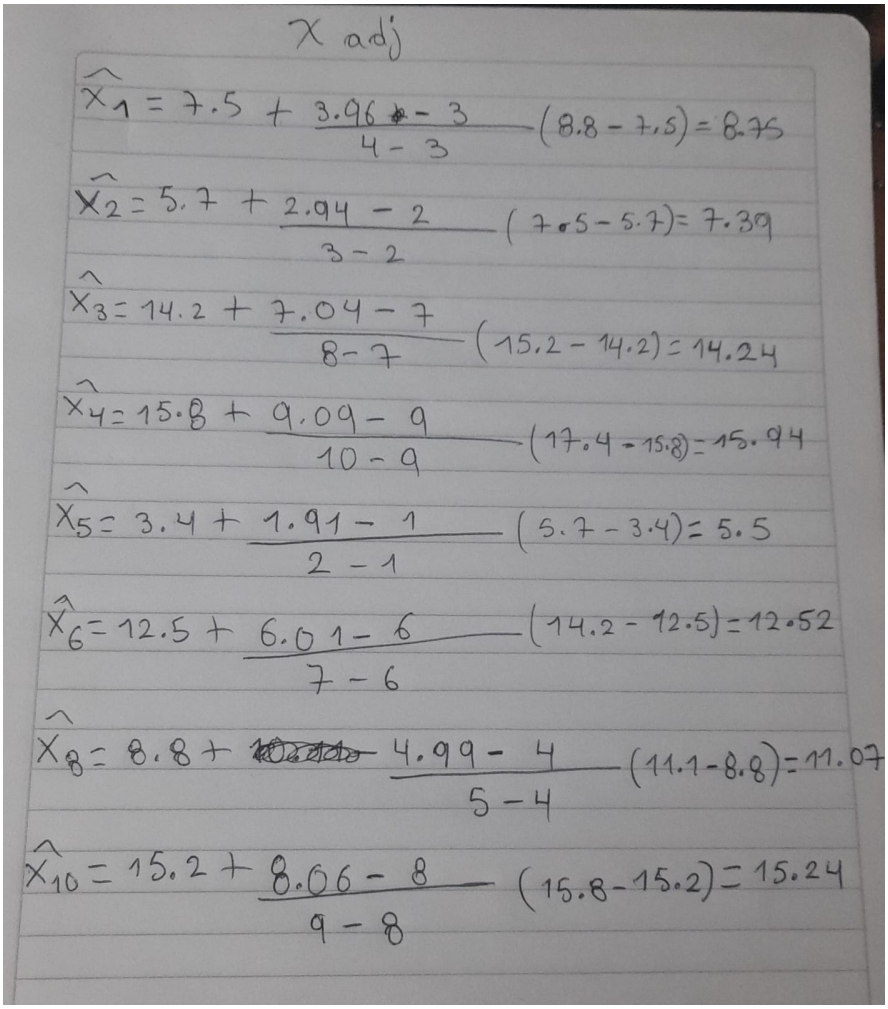

```
rangos$y.adj = c(154.03636,  90.37576, 192.65455,
                 246.43636, 117.57576, 156.96364,
                 254.45455, 142.43636,  47.90909, 207.15152)
```

$X_i s$ ajustados

Teniendo en cuenta los valores obtenidos, tenemos lo siguiente:

- Todas las corresponde al caso donde $\widehat{R(x_0)}$ esta entre 2 rangos de valores adyancentes, para este caso:

$$\widehat{X_i} = x_j + \frac{\widehat{R(x_i)} - R(x_j)}{R(x_k) - R(x_j)}(x_j - x_j)$$



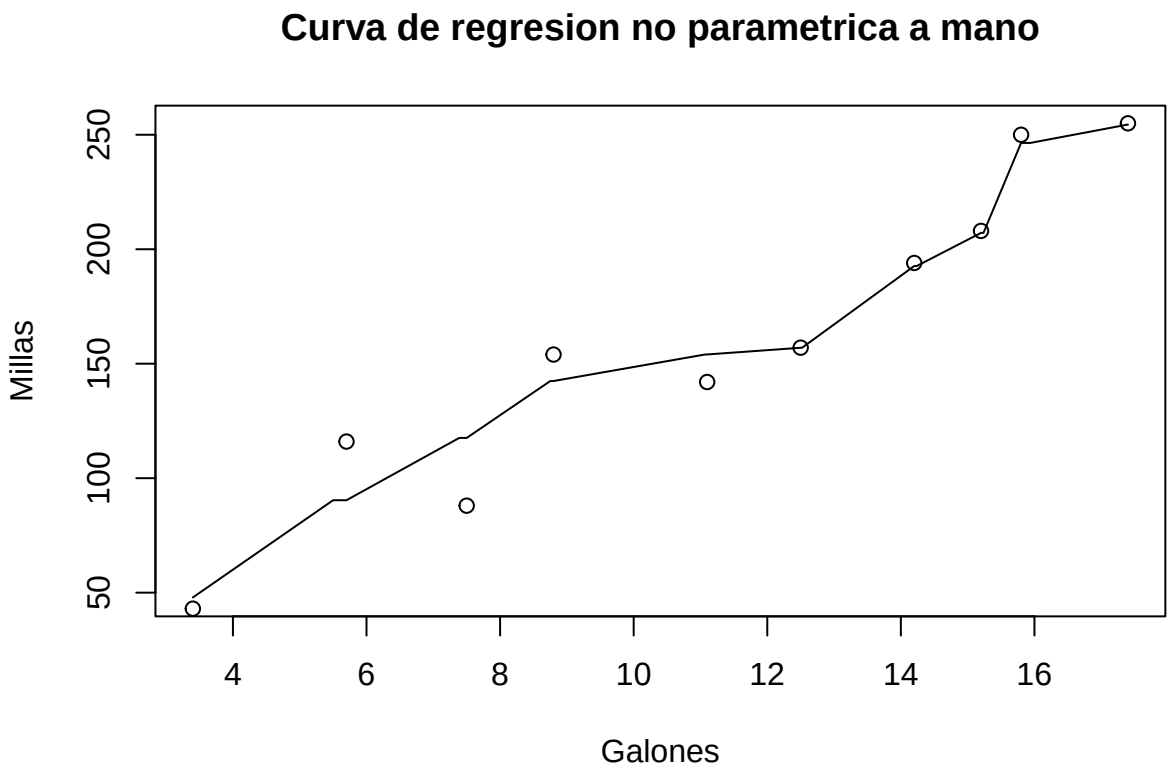
- La fila 7 al caso donde $\widehat{R(x_0)}$ es mayor al maximo de los rangos de X, por lo tanto se toma el valor mayor observado de X, que en este caso es **17.4**
- La fila 9 al caso donde $\widehat{R(x_0)}$ es menor al minimo de los rangos de X, por lo tanto se toma el valor menor observado de X, que en este caso es **3.4**

```
rangos$x.adj = c(8.751553, 7.388199, 14.237267,
                 15.939130, 5.5, 12.521118,
                 17.4, 11.071429, 3.4, 15.237267)
```

Paso 5 Ahora grafiquemos

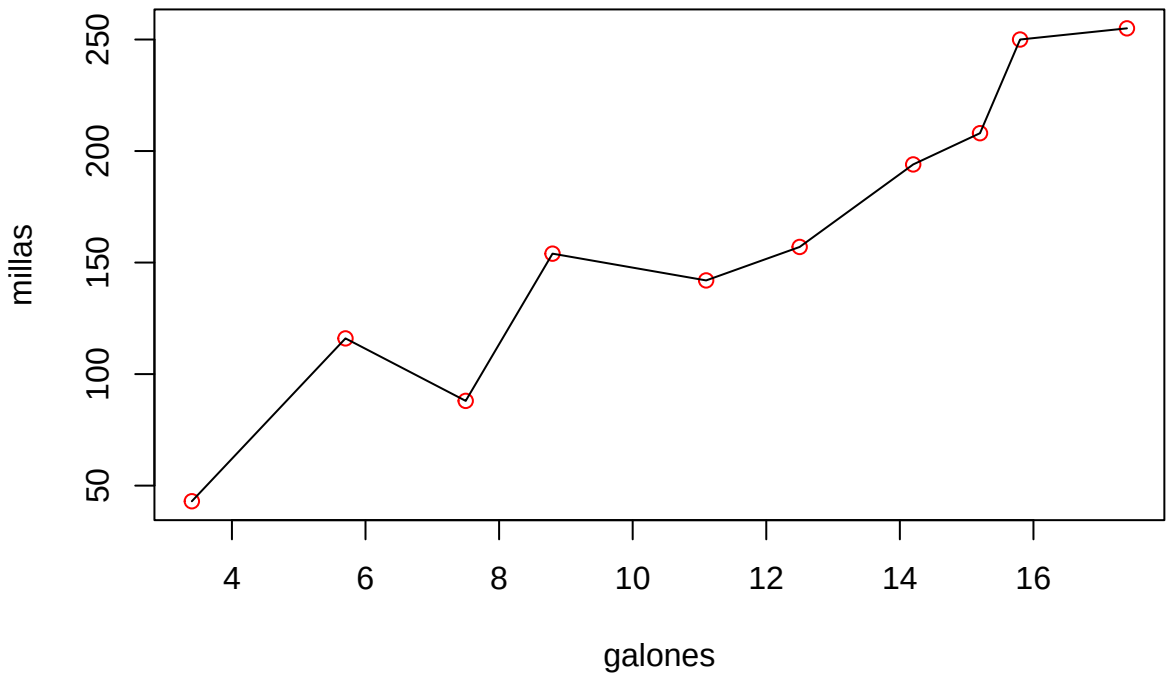
```
x1 = c(rangos$X, rangos$x.adj)
y1 = c(rangos$y, rangos$y.adj)

x1 = x1[order(x1)]
y1 = y1[order(y1)]

plot(x1, y1, type = "l", xlab = "Galones", ylab = "Millas",
     main = "Curva de regresion no parametrica a mano")
```

```
points(galones, millas)
```



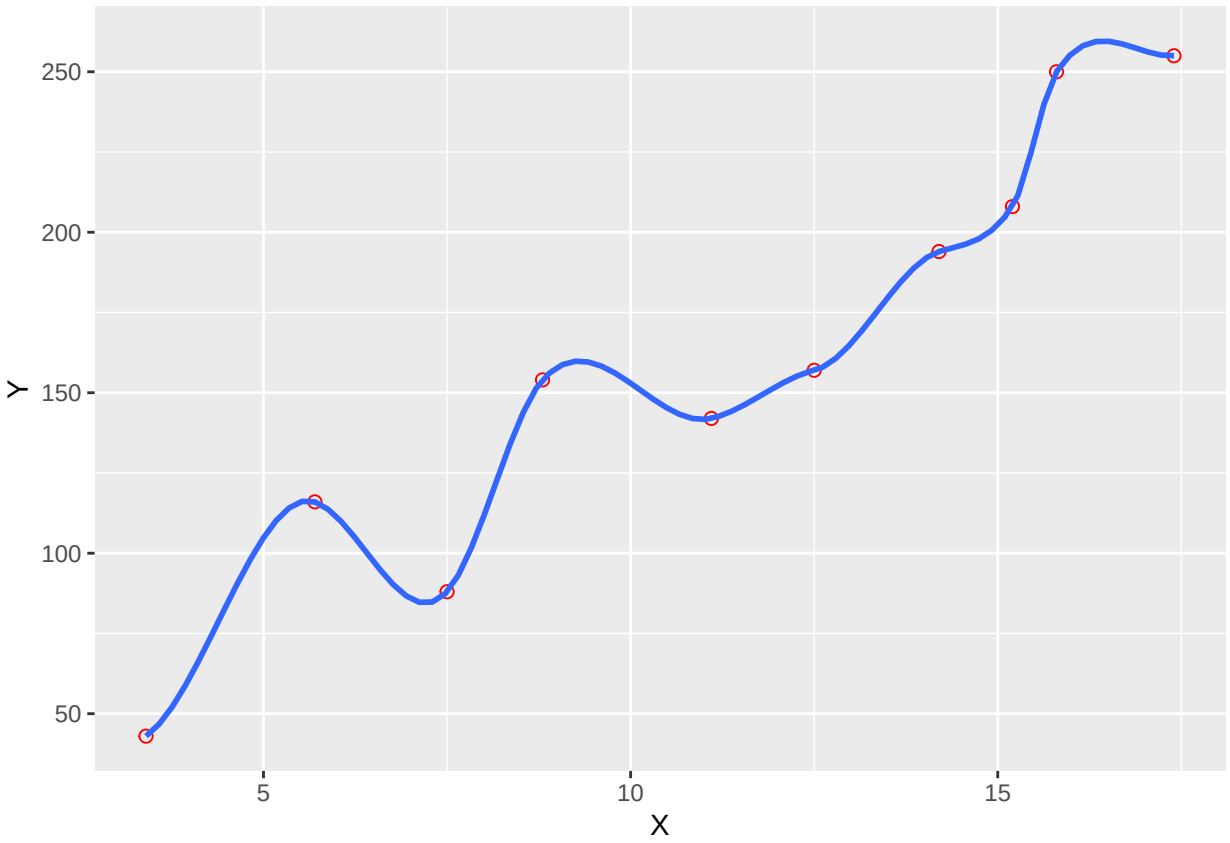
Con R

```
lw1 = loess(millas ~ galones, span = 0.35)
plot(millas ~ galones, pch = 1, cex = 1, col = "red")

por = data.frame(galones, lw1$fitted)
bla = por[order(por$galones),]

lines(bla$galones, bla$lw1.fitted)
```



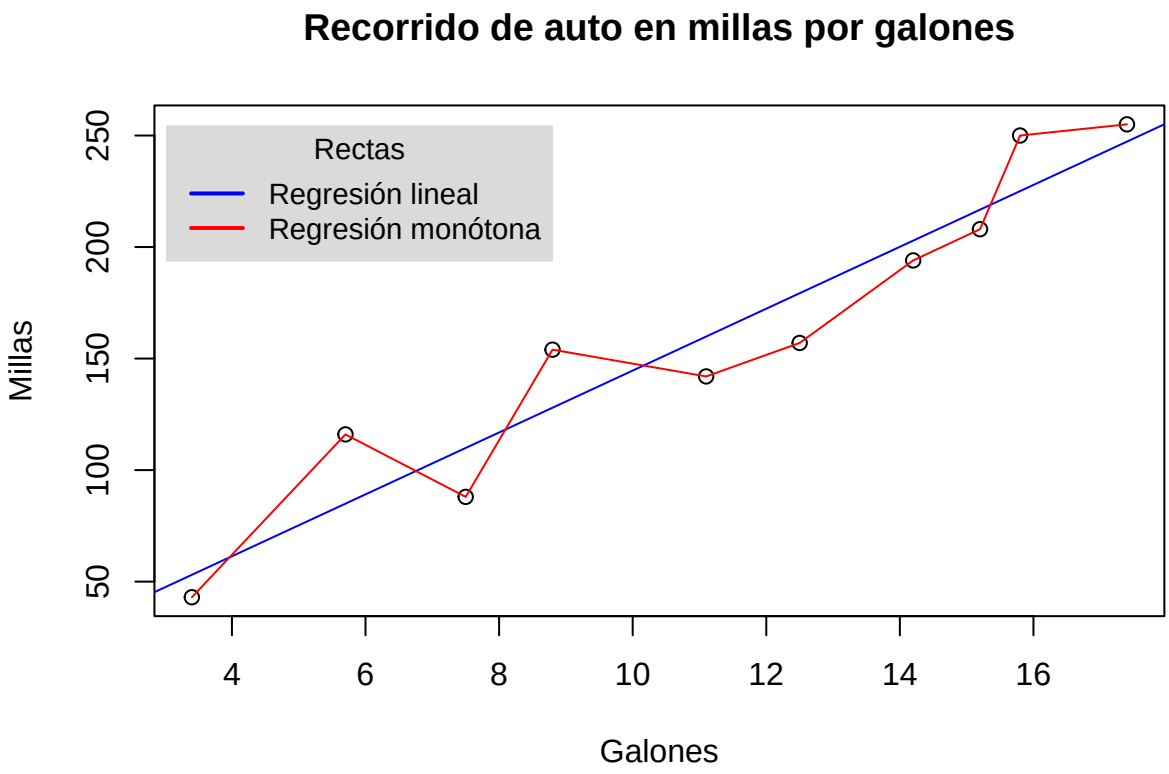

```
library(ggplot2)

ggplot(rangos, aes(X, Y)) +
  geom_point(size = 2, pch = 1, col = "red") +
  geom_smooth(span = 0.35, method = "loess")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Literal E

```
plot(galones, millas, xlab = "Galones", ylab = "Millas",
     title("Recorrido de auto en millas por galones"))

abline(regresion, col="blue")

por = data.frame(galones, lw1$fitted)
bla = por[order(por$galones),]

lines(bla$galones, bla$lw1.fitted, col= "red")

legend(x = 3, y = 255, legend = c("Regresión lineal", "Regresión monótona"),
       bg = rgb(0, 0, 0, alpha = 0.15), box.col = "white", lty=c(1,1), cex = 0.9, col = c("blue","red"))
```



Como podemos ver la regresion monotona hace un sobreajuste de los datos y esto no es algo idoneo, por otro lado la regresion lineal por minimos cuadrados no tiene este problema asi que por este lado el mejor modelo sera el modelo de regresion lineal.

Literal F

Para estimar el millaje para 16 galones se procede de la siguiente forma:

Primero observemos que 16 se encuentra entre los galones 15.8 y 17.4 que se muestran en la tabla inicial los cuales seran nuestro x_i y x_j respectivamente.Por otro lado tenemos la siguiente fórmula:

$$R(x_0) = R(x_i) + \frac{x_0 - x_i}{x_j - x_i}(R(x_j) - R(x_i)) = 9 + \frac{16 - 15.8}{17.4 - 15.8}(10 - 9) = 9.125$$

Luego encontramos el rango en y_0 asi:

$$R(y_0) = a_2 + b_2 R(x_0) = -36.53333 + 7.642424 * 9.125 = 33.203789$$

Luego se toma y_i y y_j como 250 y 255 respectivamente que serian las millas recorridas correspondientes a los 15.8 y 17.4 galones usados para el calculo anterior.

$$\widehat{Y_i} = y_i + \frac{R(y_0) - R(y_i)}{R(y_j) - R(y_i)}(y_j - y_i) = 250 + \frac{33.203789 - 32.23}{39.87 - 32.23}(255 - 250) = 250.63729$$

Finalmente se concluye que con 16 galones se recorren aproximadamente 250.63729 millas.