Tarea 1

Análisis de datos Multivariados

Christian Badillo, Luis Nuñez, Luz Maria Santana, & Sealtiel Pichardo

Tabla de contenidos

1 Exploración gráfica y datos faltantes

1.1 Escalas de Medición

En la tabla 1 se muestran las variables junto a su escala de medición y el dato que se reporta en la página. Como se puede observar, a pesar de que la variable "Salud autorreportada" tiene una escala de medición de tipo ordinal, se reporta el promedio de los puntajes obtenidos por las personas, siendo más adecuado el reportaje de la mediana de los mismos.

Tabla 1: Variables y escala de medición de algunos indicadores.

Variable	Escala de medición	Dato reportado
Hogares con acceso a banda ancha	Razón	Del total de hogares, el porcentaje con acceso a banda ancha
Contaminación del aire	Razón	Nivel de contaminación del aire en PM2.5 microgramos por metro cúbico
Participación electoral	Razón	Porcentaje de personas con participación electoral del total de población adulta
Esperanza de vida al nacer	Razón	Número de años vividos (por la generación estudiada) entre el tamaño de la generación estudiada
Años promedio de escolaridad	Razón	Promedio. Suma del número acumulado de años estudiados por un conjunto de personas, entre el número de individuos que componen al estudio
Salud autorreportada	Ordinal	Promedio de los resultados obtenidos con un instrumento tipo escala Likert con unidades desde 0 (totalmente insatisfecho) hasta 10 (totalmente satisfecho)
Tasa de obesidad	Razón	Del total de población mayor de 20 años, el porcentaje de personas cuyo índice de masa corporal (IMC) fue mayor o igual a 30
Deserción escolar	Razón	Porcentaje de alumnos que abandonan la escuela de un nivel educativo, respecto a la matrícula de inicio de curso del mismo nivel

1.2 Valores Faltantes

En la tabla 2 se presentan los datos faltantes, a qué estado pertenecen y en cuál variable se encuentra el dato faltante. Hay un total de 15 datos faltantes, 15 estados tienen una observación faltante en alguna variable. En la tabla 3 se puede observar que la variable que tiene más datos faltantes es "Porcentaje de la población en situación de pobreza extrema" y los estados a los que pertenecen son a Nuevo León y Puebla.

Tabla 2: Estados con datos faltantes.

Estado	Variable	Obvs. Faltantes
Aguascalientes	Contaminación del aire	1
Campeche	Tasa de incidencia delictiva	1
Colima	Razón de mortalidad materna defunciones por cada 100 mil nacidos vivos	1
Chiapas	Confianza en la policía	1
Durango	Participación cívica y política	1
Guanajuato	Porcentaje de viviendas con techos de materiales resistentes	1
Jalisco	Salud autorreportada	1
Michoacán de Ocampo	Calidad de la red social de soporte	1
Nuevo León	Porcentaje de la población en situación de pobreza extrema	1
Oaxaca	Niveles de educación	1
Puebla	Porcentaje de la población en situación de pobreza extrema	1
Sinaloa	Tasa de informalidad laboral	1
Tamaulipas	Satisfacción con la vida	1
Tlaxcala	Población ocupada trabajando más de 48 horas	1
Zacatecas	Tasa de desocupación	1

Tabla 3: Valores faltantes en cada variable

Variable	Estado	Obvs. Faltantes
Contaminación del aire	Aguascalientes	1
Tasa de incidencia delictiva	Campeche	1
Razón de mortalidad materna defunciones por cada 100 mil nacidos vivos	Colima	1
Confianza en la policía	Chiapas	1
Participación cívica y política	Durango	1
Porcentaje de viviendas con techos de materiales resistentes	Guanajuato	1
Salud autorreportada	Jalisco	1
Calidad de la red social de soporte	Michoacán de Ocampo	1
Porcentaje de la población en situación de pobreza extrema	Nuevo León / Puebla	2
Niveles de educación	Oaxaca	1
Tasa de informalidad laboral	Sinaloa	1
Satisfacción con la vida	Tamaulipas	1

Variable	Estado	Obvs. Faltantes
Población ocupada trabajando más de 48 horas	Tamaulipas / Tlaxcala	1
Tasa de desocupación	Zacatecas	1

El porcentaje de estados que tienen valores faltantes, respecto al total, representa el 45.45%; por otra parte, respecto a las variables, los datos faltantes representan al 40% de todos los datos. El total de datos faltantes es de 1.29% respecto al total de datos.

Tabla 4: Porcentaje de valores faltantes, por renglón

# de Renglones	Cantidad de NA	%
15	1	45.45
18	0	54.55

Tabla 5: Porcentaje de valores faltantes, por columna

# de Columnas	Cantidad de NA	%
1	2	2.85
13	1	37.15
21	0	60

Tabla 6: Porcentaje de valores faltantes, total

	Cantidad	%
Datos con NA	15	1.3
Datos sin NA	1140	98.7

1.3 Tratamiento de Datos Faltantes

Dado que tenemos pocos datos faltantes, se hizo uso de la libreria `mice` de R para imputar los datos faltantes usando el método de regresión base de la función `mice()`, con semilla `seed=20` para garantizar su reproducibilidad en el código.

En cuanto a la clasificación de datos faltantes, podemos creer que existen de dos tipos en la base de datos: CAR y MNAR. Los casos de MNAR sería viable pensarlos en casos como en Oaxaca donde hace falta información con respecto a los niveles de educación, caso similar a Chiapas donde no hay registros sobre la confianza en la policía, lo cual es estraño en un lugar con muchos conflictos en temas de seguridad; otro estado con datos faltantes sospechosos es Campeche donde la tasa de incidencia delictiva no es reportada, Puebla también puede ser un valor faltante del tipo MNAR dado que no se reporta el porcentaje de la población en situación de pobreza extrema. Dada la

naturaleza de estas variables mencionadas, creemos que es plausible que su falta no sea debida al azar.

Los valores faltantes restantes consideramos que pueden ser del tipo CAR ya que no vemos razones por la cual su información se haya modificado.

1.4 Función empírica de distribución, Boxplot, Normal QQ-Plot e Histograma

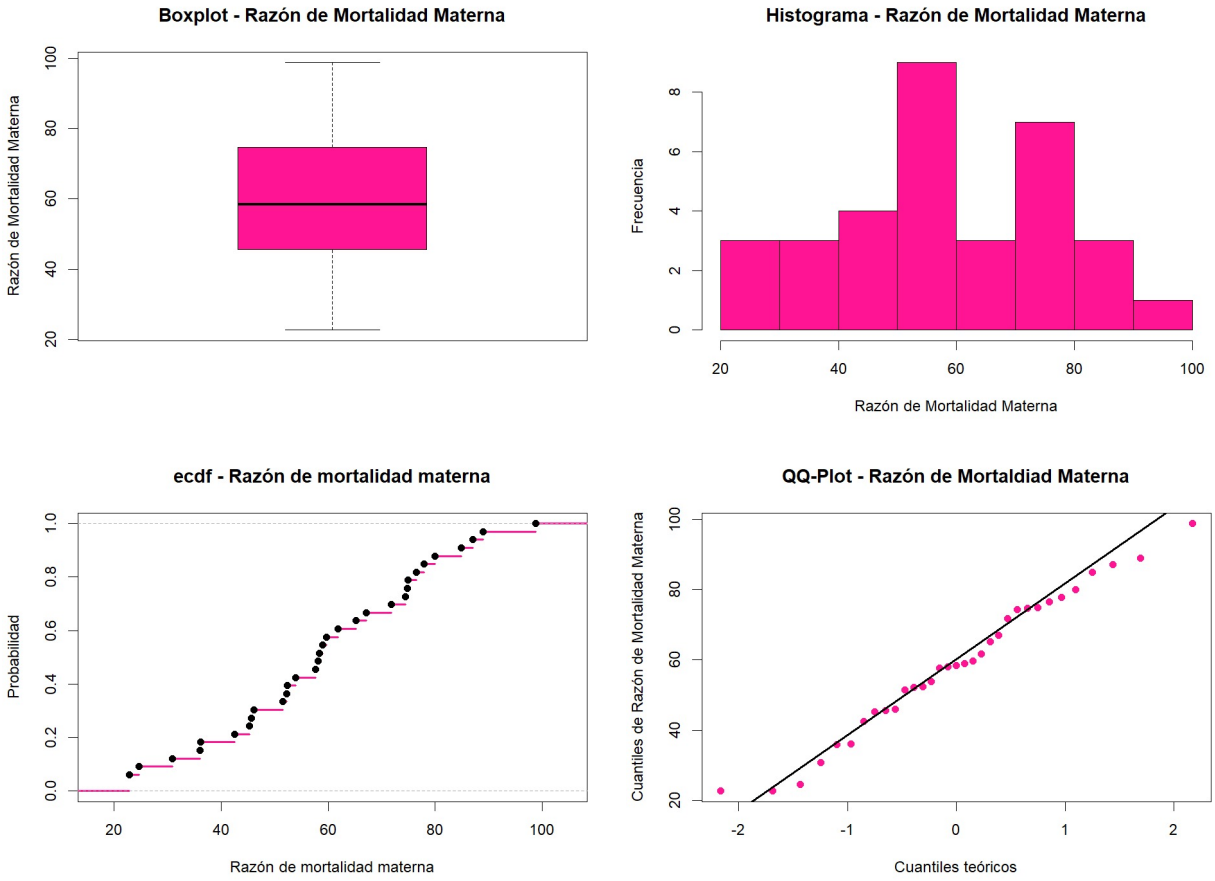


Figura 1: Diagrama de caja, histograma, función de distribución acumulada empírica y QQ-plot para la variable "razón de mortalidad materna". Para la variable Razón de mortalidad materna, nuevamente se tiene una distribución cercana a la normal ya que la QQ-plot se acerca demasiado a la línea de referencia y tiene una mayor variabilidad en los valores de los datos.

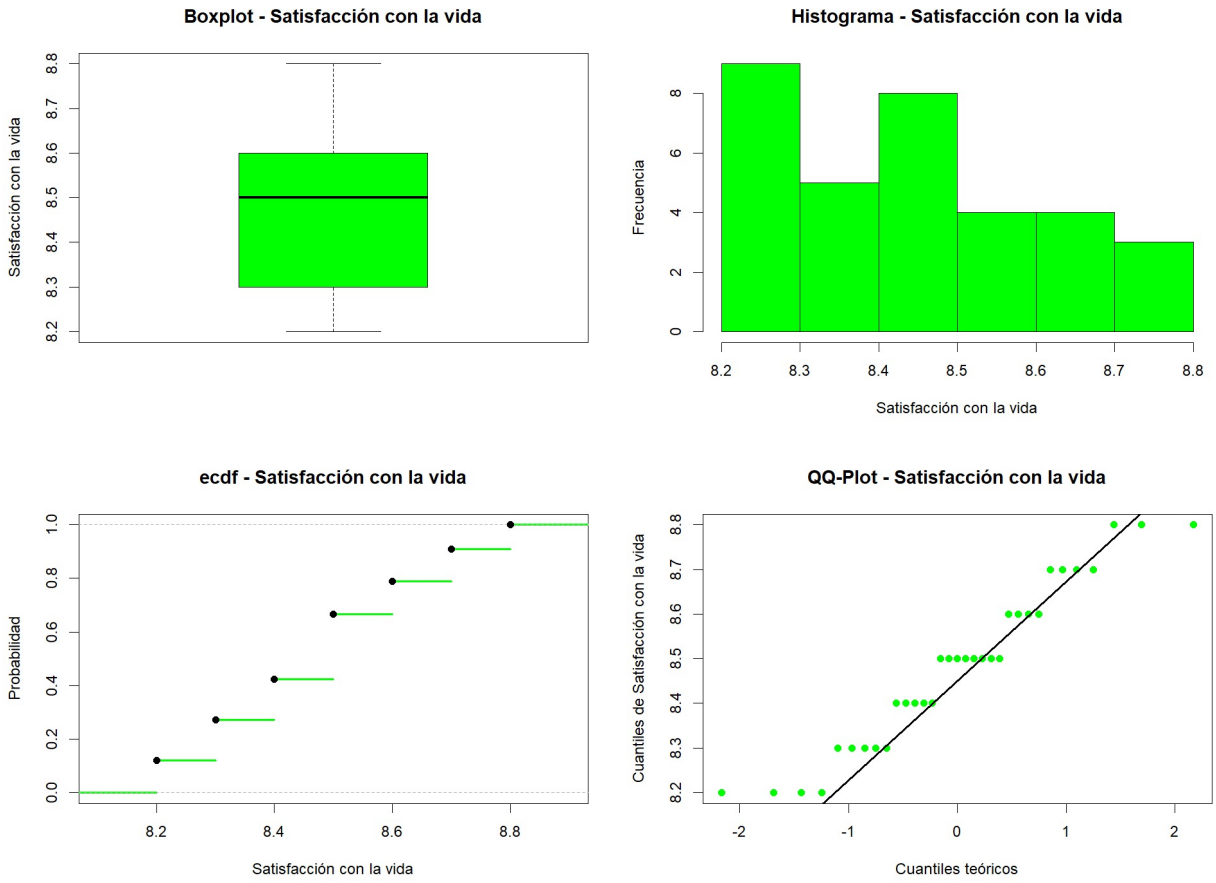


Figura 2: Diagrama de caja, histograma, función de distribución acumulada empírica y QQ-plot para la variable “satisfacción con la vida”. En el caso de la variable Satisfacción con la vida, se tiene un sesgo positivo, ya que en el histograma se presenta una cola pronunciada a la derecha, además en la boxplot, se puede observar el bigote superior más largo, confirmando el sesgo positivo.

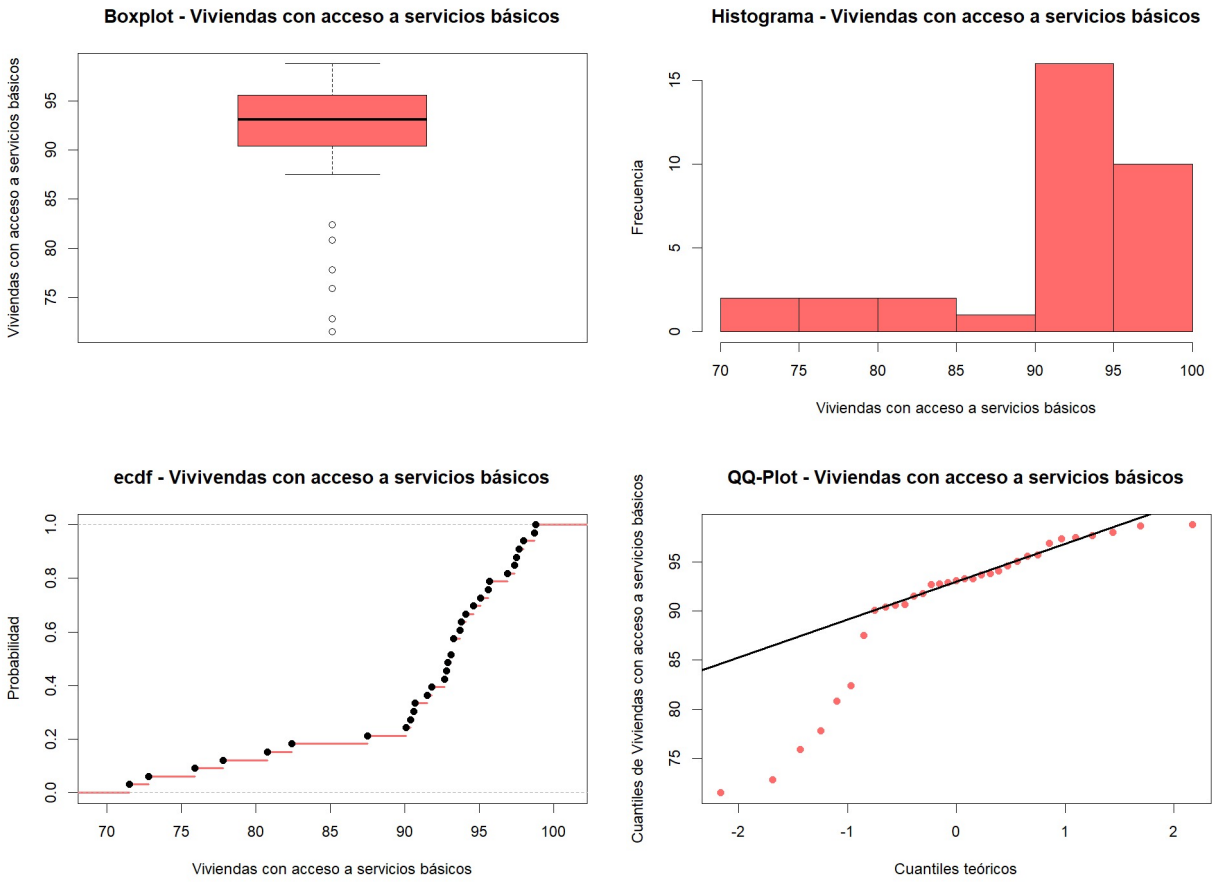


Figura 3: Diagrama de caja, histograma, función de distribución acumulada empírica y QQ-plot para la variable "viviendas con acceso a servicios básicos". En el histograma se observa una distribución con sesgo negativo ya que tiene una cola pronunciada a la izquierda. En la QQ-plot se observa que tiene valores por debajo de la normal, pero serían los referentes a los outliers vistos en la boxplot, por lo que no sería correcto definir el sesgo guiados por esa gráfica, ya que para tener un sesgo negativo, deberían ir ligeramente de la línea de referencia, como lo hacen después de los valores a la izquierda. Aunado a lo anterior, al analizar la boxplot, para cumplir con un sesgo negativo, el bigote inferior debería ser más largo, sin embargo, no se observa ese comportamiento, pero, los outliers generan ese tipo de sesgo ya que están muy por debajo del límite del bigote.

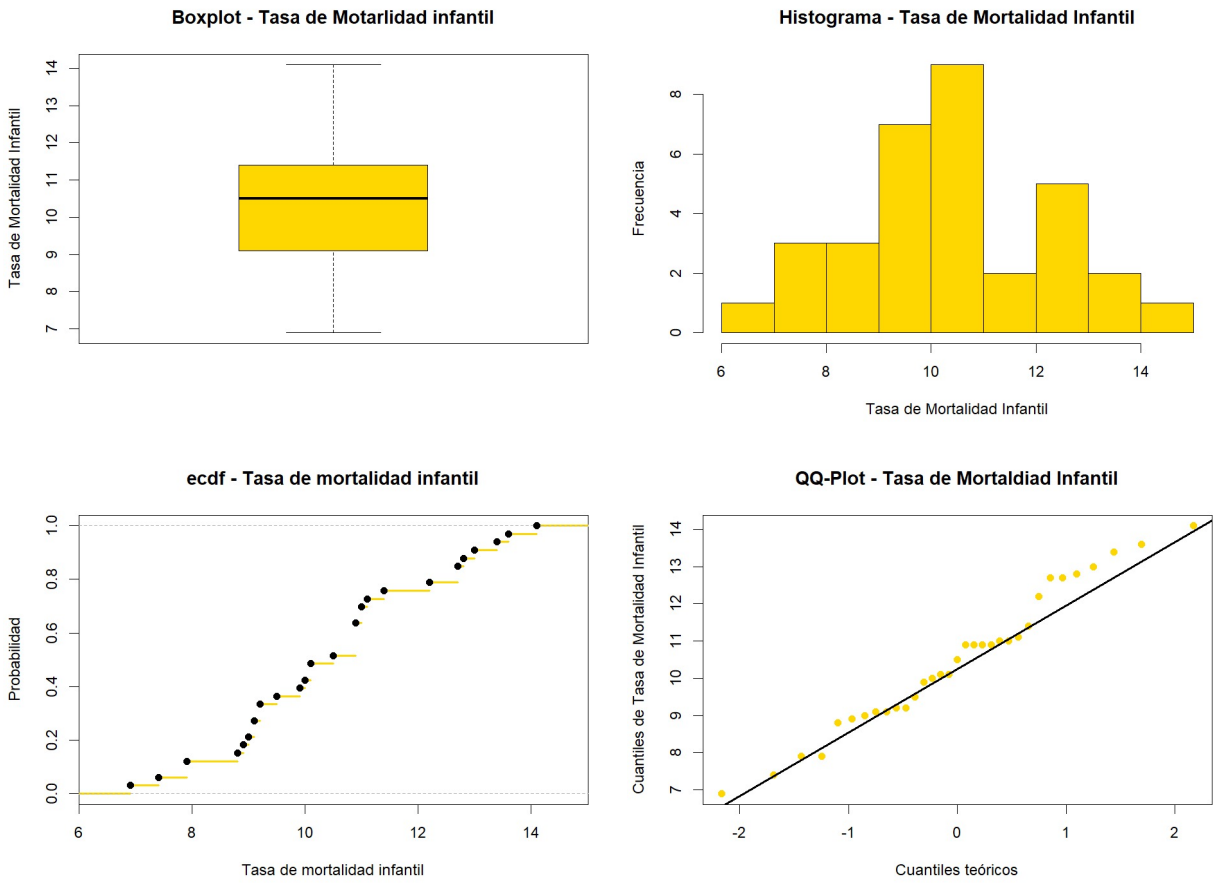


Figura 4: Diagrama de caja, histograma, función de distribución acumulada empírica y QQ-plot para la variable “tasa de mortalidad infantil”. En esta variable se presenta algo parecido al caso anterior pero con una dispersión más temprana de los datos. En este caso no se presentan outliers, sin embargo, la boxplot presenta bigotes más alargados haciendo referencia a que la variabilidad es mayor.

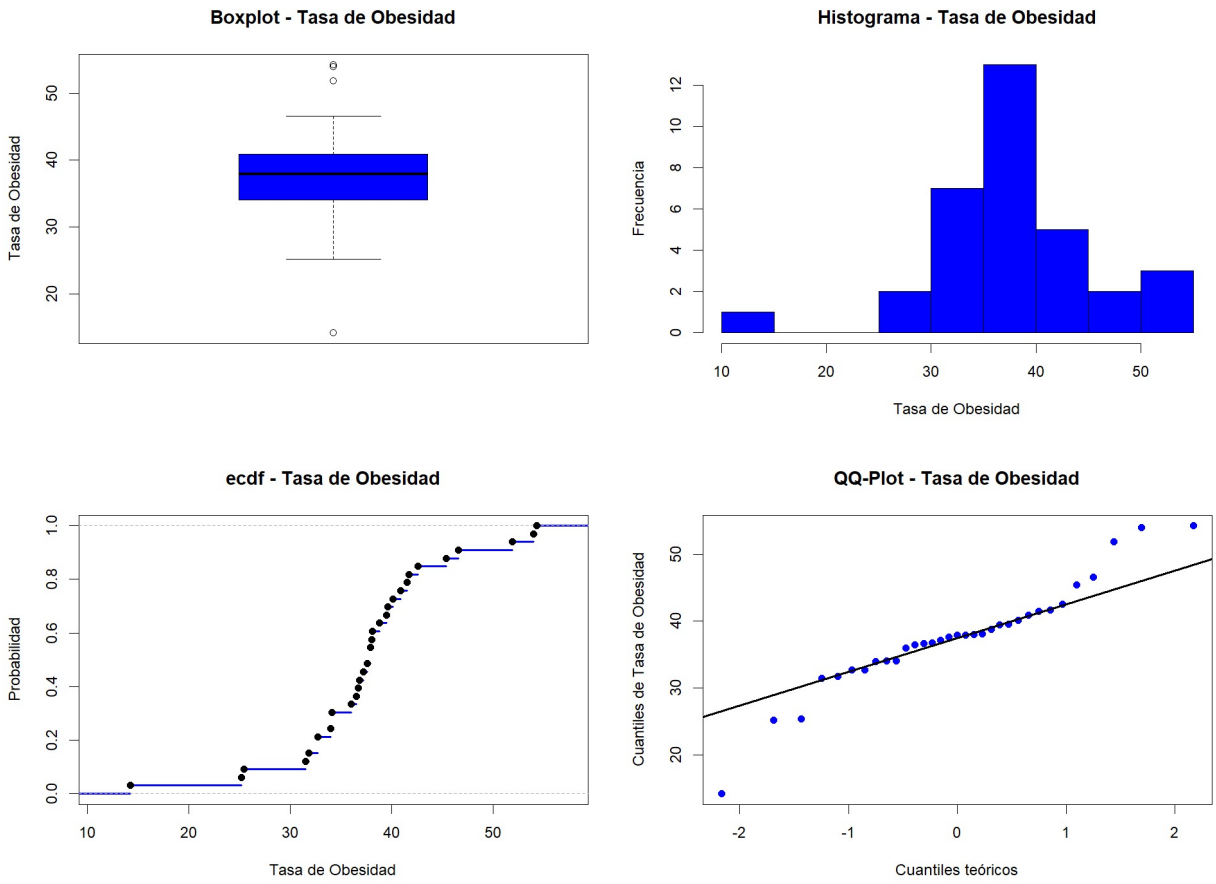


Figura 5: Diagrama de caja, histograma, función de distribución acumulada empírica y QQ-plot para la variable "Tasa de Obesidad". La variable Tasa de Obesidad presenta una distribución un poco normal ya que en la gráfica QQ-plot se puede observar como los puntos de la variable se acercan a la línea que representa la normal, sin embargo, las orillas se dispersan, además de presentar outliers (ver boxplot) que pueden aumentar dicha dispersión.

1.5 Conclusión

Cada gráfica muestra la información de los datos de diferente manera, pero ambos muestran la distribución de los datos: lo ecdf's muestran el rango de manera continua y acumulada, mientras que los boxplot's muestran la distribución de una manera más compacta. Por otro lado en ambos casos podemos visualizar la mediana: en el caso de las ecdf's es el valor correspondiente al 0.5% de probabilidad, mientras que en los boxplot se puede encontrar de una manera más sencilla al visualizar la línea que se encuentra entre el primer y tercer cuartil.

Este tipo de gráficos permiten visualizar variables con algún tipo de sesgo, recordando que una distribución sesgada negativamente es aquella con una cola pronunciada a la izquierda, por lo que en este caso la variable "Viviendas con acceso a servicios básicos" tiene un sesgo negativo lo que implica que la media es menor que la mediana en esta variable. Para determinar si una variable es sesgada positivamente, se debe observar si se tiene una cola pronunciada a la derecha, un ejemplo de ello es la variable "Satisfacción con la vida" cuya distribución observada en el histograma muestra un sesgo positivo y por tanto, indica que su media es mayor que la mediana, además de que hay valores altos que extienden la distribución a la derecha.

Y por último, usando estas visualizaciones podemos inspeccionar aquellas variables con una distribución aproximadamente normal como lo serían: la "Tasa de mortalidad infantil", la "Razón de mortalidad materna" y la "Tasa de obesidad", siendo esta última aquella con una aproximación mayor por lo descrito en su análisis.

1.6 Gráficas multivariadas.

Variables seleccionadas:

1. Vivienda: Porcentaje.de.viviendas.con.techos.de.materiales.resistentes
2. Ingresos: Gini.del.ingreso.disponible.de.los.hogares.per.capita
3. Empleo: Tasa.de.condiciones.criticas.de.ocupacion
4. Accesibilidad a servicios: Viviendas.con.acceso.a.servicios.basicos
5. Seguridad: Tasa.de.homicidio
6. Educación: Niveles.de.educacion
7. Medio ambiente: Contaminacion.del.aire
8. Compromiso cívico y gobernanza: Participacion.electoral
9. Salud: Esperanza.de.vida.al.nacer
10. Satisfacción con la vida: Satisfaccion.con.la.vida
11. Balance vida-trabajo: Satisfaccion.con.tiempo.para.ocio
12. Relaciones sociales en la comunidad: Calidad.de.la.red.social.de.soporte

Empezando con las curvas de Andrew recordemos que si las curvas se encuentran muy cerca entre si, relativamente en las dimensiones superiores también lo estarán, es decir, si los estados muestrán valores similares para cada una de las variables, entonces la proyección a estas curvas será observar curvas paralelas o bien muy parecidas.

A continuación se muestran las curvas de Andrews para las 12 variables seleccionadas, en primera instancia vemos que es muy poco distinguible el paralelismo entre entidades federales e incluso distinguir una de otra.

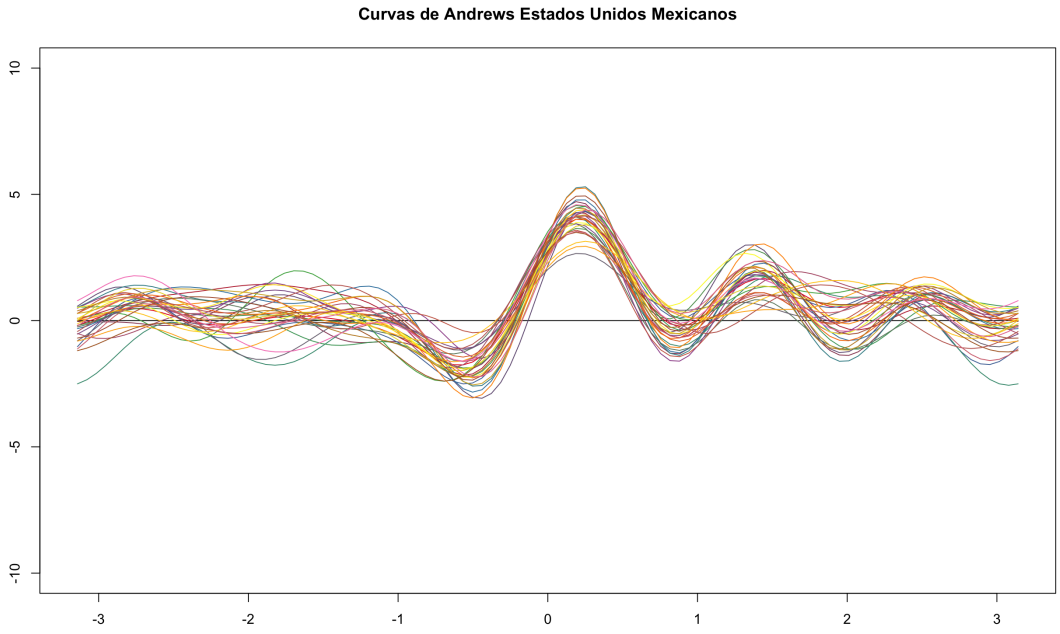


Figura 6: Curvas de Andrews. Cada Curva corresponde a un estado.

Para las caras de Chernoff sabemos que las variables son asociadas a rasgos del rostro, por decisión fijamos la sonrisa (nula) y así podremos distinguir las otras variaciones, estas son:

# Var	modified item	Var
1	“height of face”	“%VTR”
2	“width of face”	“GIDH”
3	“structure of face”	“TCCO”
4	“height of mouth”	“VASB”
5	“width of mouth”	“TH”
6	“height of eyes”	“NE”
7	“width of eyes”	“CAire”
8	“height of hair”	“PElec”
9	“width of hair”	“EspVN”
10	“style of hair”	“SV”
11	“height of nose”	“STO”
12	“width of nose”	“CalRed”
1	“width of ear”	“%VTR”
2	“height of ear”	“GIDH”

Notamos que es un poco más simple ver la variación y relación, por ejemplo si ponemos atención en la variable 10: Satisfacción con la vida “SV” que corresponde al estilo del cabello, veremos que estados como: “COL”, “COA”, “NL”, “NAY”, “SIN”; tienen el mismo estilo de cabello por lo que ese índice es muy similar entre ellos (3.7 aprox).

También podemos ver un claro contraste entre “Gto” y “NL” donde las caracteristicas del rostro son contrarias, mientras uno tiene orejas anchas (“%VTR”) y nariz ancha (“CalRed”), el otro no; eso implica que los valores de las variables en uno son de las más altas mientras que en “Gto” no es así.

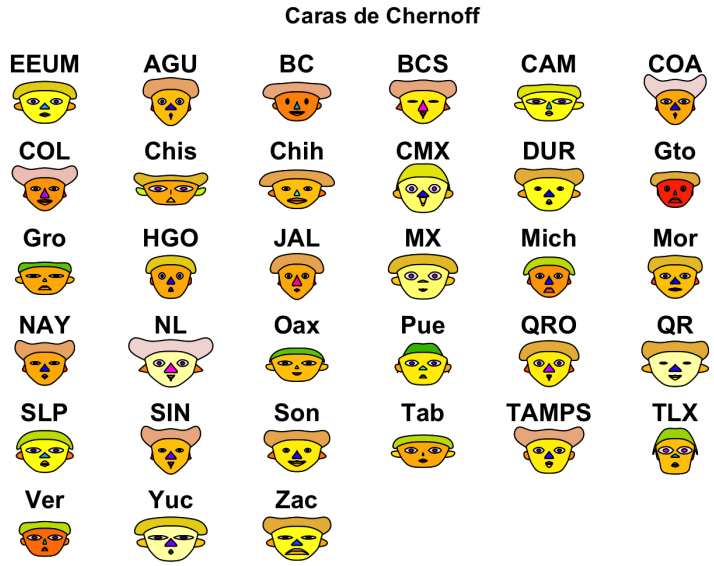


Figura 7: Caritas de Chernoff para los distintos estados.

Finalmente en el gráfico de paralelas, recordamos que tendremos 12 rectas verticales, además este grafico nos muestra claramente como cada variable tiene una unidad distinta. Los valores en cada variables se pueden apreciar de recta en recta, la correlación entre variables se vería como rectas paralelas entre los ejes, relaciones inversas como rectas cruzandose.

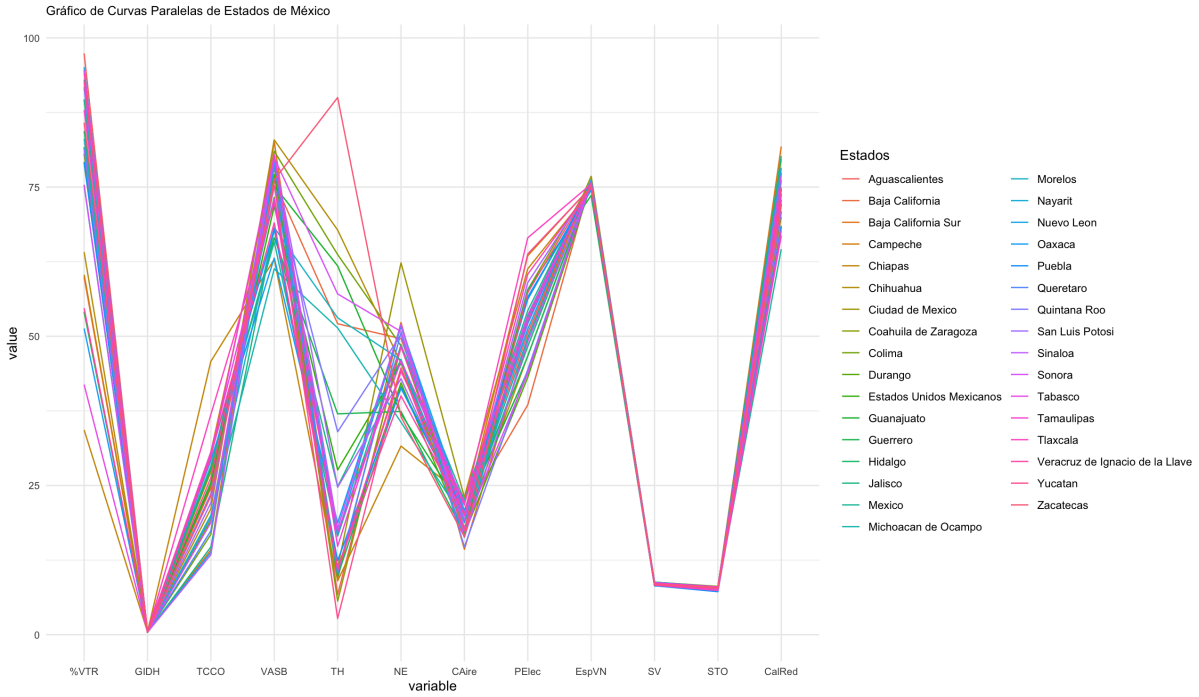


Figura 8: Gráfico de paralelas para los distintos estados

Estas gráficas nos permieten comparar las entidades federales, de mejor manera, como hemos visto cada una de ellas nos da una visión diferente de los datos,las que se parecen menos son como vimos en el grafico de rostros son “Gto” y “NL”. Una forma de agrupar las entidades federales sería por caracteisticas similares en los rostros de Chernoff

2 Matrices de Correlación y Covarianza.

Se muestra el código para obtener las matrices de covarianza y correlación.

2.1 Correlación

En este bloque código cargamos las paqueterias al igual que los datos y creamos un subconjunto de los datos con las variables solicitadas.

```
library(ggplot2)
library(tidyverse)

data.raw <- read.csv("imputacion_completa.csv", header = TRUE)
columns <- names(data.raw)

subset.columns <- c(
          "Niveles.de.educacion", "Desercion.escolar",
          "Anios.promedio.de.escolaridad",
          "Satisfaccion.con.tiempo.para.ocio",
          "Poblacion.ocupada.trabajando.mas.de.48.horas",
          "Gini.del.ingreso.disponible.de.los.hogares.per.capita")

data <- data.raw %>%
    select(subset.columns)
```

```
# Podemos obtener la matrix de correlación como:
mat.corr.data <- as.matrix(cor(data, metho="pearson"), 6, 6)
```

Tabla 8: Matriz de Correlación por medio de la función base de R

	n.edu	des.esc	anios.esc.pro	satis.t.ocio	trabajo>48h	gini.hog.pcap
n.edu	1.000	0.136	0.804	0.037	-0.124	0.010
des.esc	0.136	1.000	0.131	0.030	0.127	0.155
anios.esc.pro	0.804	0.131	1.000	0.118	-0.092	-0.126
satis.t.ocio	0.037	0.030	0.118	1.000	-0.323	-0.057
trabajo>48h	-0.124	0.127	-0.092	-0.323	1.000	-0.006
gini.hog.pcap	0.010	0.155	-0.126	-0.057	-0.006	1.000

Podemos usar operaciones con matrices para obtener el mismo resultado.

```
# O podemos obtener la matriz centradora.
n <- nrow(data)
J <- diag(n) - 1/n * matrix(rep(1, n), n, n)
```

```r
# Ahora premultiplicamos esto por los datos.
centered.data.mat <- matrix(J, n, n) %*% as.matrix(data, 33, 6)

# La varianza de cada variable se obtiene como:
cov.data <- t(as.matrix(centered.data.mat, 33, 6)) %*%
  as.matrix(centered.data.mat, 33, 6)
cov.data <- as.matrix(cov.data, 6, 6) * 1/n

## Se puede verificar que da la misma entrada que:
# cov(data) * (n-1)/n

# Ahora obtenemos la varianza de cada variable
var.data <- diag(cov.data)

# Ahora calculamos su desviación estandar
sd.data <- sqrt(var.data)
sd.data <- as.matrix(sd.data, 1, 6)

# Creamos una matrix cuyas diagonales sea la sd de cada variable
D <- matrix(rep(0, 36), 6, 6)
for (i in 1:6) {
  for (j in 1:6) {
    if (i == j) {
      D[i, j] <- sd.data[i]
    }
  }
}

# Encontramos su inversa
D.inv <- solve(D)

# Ahora obtnemos R
R <- t(D.inv)%*%cov.data%*%D.inv

# Veamos si R es igual a mat.corr.data con una tolerancia de 1.5e-8
all.equal(R, mat.corr.data, check.attributes =F)
```

```
[1] TRUE
```

Como se puede observar al comparar ambas matrices, son iguales.

Tabla 9: Matriz de Correlación por medio de operaciones matriciales.

	n.edu	des.esc	anios.esc.pro	satis.t.ocio	trabajo>48h	gini.hog.pcap
n.edu	1.000	0.136	0.804	0.037	-0.124	0.010
des.esc	0.136	1.000	0.131	0.030	0.127	0.155
anios.esc.pro	0.804	0.131	1.000	0.118	-0.092	-0.126
satis.t.ocio	0.037	0.030	0.118	1.000	-0.323	-0.057
trabajo>48h	-0.124	0.127	-0.092	-0.323	1.000	-0.006
gini.hog.pcap	0.010	0.155	-0.126	-0.057	-0.006	1.000

Las variables con mayor relación lineal son el nivel educativo y los años promedio de estudio, lo cual es lógico dado que entre mayor sea el nivel de estudios más años se habrán invertido en actividades académicas. El coeficiente gini del ingreso disponible de los hogares *per capita* no parece estar asociado con las variables de este subconjunto de los datos, caso similar con la deserción escolar. La satisfacción con el tiempo de ocio tiene una correlación negativa algo moderada con la población que trabaja más de 48 horas a la semana, lo cual oarece razonable dado que a más tiempo en trabajo, menos cantidad de tiempo disponible habrá para actividades de ocio.

2.2 Covarianza

Con el código anterior hemos obtenido la matriz de covarianza.

Tabla 10: Matriz de Covarianza

	n.edu	des.esc	anios.esc.pro	satis.t.ocio	trabajo>48h	gini.hog.pcap
n.edu	35.688	2.267	3.652	0.041	-3.002	0.001
des.esc	2.267	7.729	0.277	0.016	1.429	0.011
anios.esc.pro	3.652	0.277	0.578	0.017	-0.283	-0.002
satis.t.ocio	0.041	0.016	0.017	0.036	-0.247	0.000
trabajo>48h	-3.002	1.429	-0.283	-0.247	16.377	-0.001
gini.hog.pcap	0.001	0.011	-0.002	0.000	-0.001	0.001

Podemos ver que variables como el coeficiente gini, la satisfación por el tiempo de ocio y los años promedio de escolaridad tienen una varianza pequeña a comparación de las otras variables. Lo cual puede ser causa de su escala de medición o que realmente exista poca variabilidad en los datos.

2.3 Vector de Medias.

Para estimar el vector de medias, podemos usar la función `apply()` de R para mejor eficiencia.

```r
# Vector de medias, esto se puede hacer con
mean.vector <- apply(data, 2, mean)
mean.vector
```

```
                                    Niveles.de.educacion
                                               45.6878788
                                       Desercion.escolar
                                               11.0878788
                             Anios.promedio.de.escolaridad
                                                9.7090909
                           Satisfaccion.con.tiempo.para.ocio
                                                7.7787879
                Poblacion.ocupada.trabajando.mas.de.48.horas
                                               26.4090909
Gini.del.ingreso.disponible.de.los.hogares.per.capita
                                                0.4300606
```

2.4 Coeficiente de Variación.

Estimamos el coeficiente de variación CV para cada variable.

```r
# Usamos el vector de medias y el de desviación estandar para estimar el C
sd.variables <- apply(data, 2, sd)
sd.variables
```

```
                                    Niveles.de.educacion
                                              6.06659694
                                       Desercion.escolar
                                              2.82320270
                             Anios.promedio.de.escolaridad
                                              0.77191468
                           Satisfaccion.con.tiempo.para.ocio
                                              0.19163373
                Poblacion.ocupada.trabajando.mas.de.48.horas
                                              4.10954405
Gini.del.ingreso.disponible.de.los.hogares.per.capita
                                              0.02605995
```

```r
c.v <- sd.variables/mean.vector
c.v
```

```
                                    Niveles.de.educacion
                                              0.13278351
                                       Desercion.escolar
                                              0.25462063
```

```
                          Anios.promedio.de.escolaridad
                                             0.07950432
                      Satisfaccion.con.tiempo.para.ocio
                                             0.02463542
          Poblacion.ocupada.trabajando.mas.de.48.horas
                                             0.15561096
Gini.del.ingreso.disponible.de.los.hogares.per.capita
                                             0.06059600
```

Podemos observar que la variable con más variabilidad es la población que trabaja más de 48 horas a la semana, lo cual indicaría que existe un amplio rango de horas de trabajo en la población, lo que puede ser por necesidad económica o por falta de políticas que controlen el tiempo de trabajo, entre otros posibles factores. El nivel de educación también varia mucho así como la deserción escolar. Todas estas variables parecen reflejar bastante bien algunos de los problemas más importantes que enfrenta el país y que son áreas de oportunidad para el desarrollo de políticas públicas.

3 Datos Faltantes

Tratamiento de datos faltantes	¿En qué consiste?	¿En qué casos se recomienda usarlo? (datos perdidos tipo MCAR MAR//MNAR)
Omisión Total	Consiste en eliminar todos los casos con uno o más datos faltantes.	Se recomienda para casos en los que el mecanismo de pérdida es completamente al azar (MCAR) y cuando haya pocos datos faltantes (Pigott, 2001), ya que produce estimaciones insesgadas de las medias y varianzas (Buuren, 2018). En los demás mecanismos no se recomienda, debido a que puede producir estimaciones sesgadas de las medias, coeficientes de regresión y/o de correlación (Buuren, 2018).
Omisión Parcial	Consiste en hacer análisis, por variable, solo de los casos que están completos, omitiendo los faltantes. También llamada eliminación por pares.	Bajo MCAR produce estimaciones consistentes de la media y se recomienda solo si los datos se distribuyen como una normal multivariada o si hay bajas correlaciones entre las variables (Buuren, 2018). En el caso de matrices de covarianzas, puede producir correlaciones que estén fuera de rango, debido a que no haya el mismo número de variables (Pigott, 2001). Puede producir sesgos si el mecanismo de pérdida no es MCAR (Buuren, 2018).
Imputación con la Media	Se trata de reemplazar el valor faltante con la media de los demás valores de esa variable. O usar la moda en caso de que el tipo de variable sea categórica.	Cuando el mecanismo de pérdida no es MCAR, subestima la varianza (la reduce artificialmente) (Buuren, 2018), altera la relación entre posibles variables y sesga casi cualquier estimación, incluida la media (Kleinke, 2020).
Imputación Hot Deck	Consiste en reemplazar un dato faltante con uno que se tiene de una base de datos alterna y que tenga puntajes similares en las demás variables.	Puede producir estimaciones de la media consistentes solo cuando el mecanismo de pérdida es MCAR y en casos limitados para MAR. Pero solo cuando la información del "donador" sea un buen reflejo de las variables observadas y de aquellas observadas que tengan buena correlación con las no observadas, además de que la muestra deba ser lo suficientemente grande para encontrar un buen donador y que los datos faltantes no sean muchos (Andridge & Little, 2010).
Imputación usando Regresión	Consiste en hacer uso de información disponible en otras variables con el fin de producir imputaciones que sean más aproximadas a los valores faltantes. De construir un modelo basado en la información disponible y rellenar los datos faltantes con las predicciones de dicho modelo.	Bajo MCAR produce estimaciones insesgadas de la media. Puede ser útil para MAR solo si los factores para la pérdida de información son atrapados por el modelo de regresión usado (Buuren, 2018). Sin embargo, aumenta de manera artificial las correlaciones y disminuye la varianza al subestimar la variabilidad que pueda ser inherente a los datos. Lo que puede aumentar la probabilidad de observar relaciones espurias (Buuren, 2018), aunque esto se podría controlar agregando errores aleatorios al valor imputado (Newman, 2014).
Imputación usando Algoritmo EM	Se trata de un proceso iterativo de dos pasos, uno de estimación de la esperanza (o algún parámetro) con los datos dados para poder aproximar el valor de los datos que se imputarán para que se cumpla la estimación. Y otro de maximización para reestimar el parámetro ya con los datos imputados. Después se repite el proceso y se reajustan los datos imputados hasta que haya coincidencia con las nuevas estimaciones, asume que los datos se distribuyen como una normal multivariada.	Puede generar estimaciones insesgadas bajo el mecanismo de pérdida MCAR y MAR; y mejora cuando aumenta la cantidad de variables usadas en el modelo (Newman, 2014).
Imputación múltiple	Consiste en crear múltiples versiones de los datos, cada uno con una estimación plausible de los datos no observados (por medio de algún método de imputación). Se calcula el parámetro de interés en los m conjuntos de datos creados, se promedian y se utiliza como el valor de la estimación.	Genera estimaciones insesgadas bajo el mecanismo MCAR y MAR (Newman, 2014).

4 Algebra de Matrices

Parte I

Consideren la siguiente matriz. Para cada uno de los incisos,muestren las operaciones que realizaron para justificar su respuesta.

1. Obtengan los valores propios (calculando el polinomio característico) y los vectores asociados a cada uno de esos valores propios.

$$\mathbf{A} = \begin{pmatrix} 6 & 4 & 0 \\ 4 & 6 & 0 \\ 0 & 0 & 6 \end{pmatrix}$$

Solución:

El polinomio característico de $\mathbf{A}$ estará determinado por:

$$det(\mathbf{A} - \lambda\mathbf{I}) = 0$$

$$(\mathbf{A} - \lambda\mathbf{I}) = \begin{pmatrix} 6 & 4 & 0 \\ 4 & 6 & 0 \\ 0 & 0 & 6 \end{pmatrix} - \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix} = \begin{pmatrix} 6-\lambda & 4 & 0 \\ 4 & 6-\lambda & 0 \\ 0 & 0 & 6-\lambda \end{pmatrix}$$

$$\Rightarrow det(\mathbf{A} - \lambda\mathbf{I}) = 0$$

$$\Leftrightarrow (6-\lambda)[(6-\lambda)(6-\lambda) - 0] - 4[4(6-\lambda) - 0] + 0[\dots] = 0$$

$$\Leftrightarrow (6-\lambda)^3 - 16(6-\lambda) = 0$$

$$\Leftrightarrow (6-\lambda)\left[(6-\lambda)^2 - 16\right] = 0$$

$$\Rightarrow \lambda_1 = 6$$

o

$$(6-\lambda)^2 - 16 = 0$$

$$6 - \lambda = \pm 4$$

$$\Rightarrow \lambda_2 = 2$$

$$\Rightarrow \lambda_3 = 10$$

Por lo tanto los valores propios de la matriz $\mathbf{A}$ son:

$$\begin{cases} \lambda_1 = 6 \\ \lambda_2 = 2 \\ \lambda_3 = 10 \end{cases}$$

Ahora supongamos que $\bar{v} \neq \bar{0}$ entonces $(\mathbf{A} - \lambda\mathbf{I}) = \bar{0}$

$$\therefore (\mathbf{A} - \lambda\mathbf{I})\bar{v} = \bar{0}$$
$$= \begin{pmatrix} 6-\lambda & 4 & 0 \\ 4 & 6-\lambda & 0 \\ 0 & 0 & 6-\lambda \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \bar{0}$$

Para $\lambda_1 = 6$

$$\begin{pmatrix} 6-6 & 4 & 0 \\ 4 & 6-6 & 0 \\ 0 & 0 & 6-6 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \bar{0}$$
$$\Rightarrow \begin{pmatrix} 4y \\ 4x \end{pmatrix} = \bar{0}$$
$$\therefore \bar{v} = (0, 0, 0)$$

Para $\lambda_2 = 2$

$$\begin{pmatrix} 6-2 & 4 & 0 \\ 4 & 6-2 & 0 \\ 0 & 0 & 6-4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \bar{0}$$
$$\Rightarrow \begin{pmatrix} 4x+4y \\ 4x+4y \\ 4z \end{pmatrix} = \bar{0}$$
$$\Leftrightarrow \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -y \\ -x \\ 0 \end{pmatrix}$$
$$\therefore \bar{v} = (-y, y, 0)$$

Para $\lambda_3 = 10$

$$\begin{pmatrix} 6-10 & 4 & 0 \\ 4 & 6-10 & 0 \\ 0 & 0 & 6-10 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \bar{0}$$

$$\Rightarrow \begin{pmatrix} -4x+4y \\ 4x-4y \\ -4z \end{pmatrix} = \bar{0}$$

$$\Leftrightarrow \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} y \\ x \\ 0 \end{pmatrix}$$

$$\therefore \bar{v} = (y, y, 0)$$

Por lo tanto, los vectores propios asociados a los valores propios de la matriz $\mathbf{A}$ son:

$$\begin{cases} \bar{v}_1 = (0,0,0) \\ \bar{v}_2 = (-y,y,0) \\ \bar{v}_3 = (y,y,0) \end{cases}$$

2. ¿Es $\mathbf{A}$ una matriz idempotente?

Una manera fácil de saberlo es si cumple:

$$\mathbf{A}^2 = \mathbf{A}$$

Veamos:

$$\begin{aligned} \mathbf{A}^2 &= \begin{pmatrix} 6 & 4 & 0 \\ 4 & 6 & 0 \\ 0 & 0 & 6 \end{pmatrix} \cdot \begin{pmatrix} 6 & 4 & 0 \\ 4 & 6 & 0 \\ 0 & 0 & 6 \end{pmatrix} \\ &= \begin{pmatrix} 36+16 & 24+24 & 0 \\ 24+24 & 16+36 & 0 \\ 0 & 0 & 36 \end{pmatrix} \\ &\neq \mathbf{A} \end{aligned}$$

No lo es !!

3. ¿Es $\mathbf{A}$ una matriz no singular?

Eso lo sabremos si $\mathbf{A}$ es invertible. Para ello el determinante debe ser distinto de cero.

$$\begin{aligned} det(\mathbf{A}) &= 6[36-0] - 4[24-0] + 0[0] \\ &= 216 - 96 \\ &= 120 \neq 0 \end{aligned}$$

Dado que el determinante de $\mathbf{A}$ es distinto de 0, entonces es no singular.

4. ¿Cuánto vale la traza de la matriz $\mathbf{A}$?

$$\begin{aligned} tr(\mathbf{A}) &= a_{11} + a_{22} + a_{33} \\ &= 6 + 6 + 6 \\ &= 18 \end{aligned}$$

5. ¿Cuánto vale el rango de la matriz $\mathbf{A}$?

Nos podemos apoyar del inciso 3 donde calculamos el determinante, que resultó distinto de cero, por lo tanto la matriz es de rango máximo,es decir **rango 3**.

6. ¿Es $\mathbf{A}$ una matriz simétrica?

Deberá cumplir que:

$$\mathbf{A}^T = \mathbf{A}$$

Veamos:

$$\begin{aligned} \mathbf{A}^T &= \begin{pmatrix} 6 & 4 & 0 \\ 4 & 6 & 0 \\ 0 & 0 & 6 \end{pmatrix}^T \\ &= \begin{pmatrix} 6 & 4 & 0 \\ 4 & 6 & 0 \\ 0 & 0 & 6 \end{pmatrix} \\ &= \mathbf{A} \end{aligned}$$

Si lo es!!

7. Determinen si esta matriz es una matriz definida positiva, semidefinida positiva o si no lo es.

Es definida positiva si cumple alguna de las siguientes condiciones:

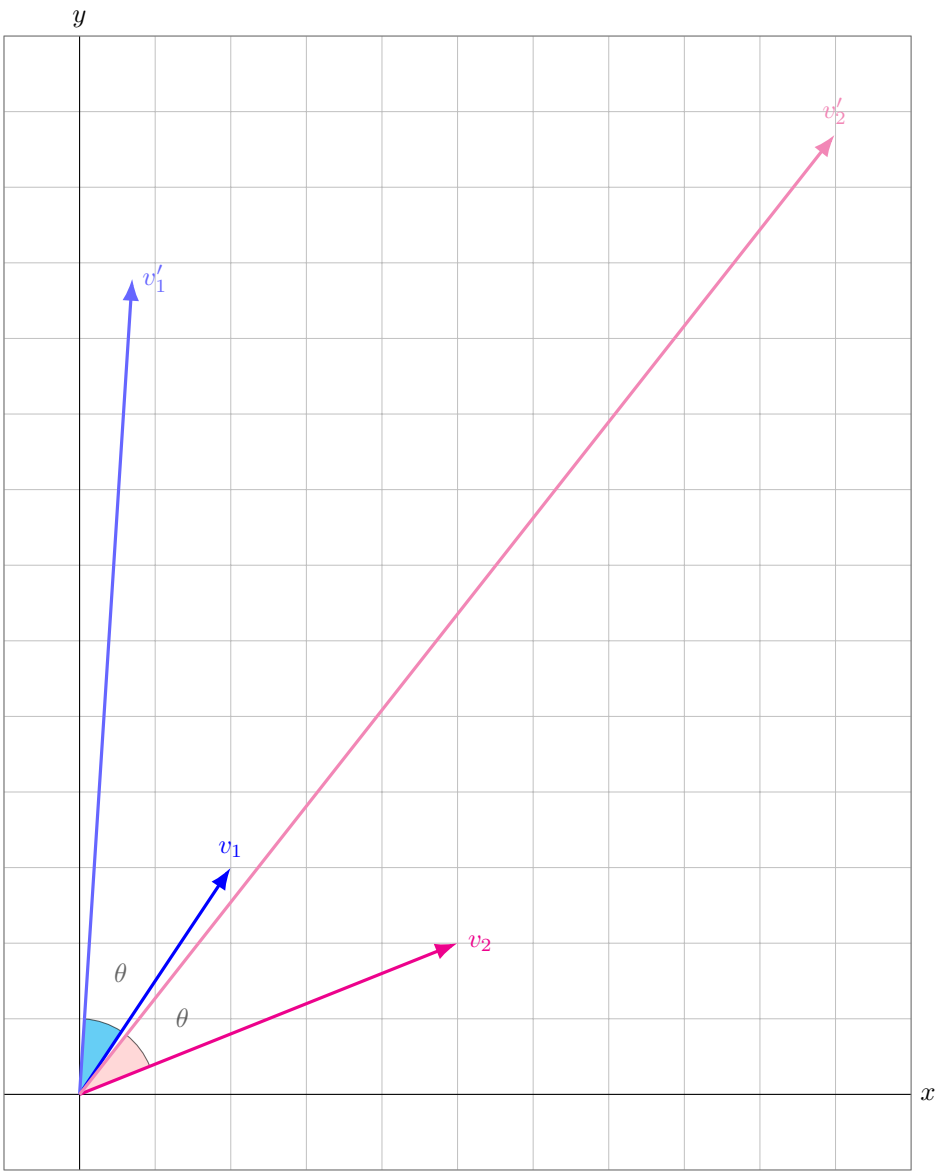
1. Tiene eigenvalores positivos
2. $\bar{x}^t \mathbf{A} \bar{x} > 0 \ \ \forall \bar{x}$
3. Los determinates de las submatrices principales de $\mathbf{A}$ son positivos

Como vimos en el inciso 1, todos los eigenvalores son positivos por lo tanto si es **DEFINIDA POSITIVA**.

Parte II

Rotar 30^o los vectores $\mathbf{v_1} = (2, 3)$ y $\mathbf{v_2} = (5, 2)$ y triplicar su tamaño.

1. Mostrar los vectores de manera gráfica antes y después de hacer la transformación.



2. Obtener las dos matrices necesarias para esta transformación.

Como hablamos de una rotación en el plano euclídiano, usaremos la matriz:

$$R(\theta) = \begin{pmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{pmatrix}$$

Por otro lado el tamaño solo es multiplicarlo por un escalar c, en este caso $c = 3$, así la operación

que debemos hacer para cada vector a transformar es:

$$\bar{\mathbf{v}'} = c\mathbf{R}\bar{\mathbf{v}}$$

- Para el vector $\mathbf{v_1} = (2, 3)$

$$\begin{aligned}
\mathbf{v_1'} &= 3\begin{pmatrix} cos(30) & -sin(30) \\ sin(30) & cos(30) \end{pmatrix}\begin{pmatrix} 2 \\ 3 \end{pmatrix} \\
&= \begin{pmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{pmatrix}\begin{pmatrix} 6 \\ 9 \end{pmatrix} \\
&= \begin{pmatrix} \frac{6\sqrt{3}-9}{2} \\ \frac{6+9\sqrt{3}}{2} \end{pmatrix} \\
&\approx \begin{pmatrix} 0.7 \\ 10.8 \end{pmatrix}
\end{aligned}$$

- Para el vector $\mathbf{v_2} = (5, 2)$

$$\begin{aligned}
\mathbf{v_2'} &= 3 \cdot \begin{pmatrix} cos(30) & -sin(30) \\ sin(30) & cos(30) \end{pmatrix}\begin{pmatrix} 5 \\ 2 \end{pmatrix} \\
&= \begin{pmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{pmatrix}\begin{pmatrix} 15 \\ 6 \end{pmatrix} \\
&= \begin{pmatrix} \frac{15\sqrt{3}-6}{2} \\ \frac{15+6\sqrt{3}}{2} \end{pmatrix} \\
&\approx \begin{pmatrix} 10 \\ 12.7 \end{pmatrix}
\end{aligned}$$

3. Encuentra el producto punto de $\mathbf{v_1}$ y $\mathbf{v_2}$.

$$\begin{aligned}
\mathbf{v_1} \cdot \mathbf{v_2} &= (2, 3) \cdot (5, 2) \\
&= 2(5) \ + \ 3(2) \\
&= 16
\end{aligned}$$

4. Encuentra la magnitud de $\mathbf{v_1}$ y $\mathbf{v_1}$ transformado.

- Para $\mathbf{v_1}$

$$\begin{aligned}
\|\mathbf{v_1}\| &= \sqrt{x^2 + y^2} \\
&= \sqrt{2^2 + 3^2} \\
&= \sqrt{4 + 9} \\
&= \sqrt{13}
\end{aligned}$$

- Para $\mathbf{v_1'}$

$$\begin{aligned}
\|\mathbf{v_1'}\| &= \sqrt{\left(\frac{6\sqrt{3}\ -\ 9}{2}\right)^2 + \left(\frac{6\ +\ 9\sqrt{3}}{2}\right)^2} \\
&= \sqrt{\left(\frac{3}{2}\right)^2 \left[\left(2\sqrt{3}\ -\ 3\right)^2 + \left(2\ +\ 3\sqrt{3}\right)^2\right]} \\
&= \left(\frac{3}{2}\right) \sqrt{4(3)\ -\ 12\sqrt{3}\ +\ 9\ +\ 4\ +12\sqrt{3}\ +\ 9(3)} \\
&= \frac{3}{2}\sqrt{52} \\
&= \frac{3}{2}\sqrt{4(13)} \\
&= 3\sqrt{13}
\end{aligned}$$

5 Referencias

- Andridge, R., & Little, R. (2010). A review of hot deck imputation for survey non response. *International statistical review, 78(1),* 40-64. https://doi.org/10.1111/j.1751-5823.2010.00103.x
- Buuren, S. (2018). *Flexible imputation of missing data*. CRC Press, Boca Raton, Florida.
- Kleinke, K. (2020) *Applied multiple imputation: advantages, pitfalls, new developments and applications in R*. Cham: Springer. https://doi.org/10.1007/978-3-030-38164-6
- Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational research methods, 17(4),* 372-411. https://doi.org/10.1177/1094428114548590
- Pigott, T. D. (2001). A review of methods for missing data. *Educational research and evaluation, 7(4),* 353-383. https://doi.org/10.1076/edre.7.4.353.8937