Correlación no Parametrica

Usando R

Humberto Vaquera Huerta

Verano 2023

Colegio de Postgraduados

Correlacion

Correlación de Pearson

El **coeficiente de correlación de Pearson** o la *r* de Pearson se define en estadística como la medida del grado de asociación o relación lineal entre dos variables.

asigna un valor entre - 1 y 1,

- donde 0 es *sin correlación*,
- 1 es *correlación positiva* y
- -1 es *correlación negativa*.

$$\rho(X, Y) = \frac{\mathbf{Cov}(X, Y)}{\sqrt{\mathbf{Var}(X)\mathbf{Var}(Y)}}.$$

Correlación de Pearson

Suponga se tiene observaciones de una muestra aleatoria de dos variables continuas X y Y: $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. El coeficiente de correlación de Pearson es:

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

Pasos para la prueba de hipótesis de ρ

Paso 1: Hipótesis

hipótesis nula y alternativa:

- Hipótesis nula:$H_0 : \rho = 0$
- Hipótesis alternativa: $H_a : \rho \neq 0$

Paso 2: Estadística de prueba:

$$t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Paso 3: Valor de tablas $t_{(n-2)}$ de t-student y p-value

paso 4: Decisión: Rechazo H_0 si $p - value$ menor que α

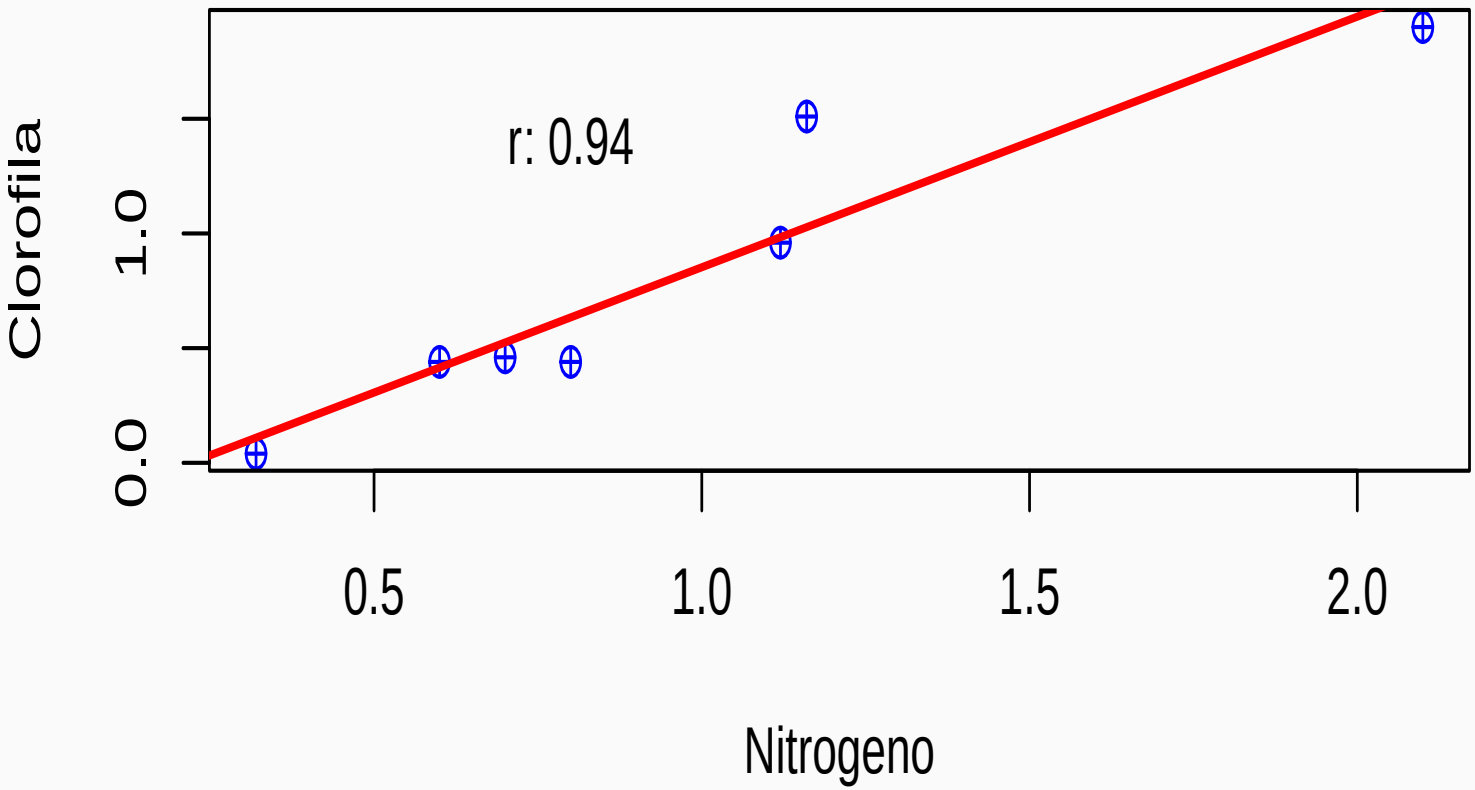
Ejemplo: Datos de arroz

Para ilustrar la asociación entre dos variables usamos datos en nitrógeno proteico soluble mg/hoja (X_1) y clorofila total mg/hoja (X_2) en hojas obtenidas de siete muestras de la variedad de arroz IR8.

Nitrogeno	Clorofila
0.6	0.44
1.12	0.96
2.1	1.9
1.16	1.51
0.7	0.46
0.8	0.44
0.32	0.04

```
Nitrogeno=c(0.6,1.12,2.1,1.16,0.7,0.8,0.32)
Clorofila=c(0.44,0.96,1.9,1.51,0.46,0.44,0.04)
plot(Nitrogeno, Clorofila, pch = 10, col = "blue")
abline(lm(Clorofila ~ Nitrogeno), col="red",lwd=2)
text(paste("r:",round(cor(Nitrogeno, Clorofila),2)),x=.8,y=1.4)
```

Ejemplo: Datos de arroz



Correlacion de Pearson

```r
Nitrogeno=c(0.6,1.12,2.1,1.16,0.7,0.8,0.32)
Clorofila=c(0.44,0.96,1.9,1.51,0.46,0.44,0.04)
datos=data.frame(cbind(Nitrogeno,Clorofila))
library(correlation)
cor_test(datos,"Nitrogeno", "Clorofila")
```

```
Parameter1 | Parameter2 |    r |       95% CI | t(5) |        p
-------------------------------------------------------------------
Nitrogeno  |  Clorofila | 0.94 | [0.65, 0.99] | 6.26 | 0.002**

Observations: 7
```

Correlacion de Pearson

```
Nitrogeno=c(0.6,1.12,2.1,1.16,0.7,0.8,0.32)
Clorofila=c(0.44,0.96,1.9,1.51,0.46,0.44,0.04)
datos=data.frame(cbind(Nitrogeno,Clorofila))
library(correlation)
corr=correlation(datos)
display(corr)
```

Table 2: Correlation Matrix (pearson-method)

Parameter1	Parameter2	r	95% CI	t(5)	p
Nitrogeno	Clorofila	0.94	(0.65, 0.99)	6.26	0.002**

p-value adjustment method: Holm (1979) Observations: 7

Supuestos: X y Y variables normales

Correlacion noparametrica

Coeficiente de correlacion Spearman

Si tiene dos variables numéricas que no están relacionadas linealmente, o si una o ambas de sus variables son variables ordinales, se puede medir la fuerza y la dirección de su relación utilizando la correlación no paramétrica. El más común de ellos es el coeficiente de **correlación de rangos de Spearman**, ρ, que considera los rangos de los valores para las dos variables.

- Spearman es equivalente a Pearson sobre los datos con rangos. Entonces,ρ esta entre -1 y 1.

Coeficiente de correlacion Spearman

Supuestos:

- Muestras aleatorias
- Observaciones independientes
- La relación entre las dos variables es monótona (evaluada visualmente con un diagrama de dispersión).

$$r_{s_{xy}} = \frac{cov(rank_x, rank_y)}{SD(rank_x) \times SD(rank_y)}$$

Tranformado a rangos para calculo de Spearman

```
Nitrogeno=c(0.6,1.12,2.1,1.16,0.7,0.8,0.32)
Clorofila=c(0.44,0.96,1.9,1.51,0.46,0.44,0.04)
datos=data.frame(cbind(Nitrogeno,Clorofila))
library(effectsize)
library(datawizard)
datos_rangos=ranktransform(datos)
knitr::kable(datos_rangos)
```

Nitrogeno	Clorofila
2	2.5
5	5.0
7	7.0
6	6.0
3	4.0
4	2.5
1	1.0

Coeficiente de correlacion Spearman

```
library(correlation)
corr_spearman=correlation(datos_rangos)
display(corr_spearman)
```

Table 4: Correlation Matrix (pearson-method)

Parameter1	Parameter2	r	95% CI	t(5)	p
Nitrogeno	Clorofila	0.94	(0.62, 0.99)	6.00	0.002**

p-value adjustment method: Holm (1979) Observations: 7

Obtencion directa en r

```r
Nitrogeno=c(0.6,1.12,2.1,1.16,0.7,0.8,0.32)
Clorofila=c(0.44,0.96,1.9,1.51,0.46,0.44,0.04)
cor(Nitrogeno, Clorofila,method="spearman")
```

```
[1] 0.936975
```

```r
cor.test(Nitrogeno, Clorofila,method="spearman")
```

```
    Spearman's rank correlation rho

data:  Nitrogeno and Clorofila
S = 3.5294, p-value = 0.001851
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.936975
```

Coeficiente de correlacion Spearman

```
Nitrogeno=c(0.6,1.12,2.1,1.16,0.7,0.8,0.32)
Clorofila=c(0.44,0.96,1.9,1.51,0.46,0.44,0.04)
datos=data.frame(cbind(Nitrogeno,Clorofila))
library(correlation)
corr=correlation(datos,method="spearman")
display(corr)
```

Table 5: Correlation Matrix (spearman-method)

Parameter1	Parameter2	rho	95% CI	S	p
Nitrogeno	Clorofila	0.94	(0.61, 0.99)	3.53	0.002**

p-value adjustment method: Holm (1979) Observations: 7

Coeficiente τ de Kendall

El coeficiente de correlación τ de Kendall,es una medida de asociación no paramétrica basada en el número de concordancias y discordancias en observaciones pareadas.

Suponga dos observaciones(X_i, Y_i) y (X_j, Y_j) son **concordantes** si están en el mismo orden con respecto a cada variable. Es decir, si

- si $X_i < X_j$ y $Y_i < Y_j$ o si
- $X_i > X_j$ y $Y_i > Y_j$

son **disconcordantes** si $X_i < X_j$ y $Y_i > Y_j$ o $X_i > X_j$ y $Y_i < Y_j$

se tienen **empates** si $X_i = X_j$ y $Y_i = Y_j$

Coeficiente τ de Kendall

El número total de pares que se pueden construir para un tamaño de muestra de n es $N = \binom{n}{2} = \frac{1}{2}n(n-1)$

El N se puede descomponer en estas cinco cantidades:$N = P + Q + X_0 + Y_0 + (XY)_0$

donde P es el número de pares concordantes, Q es el número de pares discordantes,X_0 es el número de pares empatados solo en la variable X, Y_0 es el número de pares vinculados solo en la variable Y, y $(XY)_0$ es el número de pares empatados tanto en X como en Y.

La tau-b de Kendall para medir la asociación de orden entre las variables X e Y viene dada por :

$$\tau_{xy} = \frac{P - Q}{\sqrt{(P + Q + X_0)(P + Q + Y_0)}}$$

Coeficiente τ de Kendall

O de manera alterna:

$$\tau_{xy} = \frac{2}{n(n-1)} \sum_{i<j} sign(x_i - x_j) \times sign(y_i - y_j)$$

$$\tau_{xy}$$

esta entre 1 y -1

Coeficiente τ de Kendall

Ejemplo:

```
Nitrogeno=c(0.6,1.12,2.1,1.16,0.7,0.8,0.32)
Clorofila=c(0.44,0.96,1.9,1.51,0.46,0.44,0.04)
cor(Nitrogeno, Clorofila,method="kendall")
```

```
[1] 0.8783101
```

```
cor.test(Nitrogeno, Clorofila,method="kendall")
```

```
    Kendall's rank correlation tau

data:  Nitrogeno and Clorofila
z = 2.7344, p-value = 0.006249
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
```

Coeficiente τ de Kendall

Ejemplo:

```
Nitrogeno=c(0.6,1.12,2.1,1.16,0.7,0.8,0.32)
Clorofila=c(0.44,0.96,1.9,1.51,0.46,0.44,0.04)
datos=data.frame(cbind(Nitrogeno,Clorofila))
library(correlation)
corr_kendall=correlation(datos,method="kendall")
display(corr_kendall)
```

Table 6: Correlation Matrix (kendall-method)

Parameter1	Parameter2	tau	95% CI	z	p
Nitrogeno	Clorofila	0.88	(0.55, 0.97)	2.73	0.006**

p-value adjustment method: Holm (1979) Observations: 7

Coeficiente τ de Kendall

Ejemplo:

```
Nitrogeno=c(0.6,1.12,2.1,1.16,0.7,0.8,0.32)
Clorofila=c(0.44,0.96,1.9,1.51,0.46,0.44,0.04)
library(Kendall)
Kendall(Nitrogeno,Clorofila)
```

```
tau = 0.878, 2-sided pvalue =0.0098091
```

Goodman Kruskal *tau*

Los estadísticos de Goodman-Kruskal son medidas de asociación entre variables categóricas.

oodman-Kruskal tau mide la asociación para tabulaciones cruzadas de variables de nivel nominal.

Goodman-Kruskal tau se basa en la asignación de categorías al azar. Mide el porcentaje de mejora en la predictibilidad de la variable dependiente (variable de columna o fila) dado el valor de otras variables (variables de fila o columna).

Goodman Kruskal *tau*

```
library(DescTools)
tab <- as.table(rbind(c(26,26,23,18,9),c(6,7,9,14,23)))
# Goodman Kruskal's tau-a C|R
GoodmanKruskalTau(tab, direction="column", conf.level=0.95)
```

```
       tauA      lwr.ci      upr.ci
0.041216580 0.009920576 0.072512583
```

```
# Goodman Kruskal's tau-a R|C
GoodmanKruskalTau(tab, direction="row", conf.level=0.95)
```

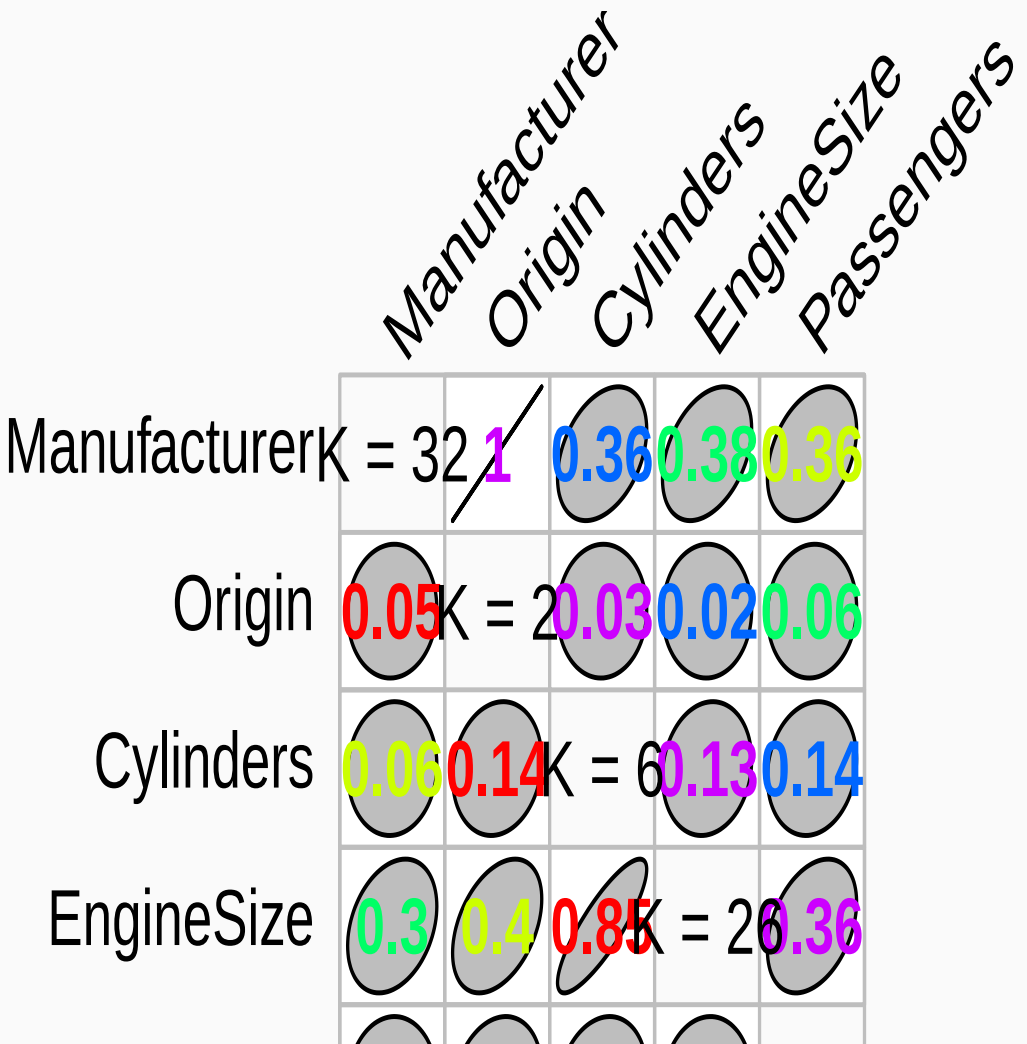
```
      tauA     lwr.ci     upr.ci
0.16523315 0.04921484 0.28125146
```

Goodman Kruskal *tau*

data for a clinical trial of a drug therapy to control pain. Theclinical trial investigates whether adverse responses increase with larger drug doses. Subjects re-ceive either a placebo or one of four drug doses. An adverse response is recorded asAdverse=‘Yes’;otherwise, it is recorded asAdverse=‘No’. The number of subjects for each drug dose and responsecombination is contained in the variableCount

Goodman Kruskal *tau*

```
library(MASS)
library(GoodmanKruskal)
varSet1 <- c("Manufacturer", "Origin", "Cylinders", "EngineSize", "Passengers")
CarFrame1 <- subset(Cars93, select = varSet1)
GKmatrix1 <- GKtauDataframe(CarFrame1)
plot(GKmatrix1)
```



Ejemplo de aplicacion

Datos de Aquifero.

The Ogallala aquifer was investigated to estimate relations between uranium and other concentrations in its waters. Below are the concentrations of uranium and total dissolved solids.

```
urants2 <- read.csv("D:/cursos/no param/Curso Noparametrica/uran
head(urants2)
```

```
  X       TDS URANIUM HCO3 DEFINITION
1 1  682.6499  0.9315    0      <=50%
2 2  819.1199  1.9380    0      <=50%
3 3  303.7600  0.2919    0      <=50%
4 4 1151.3999 11.9042    0      <=50%
5 5  582.4199  1.5674    0      <=50%
6 6 1043.3899  2.0623    0      <=50%
```

Datos

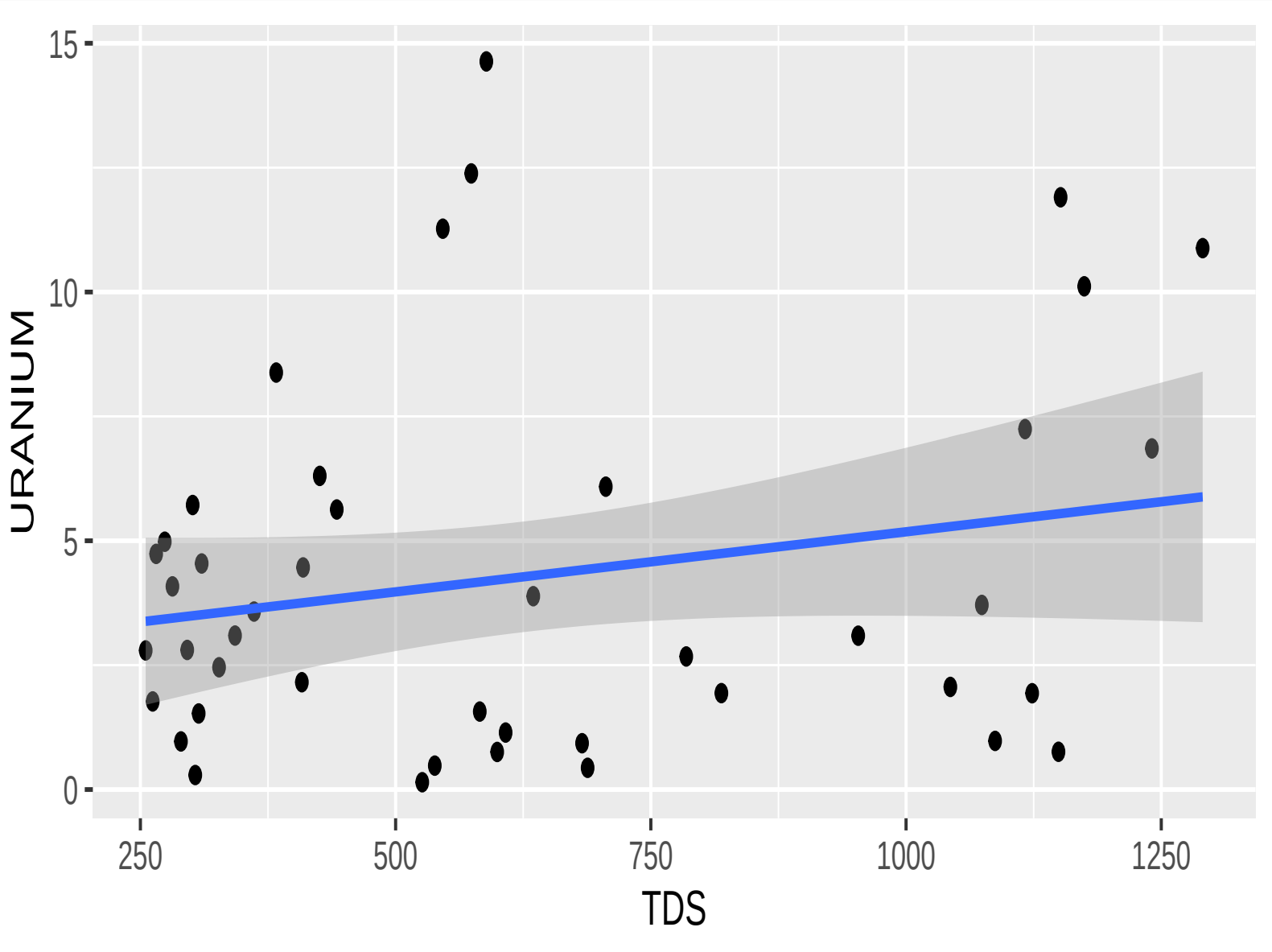
TDS	URANIUM	TDS	URANIUM	TDS	URANIUM	TDS	URANIUM
682.649902	0.93149996	1116.58984	7.24459934	599.5	0.75509989	588.859985	14.6341991
819.119873	1.93799996	301.199951	5.71899986	1240.81006	6.85589981	574.109985	12.3834992
303.76001	0.29189998	265.449951	4.73659992	538.349976	0.48059994	307.089966	1.52909994
1151.3999	11.9041977	295.879944	2.80569983	607.75	1.14519978	409.369995	4.46469975
582.419922	1.56739998	442.359985	5.62899971	705.889893	6.08759975	327.069946	2.45739985
1043.38989	2.06229973	342.709961	3.09499979	1290.56982	10.8822994	425.689941	6.30419922
634.839966	3.88579989	361.299988	3.57739973	526.089966	0.14730001	310.049988	4.54409981
1087.25	0.97719991	262.069946	1.77109981	784.679932	2.67409945	289.75	0.96719992
1123.51001	1.93539977	546.219971	11.2723999	953.139893	3.09179974	408.179993	2.15679979
688.089966	0.43669999	273.889954	4.98069954	1149.31006	0.75919986	383.039978	8.38099861
1174.5398	10.1141987	281.379944	4.08329964	1074.21997	3.7100997	255.189972	2.7956996

Referencia de datos

Helsel, D.R., Hirsch, R.M., Ryberg, K.R., Archfield, S.A., and Gilroy, E.J., 2020, Statistical methods in water resources: U.S. Geological Survey Techniques and Methods, book 4, chapter A3, 458 p., https://doi.org/10.3133/tm4a3. [Supersedes USGS Techniques of Water-Resources Investigations, book 4, chapter A3, version 1.1.]

grafica

```
library(ggplot2)
ggplot(urants2, aes(TDS,URANIUM)) + geom_point() + geom_smooth(method = "lm")
```



Correlaciones Pearson

```
datos2=urants2[,3:4]
library(correlation)
corr=correlation(datos2)
display(corr)
```

Table 7: Correlation Matrix (pearson-method)

Parameter1	Parameter2	r	95% CI	t(42)	p
URANIUM	HCO3	0.23	(-0.07, 0.49)	1.53	0.133

p-value adjustment method: Holm (1979) Observations: 44

Correlaciones Spearman

```
datos2=urants2[,3:4]
library(correlation)
corr=correlation(datos2, method="spearman")
display(corr)
```

Table 8: Correlation Matrix (spearman-method)

Parameter1	Parameter2	rho	95% CI	S	p
URANIUM	HCO3	0.34	(0.03, 0.58)	9435.88	0.026*

p-value adjustment method: Holm (1979) Observations: 44

Correlaciones Kendall

```
datos2=urants2[,3:4]
library(correlation)
corr=correlation(datos2, method="kendall")
display(corr)
```

Table 9: Correlation Matrix (kendall-method)

Parameter1	Parameter2	tau	95% CI	z	p
URANIUM	HCO3	0.28	(0.08, 0.45)	2.20	0.028*

p-value adjustment method: Holm (1979) Observations: 44