Análisis Multivariado

ID 033521 - Clase 4901

Lina Maria Acosta Avena

Ciencia de Datos
Departamento de Matemáticas
Pontificia Universidad Javeriana

Semana 3: 29/07/24 – 03/08/24

Introducción

Recuerde que, para **evaluar** la **normalidad**,

✓ en el caso **univariado** tenemos

- **Gráficos**: Histograma, Boxplots y Q-Q plot (n grandes).
- **Pruebas de Hipótesis**: Shapiro-Wilk, Kolmogorov-Smirnoff, Anderson-Darling, entre otras.

✓ en el caso **bivariado** podemos (informalmente)

- construir el **diagrama de dispersión** entre las dos variables y observar si se forma una elipse (sospecha de normalidad bivariada).
- construir el **Q-Q plot** con los valores de las **distancias de Mahalanobis** con las respectivas ordenadas de los cuantiles de la distribución chi-cuadrado.

Introducción

Observaciones:

- *En el Q-Q plot se grafican los cuantiles muestrales contra los cuantiles (teóricos) que se esperaría observar si las observaciones realmente provienen de una distribución particular.*
- *Si en el Q-Q plot se forma una recta, hay sospecha de la distribución en cuestion, en caso contrario, hay indicios de desvíos de la distribución en cuestion.*
- *En este gráfico también se pueden detectar outliers.*

Introducción

Tenga presente que normalidad univariada o bivariada NO implican que el vector aleatorio $\mathbf{X}_{p\times 1}$ siga una normal mutivariada. Sin embargo, Si las observaciones fueran generadas por un distribución **normal multivariada**, TODAS las distribuciones **bivariadas** deberían ser **normales** y los **contornos** de densidad constante deberían ser **elipses**.

Introducción

Example

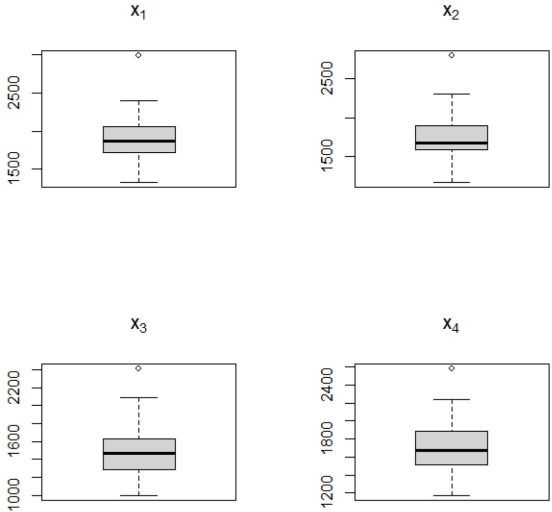
Los datos del archivo `Rigidez.xlsx` corresponden a 4 medidas de rigidez de $n = 30$ tablas. La primera medición (x_1) implica enviar una onda de choque hacia abajo por la tabla, la segunda medición (x_2) se determina mientras se hace vibrar a la tabla y las dos últimas mediciones (x_3, x_4) se obtienen a partir de pruebas estáticas.

Obs: Example 4.14 de Johnson and Wichern (2014), Applied Multivariate Statistical Analysis, 6th Edition, pp. 186

```
# --- Example 4.14 from Johnson and Wichern --- #
require(tidyverse)
Rigidez <- read_excel("D:/Desktop/Rigidez.xlsx")
Datos<-Rigidez %>% as.data.frame()

# --- Boxplot
par(mfrow=c(2,2))
boxplot(Datos$x1, main=expression(x[1]),xlab="",ylab="")
boxplot(Datos$x2, main=expression(x[2]),xlab="",ylab="")
boxplot(Datos$x3, main=expression(x[3]),xlab="",ylab="")
boxplot(Datos$x4, main=expression(x[4]),xlab="",ylab="")
```

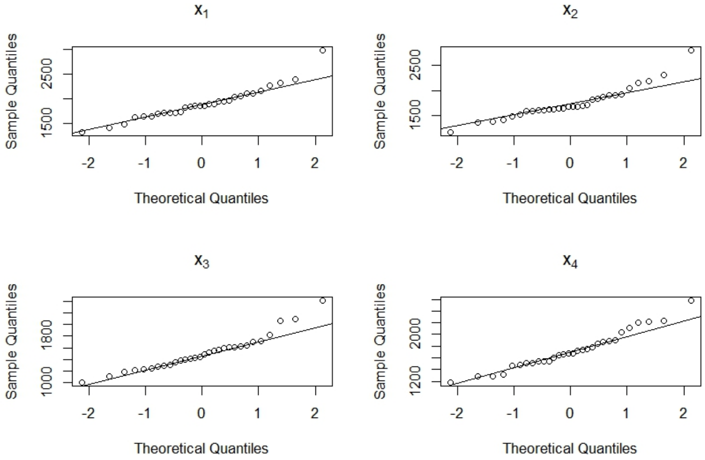
Introducción



Introducción

```
# --- QQplot
par(mfrow=c(2,2))
qqnorm(Datos$x1, main=expression(x[1]))
qqline(Datos$x1)
qqnorm(Datos$x2, main=expression(x[2]))
qqline(Datos$x2)
qqnorm(Datos$x3, main=expression(x[3]))
qqline(Datos$x3)
qqnorm(Datos$x4, main=expression(x[4]))
qqline(Datos$x4)
```

Introducción



Introducción

La **linealidad del Q-Q plot** se puede medir calculando el **coeficiente de correlación** para los puntos del gráfico:

$$r_Q = \frac{\sum_{j=1}^{n} \left(x_{(j)} - \overline{x}\right)\left(q_{(j)} - \overline{q}\right)}{\sqrt{\sum_{j=1}^{n} \left(x_{(j)} - \overline{x}\right)^2}\sqrt{\sum_{j=1}^{n} \left(q_{(j)} - \overline{q}\right)^2}}$$

donde

- $x_{(j)}$ es la $j-$ésima observación ordenada
- $q_{(j)}$ es el cuantil de la normal para la posición $(j - 0.5)/n$
- $\overline{x}$ es la media de $x_{(j)}$
- $\overline{q}$ es la media de $q_{(j)}$

Introducción

La hipótesis de normalidad

$$\begin{aligned}\mathsf{H}_0 &: \text{Los datos } \underline{\text{provienen}} \text{ de una poblacion } \underline{\text{Normal}}\\ \mathsf{H}_1 &: \text{Los datos } \underline{\text{NO provienen}} \text{ de una Poblacion } \underline{\text{Normal}}\end{aligned}$$

es **rechazada** a un nivel de significancia α, cuando

$$r_Q < r_Q(\alpha, n)$$

donde $r_Q(\alpha, n)$ son los valores de la Tabla 4.2 de Johnson and Wichern (2013), Applied Multivariate Statistical Analysis, pp. 181:

Introducción

Sample size n	Significance levels α .01	.05	.10
5	.8299	.8788	.9032
10	.8801	.9198	.9351
15	.9126	.9389	.9503
20	.9269	.9508	.9604
25	.9410	.9591	.9665
30	.9479	.9652	.9715
35	.9538	.9682	.9740
40	.9599	.9726	.9771
45	.9632	.9749	.9792
50	.9671	.9768	.9809
55	.9695	.9787	.9822
60	.9720	.9801	.9836
75	.9771	.9838	.9866
100	.9822	.9873	.9895
150	.9879	.9913	.9928
200	.9905	.9931	.9942
300	.9935	.9953	.9960

Introducción

Para los datos del ejemplo tenemos

	X_1	X_2	X_3	X_4
r_Q	0.9599	0.9504	0.9635	0.9803

Observe que para los niveles de significancia del 5 % y del 10 %, sólo X_4 provendría de una población normal; mientras que al 1 %, las 4 variables serían normales.

Introducción

```
#--- linealidad del QQplot - r_Q
x_j<-sort(Datos$x1)
x_barra<-mean(x_j)
q<-c()
for(j in 1:30){
               q[j]<-qnorm((j-0.5)/30)
              }
q_barra<-mean(q)
num_rQ<-sum((x_j-x_barra)%*%(q-q_barra))
den_rQ<-(sqrt(sum((x_j-x_barra)^2)))*
        (sqrt(sum((q-q_barra)^2)))
r_Q<-num_rQ/den_rQ
```

Introducción

Por otro lado, podemos probar (**formalmente**) las hipótesis de **normalidad univariada** (para cada X_i) usando la prueba de **Shapiro-Wilk**:

```
shapiro.test(Datos$x1); shapiro.test(Datos$x2)
shapiro.test(Datos$x3); shapiro.test(Datos$x4)
```

De ahí se tiene que

$$p-\mathrm{valor}_{x_1} = 0.05118 \qquad p-\mathrm{valor}_{x_2} = 0.01746$$
$$p-\mathrm{valor}_{x_3} = 0.05805 \qquad p-\mathrm{valor}_{x_4} = 0.33370$$

Observe que para un nivel de significancia del 5 % las distribuciones marginales de X_1, X_3 y X_4 son normales. X_2 es normal a un nivel de significancia del 1 %.

Pruebas de Bondad y Ajuste a la Normal Multivariada

Pruebas de Bondad y Ajuste a la Normal Multiv.

Si las observaciones fueran generadas por un distribución **normal multivariada**:

- ✓ **Cada** una de las distribuciones **bivariadas** serían **normales** y los **contornos** de densidad constante deberían ser **elipses**.
- ✓ El **scatterplot** de **cada par** de variables debería mostrar un patrón aproximadamente **elíptico**.
- ✓ el conjunto de **puntos bivariados x** tal que

$$(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \leq \chi^2_2(\alpha)$$

  tendrá una **probabilidad** α.

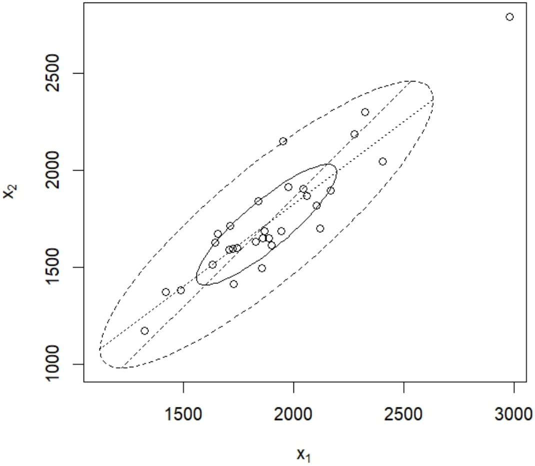
Por ejemplo, si $\underline{\alpha = 0.5}$, **se esperaría que alrededor del 50 % de las observaciones caigan dentro de la elipse**

$$(\mathbf{x} - \overline{\mathbf{x}})^\top S^{-1} (\mathbf{x} - \overline{\mathbf{x}}) \leq \chi^2_2(0.5)$$

En caso contrario, la normalidad es considerada sospechosa. En otras palabras, si los datos proceden de una distribución normal, el 50 % de las distancias calculadas deberian ser $\leq 1.39 = \chi^2_2$.

```
qchisq(0.5,2)
[1] 1.386294
```

Pruebas de Bondad y Ajuste a la Normal Multiv.



Pruebas de Bondad y Ajuste a la Normal Multiv.

```
 #--- Gráficos de elipses
require(MVA)
x1x2<-data.frame(Datos$x1,Datos$x2)
y1<-as.matrix.data.frame(x1x2)
bvbox(y1, method ="robust",
      xlab = expression(x[1]),
      ylab = expression(x[2]))
```

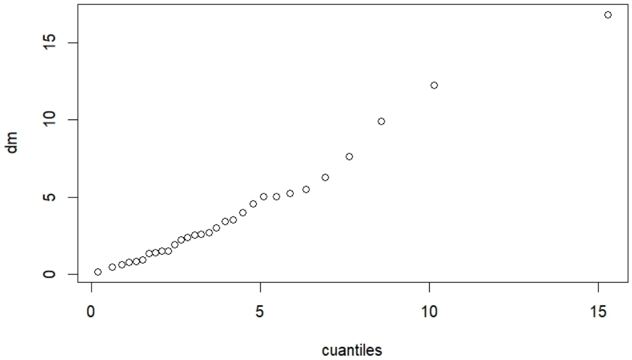
Más **formalmente**, para evaluar la **normalidad conjunta** (p variables)

- ✓ se pueden calcular las **distancias cuadráticas generalizadas**

$$d_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})^\top \mathrm{S}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}}) \qquad j = 1, \ldots, n$$

las cuales deben seguir una **distribución** χ_p^2. **Bajo normalidad**, el **gráfico** debería mostrar un **patrón lineal a través del origen y pendiente 1**. Patrones por fuera de la linealidad sugieren falta de normalidad.

Pruebas de Bondad y Ajuste a la Normal Multiv.



```
# ---  QQplot normal multivariada
X<-as.matrix(Datos)
Xbarra<-colMeans(X)
S<-cov(X)
dm<-mahalanobis(X,Xbarra,S)
cuantiles<-qchisq(ppoints(length(X)),df=4)
qqplot(cuantiles,dm)
```

Pruebas de Bondad y Ajuste a la Normal Multiv.

✓ se puede probar las **hipótesis**

$$\begin{aligned} \mathsf{H}_0 &: \text{Los datos \underline{provienen} de una poblacion \underline{Normal Multivariada}} \\ \mathsf{H}_1 &: \text{Los datos \underline{NO provienen} de una poblacion \underline{Normal Multivariada}} \end{aligned}$$

o matemáticamente

$$\begin{aligned} \mathsf{H}_0 &: \mathbf{X}_{p\times 1} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \mathsf{H}_1 &: \mathbf{X}_{p\times 1} \nsim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \end{aligned}$$

Pruebas de Bondad y Ajuste a la Normal Multiv.

✓ se puede probar las **hipótesis**

$$
\begin{aligned}
\mathsf{H}_0 &: \text{Los datos } \underline{\text{provienen}} \text{ de una poblacion } \underline{\text{Normal Multivariada}} \\
\mathsf{H}_1 &: \text{Los datos } \underline{\text{NO provienen}} \text{ de una poblacion } \underline{\text{Normal Multivariada}}
\end{aligned}
$$

o matemáticamente

$$
\begin{aligned}
\mathsf{H}_0 &: \mathbf{X}_{p\times 1} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\mathsf{H}_1 &: \mathbf{X}_{p\times 1} \nsim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})
\end{aligned}
$$

Existen **varias estadísticas de pruebas** y **varios paquetes** para probar estas hipótesis, dentro de los cuales tenemos:

Pruebas de Bondad y Ajuste a la Normal Multiv.

```
# ----- Prueba de Shapiro ----- #
require(mvShapiroTest)
mvShapiro.Test(X)

# ----- Otras Pruebas ----- #
require(MVN)
mvn(X, mvnTest="mardia")    # test de Mardia
mvn(X, mvnTest="hz")        # test de Henze-Zirkler
mvn(X, mvnTest="royston")   # test de Royston
mvn(X, mvnTest="dh")        # test de Doornik-Hansen
```

Más detalles en Selcuk et al (2014), MVN An R Package for Assessing Multivariate Normality, The R Journal, vol. 6(2)

Pruebas de Bondad y Ajuste a la Normal Multiv.

En el QQplot anterior, los dos **puntos más distantes** corresponden a las **distancias** más **grandes** (observaciones 9 y 16).

```
> dm
 [1]  0.5986403  5.4765410  7.6242571  5.2337826  1.3985786
 [6]  2.2198583  5.0205375  1.4854047 12.2667558  0.7650085
[11]  1.9291087  0.4650468  2.6967437  0.1285018  0.9452695
[16] 16.8468161  3.5024405  3.9904867  1.3608102  1.4745213
[21]  9.9436642  5.0536019  0.7971562  2.5484135  4.5785994
[26]  3.4046873  2.3799906  2.9944696  6.2880650  2.5822428
```

Las observaciones correspondientes a estas distancias son llamadas **atípicas** y aunque parecieran no pertenecer al patrón de variabilidad que siguen las otras observaciones, **hacen parte del grupo** y pueden conducir a una **mejor comprensión del fenómeno estudiado**.

Detección de Datos Atípicos

Detección de Datos Atípicos

¿Cómo **detectar** observaciones **atípicas**?

✓ En el caso **univariado** se pueden construir **Boxplot** para cada variable.

✓ En el caso **bivariado** el **diagrama de dispersión** para cada par de variables.

✓ En el caso **multivariado**:

- se hacen las **estandarizaciones**[1] para cada variable

$$z_{jk} = \frac{x_{jk} - \overline{x}_k}{\sqrt{s_{kk}}}$$

y se **examinan conjuntamente todos** los nk valores para detectar aquellos inusuales (muy grandes o muy pequeños).

[1] cuando las unidades de medida son muy diferentes

Detección de Datos Atípicos

Observación: si $n = 100$ y $p = 5$, tendremos $np = 500$. Para una ***normal estándar***

$$P[|Z| > 3] = 0.0026$$

luego ***se esperaria*** *que 1 o 2 sean menores que -3 y mayores que 3, puesto que*

$$500 \times 0.0026 = 1.3$$

```
# --- P[|Z|>3]
1-(pnorm(3)-pnorm(-3))
[1] 0.002699796
```

Detección de Datos Atípicos

- se calculan las **distancias generalizadas**

$$d_j^2 = (\mathbf{x}_j - \overline{\mathbf{x}})^\top \mathrm{S}^{-1}(\mathbf{x}_j - \overline{\mathbf{x}}) \qquad j = 1, \ldots, n$$

y se **examinan** aquellas que sean **inusualmente más grandes**. *Observación: Si $n = 100$ y $p = 5$, se debería esperar que 5 observaciones excedan el percentil* 0.05*-superior de la distribución chicuadrado.*

Note que $d_{16}^2 = 16.85 > 14.86 = \chi_4^2(0.005)$[2] por lo que se considera una observación atípica multivariada.

[2] qchisq(0.005,4,lower.tail = F)

Detección de Datos Atípicos

Podemos **detectar** los **outliers multivariados** gráficamente:

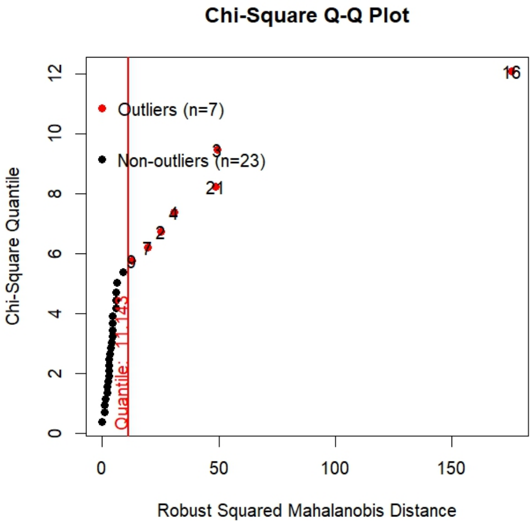
```
# -----  Outliers ----- #
require(MVN)
Out<-mvn(Datos, mvnTest = "mardia",
        multivariateOutlierMethod ="quan" ) #QQplot
Out<-mvn(Datos, mvnTest = "mardia",
       multivariateOutlierMethod ="adj") #QQplot ajust.
```

El gráfico

- considera una **Distancia** de **Mahalanobis Robusta**, esto es, considera vector de medias y matriz de covarianas robustas.
- crea un **punto de corte** a partir del cual **identifica** los **outlier**.

Pontificia Universidad JAVERIANA Bogotá

Detección de Datos Atípicos



Transformaciones

¿Como proceder **si no se cumple** el supuesto de **normalidad**?

✓ Si se ignora y se **continúa** como si se cumpliera, las conclusiones pueden ser incorrectas.

Transformaciones

¿Como proceder **si no se cumple** el supuesto de **normalidad**?

- ✓ Si se ignora y se **continúa** como si se cumpliera, las conclusiones pueden ser incorrectas.
- ✓ Hacer **transformaciones** sobre los datos originales para tratar de **aproximarlos** a la **normalidad**. Éstas transformaciones pueden ser sugeridas por los mismos datos o por consideraciones teóricas.

Transformaciones

Transformaciones

Dentro de las **consideraciones teóricas** están

Escala original	Escala transformada
x es de **conteo**	$\sqrt{x}$
$\hat{p}$ es una **proporción**	$\mathrm{logit}(\hat{p}) = \frac{1}{2}\ln\left(\frac{\hat{p}}{1-\hat{p}}\right)$
r **correlaciones**	$z(r)^3 = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$

[3] Transformación de Fisher

Existe una **familia de transformaciones** potenciales, conocida como **transformación** de **Box-Cox**:

$$x^{\lambda} = \begin{cases} \frac{x^{\lambda}-1}{\lambda} & , \lambda \neq 0 \\ \ln(\lambda) & , \lambda = 0 \end{cases}$$

la cual es **continua** en λ para $x > 0$.

Transformaciones

La **solución Box-Cox** para escoger la **transformación** λ **adecuada** es aquella que maximiza el logaritmo de la función de verosimilitud de una normal después de haberla maximizado con respecto a los parámetros. Esto es:

$$l(\lambda) = -\frac{n}{2} \ln \left[\frac{1}{n} \sum_{j=1}^{n} \left(x_j^{\lambda} - \overline{x^{\lambda}} \right)^2 + (\lambda - 1) \sum_{j=1}^{n} \ln(x_j) \right]$$

donde

$$\overline{x^{\lambda}} = \frac{1}{n} \sum_{j=1}^{n} x_j^{\lambda}$$

es la media de las observaciones transformadas.

Transformaciones

En R, obetenemos el **valor de** $\underline{\lambda}$ usando

```
# ----- Transformación BoxCox ----- #
require(car)
powerTransform(Datos$x2)
Estimated transformation parameter
  Datos$x2
-0.7756761
```

observe que para X_2, $\lambda = -0.77$.

Ejercicio: Haga la transformación y repita la verificación de normalidad.

Transformaciones

Es evidente que la **transformación de Box-Cox** antes presentada es para el caso **univariado**.

Transformaciones

Es evidente que la **transformación de Box-Cox** antes presentada es para el caso **univariado**.

¿Cómo hacer la **transformación** en el caso **multivariado**?

Transformaciones

Es evidente que la **transformación de Box-Cox** antes presentada es para el caso **univariado**.

¿Cómo hacer la **transformación** en el caso **multivariado**?

Para el caso multivariado, se **aplica el procedimiento anterior para cada una de las variables**. Asi que se tendrán

$$\lambda_1, \lambda_2, \ldots, \lambda_p$$

transformaciones y por lo tanto la $\underline{j-\text{ésima}}$ **observación transformada** es:

Transformaciones

$$x_j^{\lambda} = \begin{bmatrix} \frac{x_{j1}^{\lambda_1} - 1}{\lambda_1} \\ \\ \frac{x_{j2}^{\lambda_2} - 1}{\lambda_2} \\ \\ \vdots \\ \\ \frac{x_{jp}^{\lambda_p} - 1}{\lambda_p} \end{bmatrix}$$

Observación: Aunque normalidad univariada no implica normalidad multivariada, en aplicaciones prácticas esto puede ser suficiente.

Correlación Parcial y Múltiple

Correlación Parcial y Múltiple

Cuando las variables X_i e X_k son **normales univariadas** se puede probar la **significancia** de los **coeficientes de correlación** a través de una prueba de hipótesis:

$$\mathrm{H}_0 : \rho_{ik} = 0$$
$$\mathrm{H}_1 : \rho_{ik} \neq 0$$

Usando la **estadística de prueba**

$$t = r_{ik}\sqrt{\frac{n-2}{1-r_{ik}^2}} \overset{H_0 \text{ verd.}}{\sim} t_{(n-2)} \qquad ; \qquad r_{ik} = \frac{s_{ik}}{s_{ii}s_{kk}}$$

Correlación Parcial y Múltiple

y el $p-$valor

$$p - \text{valor} = 2\, P\left[t_{(n-2)} > |t|\right]$$

Observación: el procedimiento se debe hacer para todos los pares de variables.

Para los **datos de Rigidez**, se deben hacer **6 pruebas de hipótesis**.

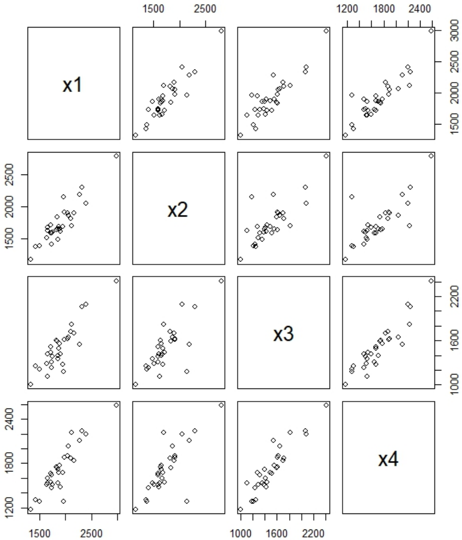
Correlación Parcial y Múltiple

Antes de hacer las pruebas de hipótesis, vamos a dar una mirada a las **correlaciones muestrales** entre éstas variables:

$$R = \begin{bmatrix} 1.00 & 0.91 & 0.89 & 0.90 \\ 0.91 & 1.00 & 0.79 & 0.79 \\ 0.89 & 0.79 & 1.00 & 0.92 \\ 0.90 & 0.79 & 0.92 & 1.00 \end{bmatrix}$$

Observe que **todas los pares** de variables tienen una **correlación** (muestral) **lineal positiva y fuerte**.

Correlación Parcial y Múltiple



Correlación Parcial y Múltiple

Verificamos si esas correlaciones son significativas o no.

```
# ----- Prueba de hipótesis para rho_ik ----- #
cor.test (Datos$x1, Datos$x2,
          alternative = "two.sided",
          método = "pearson")
```

Tenemos

	ρ_{12}	ρ_{13}	ρ_{14}	ρ_{23}	ρ_{24}	ρ_{34}
$p-$valor	1.807e-12	7.429e-11	1.688e-11	2.324e-07	2.337e-07	3.637e-13

Observe que en todos los casos

$$p - \text{valor} < 0.05 = \alpha \,,$$

por lo tanto todas las **correlaciones** son **significativas**.

Correlación Parcial y Múltiple

La aplicación de las **técnicas multivariadas** exigen que las **variables** estén **correlacionadas** entre sí de algún modo. Esto no significa que todas deban ser correlacionadas, sólo algunas. Pues, si ellas NO están correlacionadas son independientes y por lo tanto se puede hacer el análisis de forma separada.

Correlación Parcial y Múltiple

Cuando

$$\mathbf{X}_{p\times 1} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

se pueden construir **pruebas de hipótesis** para evaluar la **matriz de correlación poblacional**

$$\mathbf{P}_{p\times p} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}_{p\times p}$$

Correlación Parcial y Múltiple

esto es

$$
\begin{aligned}
\mathrm{H}_0 &: \mathbf{P}_{p\times p} = \mathrm{I}_{p\times p} \\
\mathrm{H}_1 &: \mathbf{P}_{p\times p} \neq \mathrm{I}_{p\times p}
\end{aligned}
$$

En este caso, en H_0 se está asumiendo que $\rho_{ik} = 0 \ \forall i \neq k, \ i, k = 1, 2, \ldots, p$. Es decir, que las k variables no son correlacionadas.

esto es

$$\begin{aligned} \mathrm{H}_0 &: \mathbf{P}_{p\times p} = \mathrm{I}_{p\times p} \\ \mathrm{H}_1 &: \mathbf{P}_{p\times p} \neq \mathrm{I}_{p\times p} \end{aligned}$$

En este caso, en H_0 se está asumiendo que $\rho_{ik} = 0 \; \forall i \neq k, \; i, k = 1, 2, \ldots, p$. Es decir, que las k variables no son correlacionadas.

Para probar estas hipótesis tenemos la **prueba de Bartlett**, también conocida como **prueba de Esfericidad**, cuya estadística de prueba está dada por:

$$T = -\left[n - 1 - \frac{2p+5}{6}\right] \ln\left(|\,\mathrm{R}\,|\right) \overset{H_0\,\text{verd}}{\sim} \chi_{\frac{1}{2}p(p-1)}$$

Para los datos de rigidez:

$$\begin{aligned} \mathrm{H}_0 &: \mathbf{P}_{4\times 4} = \mathrm{I}_{4\times 4} \\ \mathrm{H}_1 &: \mathbf{P}_{4\times 4} \neq \mathrm{I}_{4\times 4} \end{aligned}$$

```
# ----- Prueba de Esfericidad ----- #
require(psych)
cortest.bartlett(cor(Datos), n=30)

$chisq
[1] 147.9593
```

```
$p.value
[1] 2.089248e-29

$df
[1] 6
```

Observe que

$$p - \mathrm{valor} = 2.089248e - 29 < 0.05 = \alpha$$

entonces **hay suficiente evidencia en la muestra** para decir que

$$\mathbf{P}_{4\times 4} \neq \mathrm{I}_{4\times 4}$$

Correlación Parcial y Múltiple

Ejercicio:
Los datos `USairpollution` del paquete `HSAUR3` fueron colectados para un estudio sobre la contaminación del aire en ciudades de los Estados Unidos. Se obtuvieron las siguientes variables para 41 ciudades:

- **SO2:** contenido de SO2 del aire en microgramos por metro cúbico,
- **temp:** temperatura media anual en grados Fahrenheit,
- **manu:** número de empresas manufactureras que emplean a 20 o más trabajadores,
- **popul:** tamaño de la población (censo de 1970) en miles,

- **wind:** velocidad media anual del viento en millas por hora,
- **precip:** precipitación media anual en pulgadas;
- **predays:** número medio de días con precipitaciones al año

La **pregunta de mayor interes** era ¿cómo se relaciona el nivel de contaminación medido por la concentración de dióxido de azufre con las otras seis variables? la cual sugiere la aplicación de una **regresión lineal múltiple**, donde **SO2** sería la variable de **respuesta** y las **otras 6** variables serian las variables independientes o **explicativas**

Correlación Parcial y Múltiple

recuerde que en en regresión lineal multiple, la variable **respuesta es aleatoria y las explicativas son son fijas**. Como en la práctica, rara vez las variables son fijas, los resultados de este modelo se interpretan como condicionales a los valores observados de las variables explicativas. Por lo tanto, trataremos todas las variables como aleatorias.

a. verifique correlación univariada y multivariada, utilice gráficos y pruebas formales.
b. verifique normalidad univariada y multivariada, usando gráficos y pruebas formales.