

ANALISIS DE DATOS LONGITUDINALES Y MEDIDAS REPETIDAS

TAREA_1

UNIVERSIDAD NACIONAL DE COLOMBIA

ANALISIS DE DATOS LONGITUDINALES Y MEDIDAS REPETIDAS

TAREA 1

ANDRES HERNANDDO CERQUERA MEJIA

C.c 1075235138

Estudiante maestría estadística

Medellín-Colombia

2021

Contenido

OBJETIVOS	3
INTRODUCCION	4
DESARROLLO	5
ANALISIS DESCRIPTIVO	5
MANOVA	11
ANOVA.....	12
MODELO LMM.....	13
PUNTO_2.....	15
ANALISIS DESCRIPTIVO	16
Modelo LMM.....	23
CONCLUSIONES	25
BIBLIOGRAFIA.....	25

OBJETIVOS

Objetivo general

- Utilizar e interpretar los parámetros y resultados de los distintos métodos y modelos lineales mixtos vistos en clase.

Objetivos específicos

- Realizar un análisis descriptivo de las bases de datos dadas en clase.
- comprender y deducir los modelos lineales mixtos, ANOVA y MANOVA en datos de longitudinales y de medidas repetidas.

INTRODUCCION

Un aspecto fundamental del estudio longitudinal es conocer no sólo los cambios o perfiles individuales, sino también determinar si el cambio es significativo y si se dan diferencias entre los distintos sujetos de la muestra.

Es decir, se trata también de abordar y estudiar el cambio en función del tiempo, razón por la cual se obtienen datos longitudinales de una muestra dada de sujetos que es medida repetidas veces en la misma variable de respuesta

El presente trabajo aborda los conceptos de análisis longitudinales y medidas repetidas en el ámbito de la estadística y el desarrollo de modelos que ayudan a su interpretación y entendimiento con el software estadístico R-studio y SAS academy, en primera instancia se analizarán dos bases de datos una balanceada y otra desbalanceada y se hará uso de modelos lineales mixtos con intercepto aleatorio y así como un enfoque basado en el ANOVA(análisis univariante de la varianza) y MANOVA(análisis multivariante de la varianza) la principal limitación de estos dos últimos modelos es el requerimiento de datos completos y balanceados.

El modelo mixto combina dos modelos de la regresión y tiene dos clases de parámetros, los parámetros de efectos fijos y los parámetros de efectos aleatorios. Es decir, estiman tanto los valores esperados de las observaciones (efectos fijos) como las varianzas y covarianzas de las observaciones (efectos aleatorios).

DESARROLLO

1) A person is conducting an experiment to estimate the woman's average time to cover a flat distance of 35 meters (no running, just walking at a normal pace). Eleven women were asked to walk 35 meters three different times and the time to cover 35 meters by walking at a normal pace was recorded at each trial. The collected data is as follows,

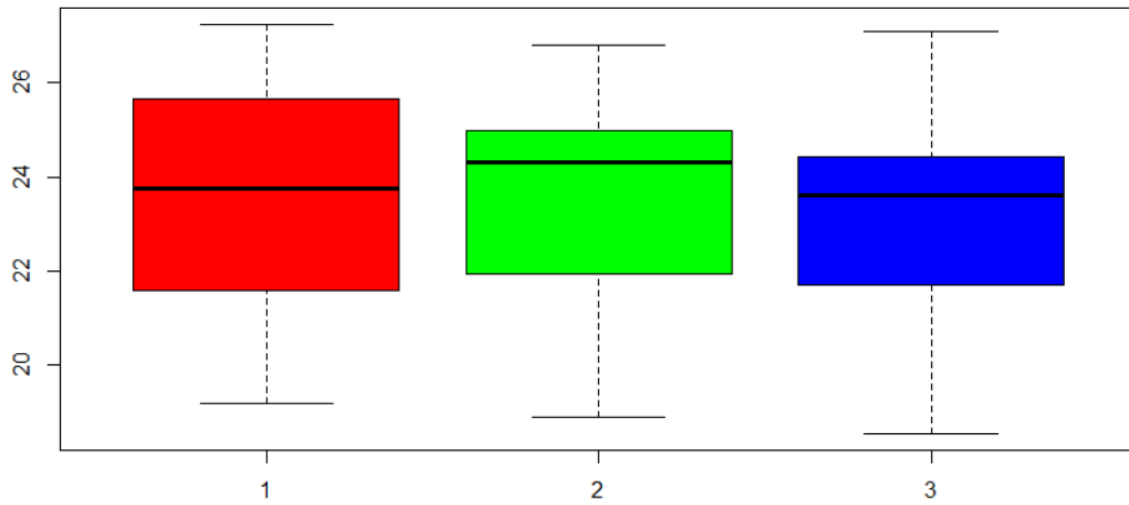
ID	Genero	T1	T2	T3
1	M	23,74	24,3	23,88
2	M	19,91	21,17	20,21
3	M	25,83	25,33	25
4	M	27,23	25,24	25,16
5	M	20,21	19,61	19,24
6	M	19,19	18,91	18,56
7	M	23,26	22,74	23,39
8	M	25,49	24,75	23,6
9	M	23	22,8	23,6
10	M	27,2	26,8	27,1
11	M	24,3	24,7	23,2

ANALISIS DESCRIPTIVO

A continuación se hace un box-plots con el siguiente analisis descriptivo:

Tabla 1

T1	T2	T3
Min. :19.19	Min. :18.91	Min. :18.56
1st Qu.:21.61	1st Qu.:21.95	1st Qu.:21.70
Median :23.74	Median :24.30	Median :23.60
Mean :23.58	Mean :23.30	Mean :22.99
3rd Qu.:25.66	3rd Qu.:25.00	3rd Qu.:24.44
Max. :27.23	Max. :26.80	Max. :27.10



Se deduce de la tabla 1 y del grafico del box plot que los valores del promedio y la mediana no difieren durante los tres tiempos que se les tomaron a los participantes ya que se obtuvo un promedio de 23.58 para T1, 23.30 para T2 y 22.99 para T3, además los valores mínimo y máximo se encuentran entre un rango muy similar durante los tres tiempos sin presencia de valores atípicos que pudieran afectar los modelos por ejemplo el ANOVA.

Matriz de varianzas y covarianzas (Tabla 2)

Los elementos de la diagonal de la matriz contienen las varianzas de las variables, mientras los valores fuera de la diagonal contienen las covarianzas entre los pares de variables, respecto a la covarianza podemos decir que hay una relación lineal positiva entre los pares de variables y las varianzas entre T1, T2, T3 no son abismales entre ellas.

Tabla 2

	T1	T2	T3
T1	8.021816	6.824979	7.058599
T2	6.824979	6.318547	6.342097
T3	7.058599	6.342097	6.879707

PRUEBA DE NORMALIDAD DE SHAPIRO-WILK

Las pruebas de normalidad son importantes para el desarrollo de la prueba ANOVA y MANOVA para tal razón se van a efectuar tanto la shapiro-wilk y el qq-plot.

A continuación, se plantea la prueba de Shapiro-Wilk

Las hipótesis estadísticas son las siguientes:

H0: La variable presenta una distribución normal

H1: La variable presenta una distribución no normal

Toma de decisión:

Sig(p valor) > alfa: No rechazar H0 (normal).

Sig(p valor) < alfa: Rechazar H0 (no normal)

Donde alfa representa la significancia, que en este ejemplo hipotético es igual al 5% (0,05).

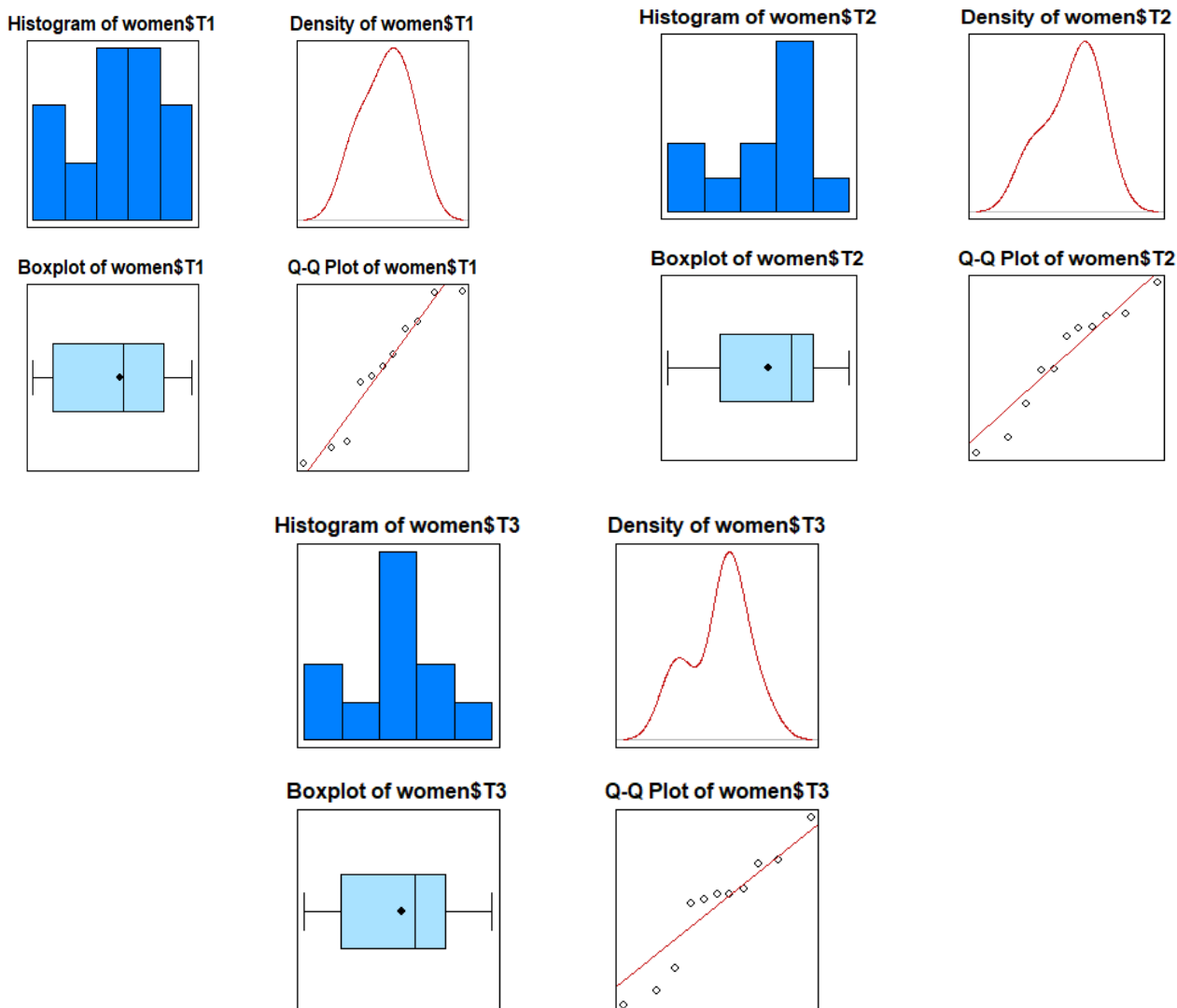
```
Shapiro-wilk normality test
data:  women$T1
W = 0.92601, p-value = 0.3719
```

```
Shapiro-wilk normality test
data:  women$T2
W = 0.92746, p-value = 0.3858
```

```
Shapiro-wilk normality test
data:  women$T3
W = 0.91632, p-value = 0.2892
```


Se concluye de las pruebas de shapiro wilk que las tres variables tienen un comportamiento normal ya que el p-value de las tres variables T1(0.3719),T2(0.3858),T3(0.289) es mayor que el alfa (0.05).

Sabiendo que ya las tres variables tienen un comportamiento normal se presenta ahora una descripción de los histogramas, densidades y qq-plot:



De los histogramas y densidades nos permite inferir un comportamiento normal de los datos, de acuerdo al histograma las variables T1 y T2 parecieran ser un poco sesgada ala izquierda, a continuación, se presentan los gráficos de dispersión y la matriz de correlación.

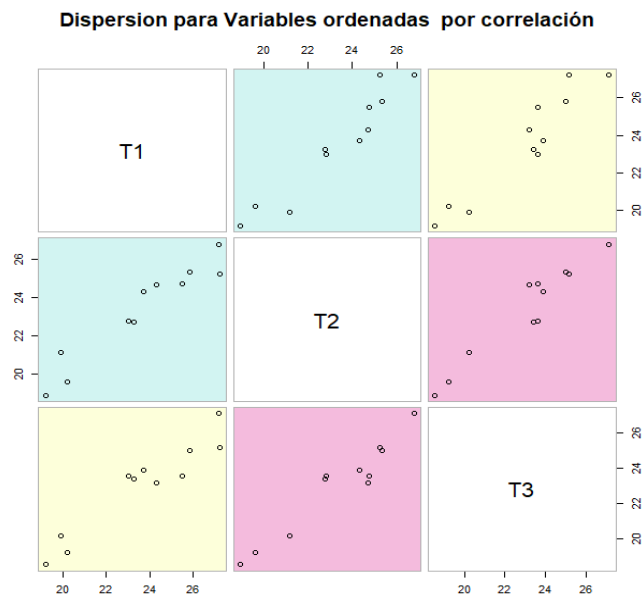
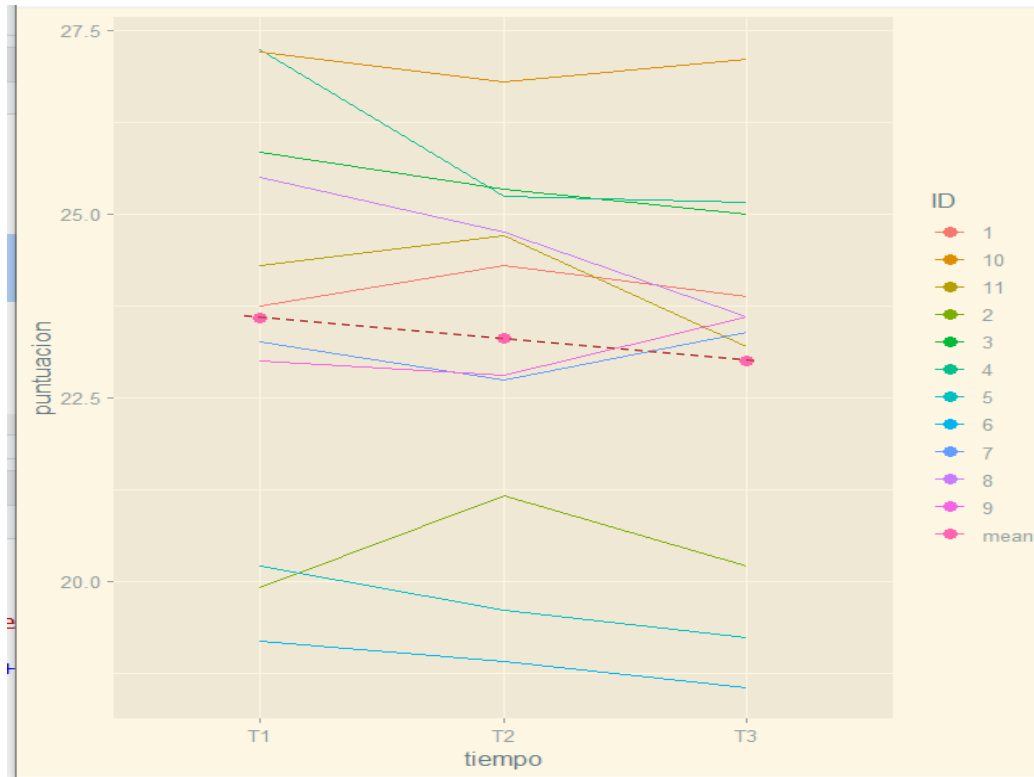


Tabla 3

	T1	T2	T3
T1	1.00	0.96	0.95
T2	0.96	1.00	0.96
T3	0.95	0.96	1.00

De la **tabla 3** se deduce una correlación lineal alta y positiva entre las variables T1,T2,T3, por otra parte Del gráfico de perfiles(**FIGURA_1**) se deduce que en las medidas de tiempo que se obtuvieron para medir el desempeño de los participantes se tuvieron desempeños diferentes hecho que se puede notar en las distintas pendientes donde en algunos es negativa, y en otros empieza negativa y se vuelve positiva, para algunos participantes el desempeño en el tiempo T3 fue más rápido que en el tiempo T1 es decir que en el tiempo T3 se demoraron menos tiempo en completar los 35 metros, mientras en otros participantes el desempeño es contrario al mencionado, los puntos de tendencia del promedio(**mean**) indican claramente una tendencia decreciente lo que refleja que la mayoría de participantes se demoraron un poco más en la primera medida de tiempo y después se fueron demorando menor tiempo.

FIGURA_1



a) According of the material we have seen in class so far, What kind of design this data comes from?

Acorde a lo expresado en los objetivos y la introducción del presente trabajo el tópico a abordar son los analisis de medias repetidas de sujetos que son homogeneos, Cuando se realiza un estudio con datos longitudinales, las observaciones obtenidas se pueden organizan de dos formas o arreglos es decir formato ancho o largo, para este estudio se va a emplear el formato ancho.

c) Fit a MANOVA, ANOVA and a LMM to this data set and interpret the outputs. Establish clearly the hypotheses to be tested. What approach and Why would you recommend?

MANOVA

El análisis de varianza multivariado (Manova) con medidas repetidas es otra alternativa para el análisis de datos longitudinales.

Cabe tener en cuenta que los supuestos básicos del modelo *MANOVA* son que: a) las respuestas de los sujetos son independientes entre sí, b) la distribución de las múltiples variables dependientes es normal multivariada y c) el conjunto de datos ha de ser completo sin observaciones perdidas.

el método de análisis multivariado ayuda a probar si los vectores de medias son iguales o diferentes (hipótesis de la diferencia entre vectores de medias de datos multivariados).

Procedimiento MEANS					
Variable	N	Media	Desv. est.	Mínimo	Máximo
T1	11	23.5781818	2.8322811	19.1900000	27.2300000
T2	11	23.3045455	2.5136721	18.9100000	26.8000000
T3	11	22.9854545	2.6223095	18.5600000	27.1000000

Información sobre el nivel de las medidas repetidas					
Variable dependiente		T1	T2	T3	
Nivel de method		1	2	3	

Criterio de test MANOVA y estadísticos F exactos para la hipótesis de efecto no method H = Tipo III Matriz SSCP para T. indemethod Matriz SSCP de error E =					
S=1 M=0 N=3.5					
Estadístico	Valor	Valor F	Num DF	Den DF	Pr > F
Lambda de Wilks	0.66790437	2.24	2	9	0.1626
Traza de Pillai	0.33209563	2.24	2	9	0.1626
Traza Hotelling-Lawley	0.49722032	2.24	2	9	0.1626
Raíz más grande de Roy	0.49722032	2.24	2	9	0.1626

Del cuadro anterior se observa que el test de Manova realizado en SAS arroja los resultados concernientes a los cuatro estadísticos, esto con el objetivo de probar la siguiente prueba de hipótesis(**prueba_1**) en donde esta plantea que si se tiene un $p < 0.05$ se rechaza la hipótesis nula y por lo tanto hay diferencia entre los promedios asociados a cada medida del tiempo de los 11 participantes.

Prueba_1

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

Where $\mu_j = E(Y_{ij}), i = 1, 2, \dots, n$ and $j = 1, 2, \dots, k$.

Se infiere de los resultados que con el valor $p=0.1623$ no se rechaza la hipótesis nula y por lo tanto no hay diferencia significativa entre los promedios de las medidas del tiempo que se hicieron.

ANOVA

Con el Anova también se va a plantear la prueba de hipótesis anterior

Procedimiento GLM
Análisis de medidas repetidas de la varianza
Test de hipótesis univariante para efectos dentro del sujeto

Origen	DF	Tipo III SS	Cuadrado de la media	Valor F	Pr > F	Pr > Adj F	
						G - G	H - F
method	2	1.87589697	0.93794848	2.83	0.0828	0.0865	0.0828
Error(method)	20	6.62930303	0.33146515				

Greenhouse-Geisser Epsilon	0.9460
Huynh-Feldt Epsilon	1.1600

Para validar el test planteado toca asegurar si hay simetría compuesta o no, si el valor de la épsilon de Greenhouse-Geisser es igual a Uno hay simetría compuesta y el valor de la prueba F es correcto, hecho que se corrobora con el resultado obtenido ya que se obtiene un valor aproximado a uno que 0.946 y por lo tanto no se debe hacer corrección a los grados de libertad y Luego los promedios en las mediciones son los mismos.

MODELO LMM

Para emplear el modelo lineal mixto con intercepto aleatorio se emplea el formato largo

Procedimiento Mixed

Información del modelo	
Conjunto de datos	WORK.ORION
Variable dependiente	y
Estructura de covarianza	Simetría compuestas
Efecto de sujeto	sujeto
Método de estimación	REML
Método de varianza del residual	Perfil
Método SE de efectos fijos	Basado en el modelo
Método de grados de libertad	Between-Within

Número de observaciones	
N.º observaciones leídas	33
N.º observaciones usadas	33
N.º observaciones no usadas	0

VALORES DE EFECTOS FIJOS

Solución para efectos fijos						
Efecto	method	Estimación	Error estándar	DF	t valor	Pr > t
Intercept		22.9945	0.8019	10	28.68	<.0001
method	1	0.5836	0.2455	20	2.38	0.0275
method	2	0.3100	0.2455	20	1.26	0.2212
method	3	0

Test de tipo 3 de efectos fijos				
Efecto	DF Num	Den DF	Valor F	Pr > F
method	2	20	2.83	0.0828

Al observar la Prueba f y teniendo en cuenta $(k-1)$ y $(k-1*n-1)$

$$(k-1*n-1)=3-1*11-1= 20$$

los grados de Libertad son los mismos del anova luego se captura la misma información que es 2 y 20 ademas Los promedios son iguales.

Diferencias de medias de mínimos cuadrados							
Efecto	method	_method	Estimación	Error estándar	DF	t valor	Pr > t
method	1	2	0.2736	0.2455	20	1.11	0.2782
method	1	3	0.5836	0.2455	20	2.38	0.0275
method	2	3	0.3100	0.2455	20	1.26	0.2212

Se analizan los p y se comparan los métodos a pesar de que la prueba resultó que las medias eran iguales se destaca que la medición 1 y 3 muestra que estas son diferentes y que por tanto si hay diferencia entre esas mediciones ahora se va a corroborar con un modelo equivalente con intercepto aleatorio en el software SAS.

Modelo equivalente con intercepto aleatorio

Diferencias de medias de mínimos cuadrados							
Efecto	method	_method	Estimación	Error estándar	DF	t valor	Pr > t
method	1	2	0.2736	0.2455	20	1.11	0.2782
method	1	3	0.5836	0.2455	20	2.38	0.0275
method	2	3	0.3100	0.2455	20	1.26	0.2212

Se concluye que los promedios de tiempo medido T1 y T3 en los 35 metros para las mujeres son diferentes entre ellos además el grafico de perfiles también arrojo una diferencia y una tendencia decreciente en ellos.

PUNTO_2

2) (Taken from Fitzmaurice et al.) Subsample (N=300) of data on FEV1 and height for girls living in Topeka from the Six Cities Study of Air Pollution and Health. Source: Data courtesy of Dr. Doug Dockery. Reference: Dockery, D.W., Berkey, C.S., Ware, J.H., Speizer, F.E. and Ferris, B.G. (1983). Distribution of FVC and FEV1 in children 6 to 11 years old. American Review of Respiratory Disease, 128, 405-412.

Description:

The Six Cities Study of Air Pollution and Health was a longitudinal study designed to characterize lung growth as measured by changes in pulmonary function in children and adolescents, and the factors that influence lung function growth. A cohort of 13,379 children born on or after 1967 was enrolled in six communities across the U.S.: Watertown (Massachusetts), Kingston and Harriman (Tennessee), a section of St. Louis (Missouri), Steubenville (Ohio), Portage (Wisconsin), and Topeka (Kansas). Most children were enrolled in the first or second grade (between the ages of six and seven) and measurements of study participants were obtained annually until graduation from high school or loss to follow-up. At each annual examination, spirometry, the measurement of pulmonary function, was performed and a respiratory health questionnaire was completed by a parental guardian. The dataset contains a subset of the pulmonary function data collected in the Six Cities Study. The data consist of all measurements of FEV1, height and age obtained from a randomly selected subset of the female participants living in Topeka, Kansas. The random sample consists of 300 girls, with a minimum of one and a maximum of twelve observations over time.

Variable List:

Subject ID, Height, Age, Initial Height, Initial Age, Log(FEV1) (FEV: Forced Expiratory Volume).

The data is attached!! (Six_Cities_Pollution.txt)

ANALISIS DESCRIPTIVO

- a) Conduct **a complete descriptive analysis** , based on graphics and summary statistics of this data set.

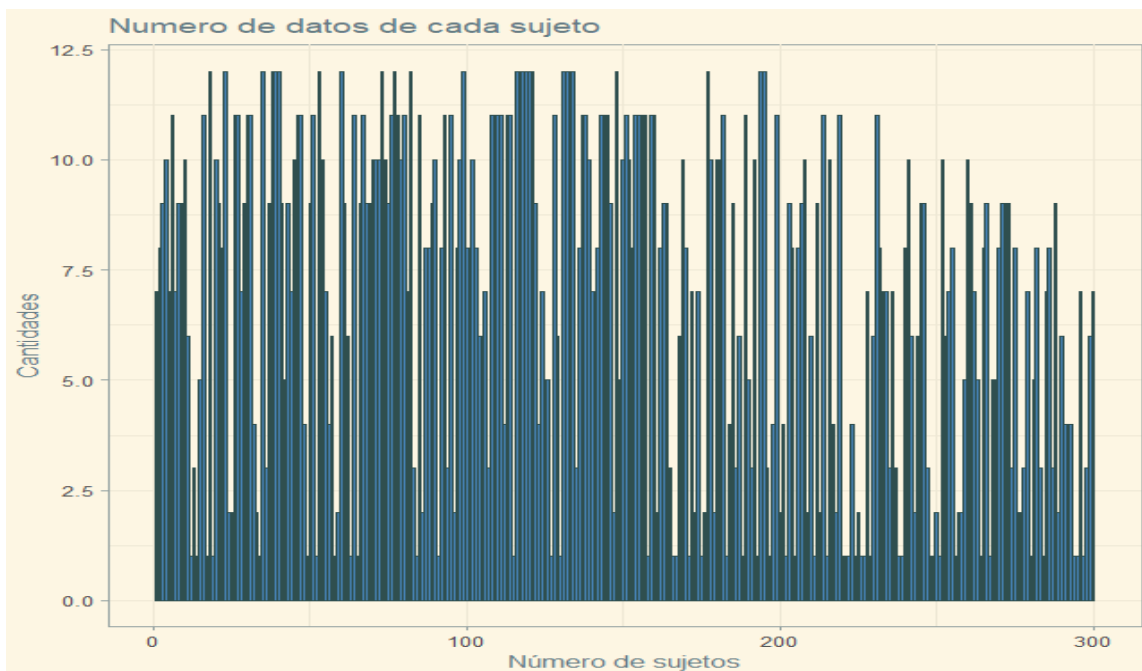
A Continuacion se presenta un resumen estadístico de las variables **TABLA 4** cabe destacar que este estudio es un periodo de seguimiento que se les hace a 300 personas de edades pequeñas hasta conforme ellas van creciendo o se retiran del estudio las edades en promedio fue hasta los 13 años con una altura en promedio de 1.496 metros.

TABLA 4

Height	Age	Init_Height	Init_Age	Log_fev
Min. :1.110	Min. : 6.434	Min. :1.110	Min. : 6.434	Min. :-0.6932
1st Qu.:1.370	1st Qu.: 9.717	1st Qu.:1.220	1st Qu.: 7.136	1st Qu.: 0.5481
Median :1.540	Median :12.595	Median :1.260	Median : 7.781	Median : 0.8671
Mean :1.497	Mean :12.566	Mean :1.276	Mean : 8.030	Mean : 0.8152
3rd Qu.:1.620	3rd Qu.:15.366	3rd Qu.:1.320	3rd Qu.: 8.449	3rd Qu.: 1.0978
Max. :1.790	Max. :18.691	Max. :1.720	Max. :14.067	Max. : 1.5953

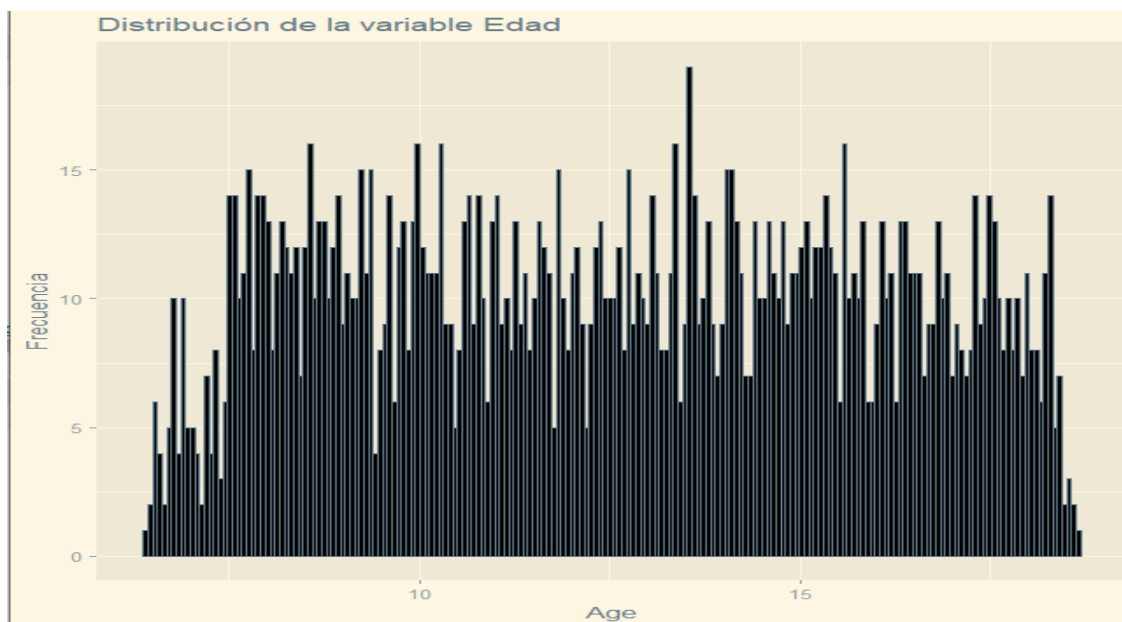
Como se deduce de la **TABLA A** las cantidades que se le tomaron a cada sujeto (300) varían ya que a algunas se les tomaron mas de 10 datos conforme transcurrían los años y a los que menos se les tomaron datos quizás abandonaron el estudio.

TABLA A (numero de sujetos vs cantidades)




Asimismo, las edades varían desde los 6 años a los 18 años durante el estudio (**tabla B**)

TABLA B AGE VS frecuencia



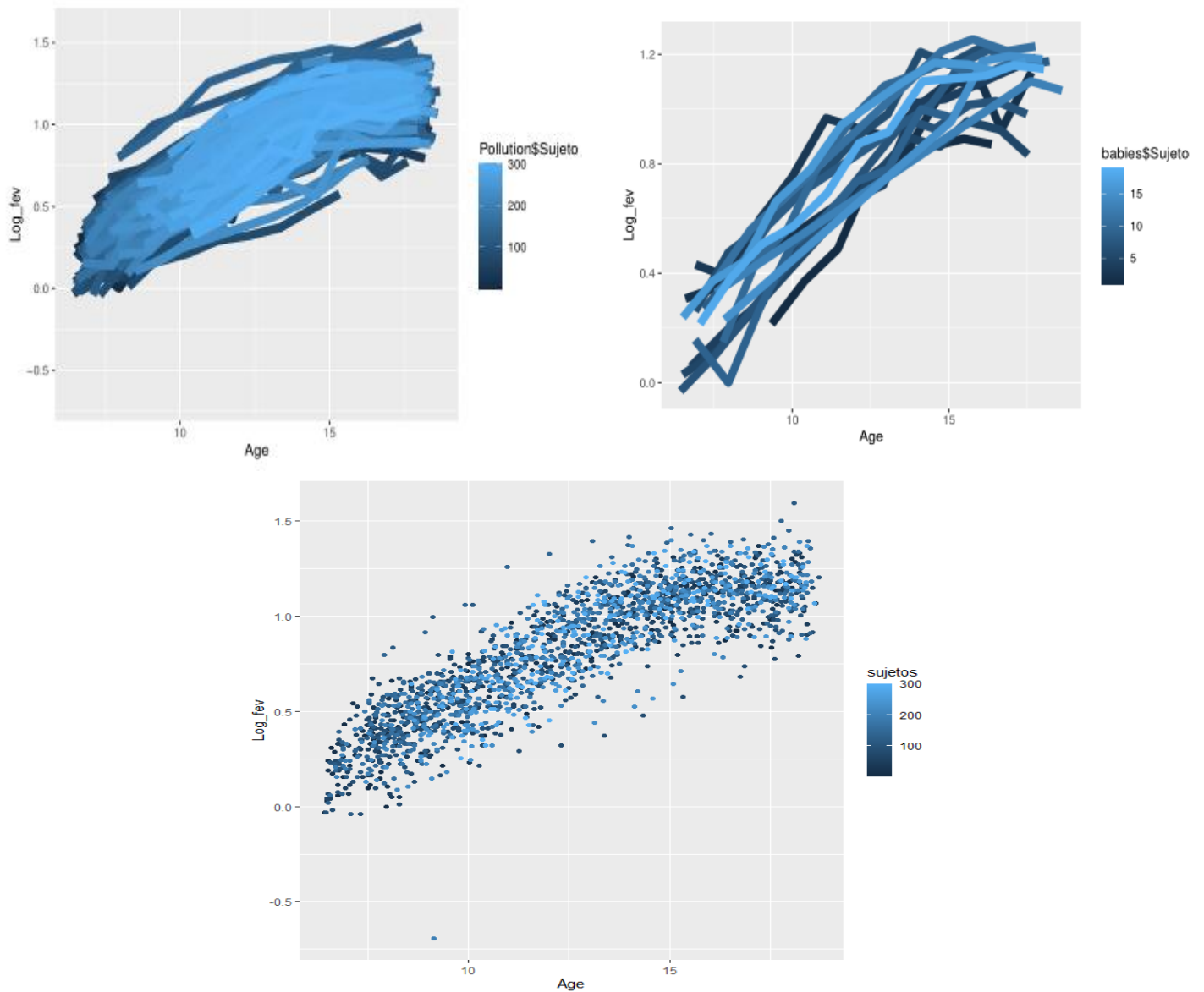
Como se mencionó anteriormente esta base de datos es desbalanceada y en rasgos generales a el sujeto 1 se le tomaron 7 medidas y su edad durante el estudio vario de los 9 a los 16 años y asi sucesivamente como se denota en la **TABLA C**

TABLA C

Sujeto	n		Sujeto	Age
1	7		1	9.3415
2	8		1	10.3929
3	9		1	11.4524
4	10		1	12.4600
5	7		1	13.4182
			1	15.4743
			1	16.3723
			2	6.5873
			2	7.6496
			2	12.7392
			2	13.7741
			2	14.6940
			2	15.8220
			2	16.6680
			2	17.6318

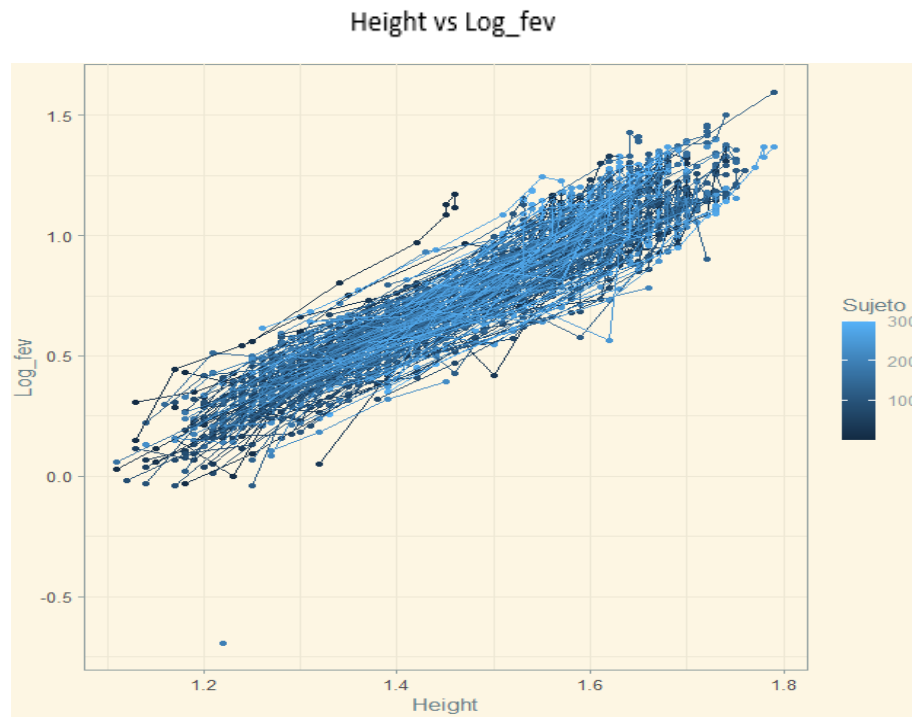
A continuación se presentan los gráficos de perfiles para mirar el comportamiento de las variables con las medidas de volumen espiratorio forzado transformado logarítmicamente(log_fev), en el set de **figuras A** se ilustra como conforme pasan los años el volumen espiratorio aumenta, algo normal ya que conforme crecen los niños el volumen crece de manera lineal y creciente para este caso, no obstante a la edad de 15 en algunos pacientes esta empieza a disminuir y en algunos casos a ser constante perdiendo la tendencia creciente que llevaban durante las primeras etapa del estudio.

FIGURAS A (Age vs Log_fev)



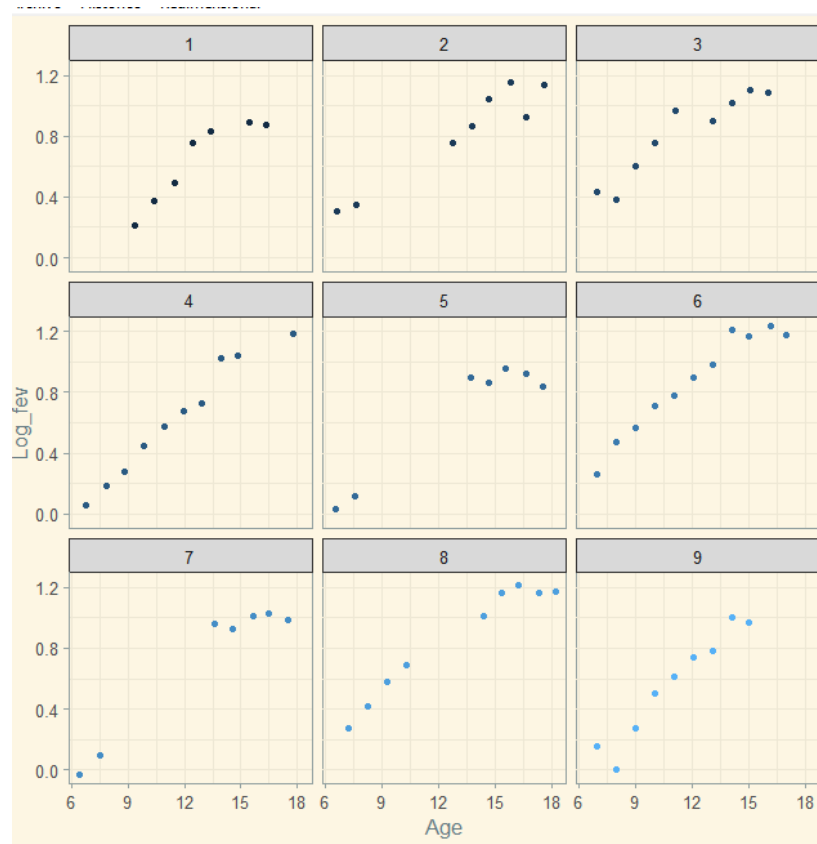
En la **figura B** se observa un comportamiento creciente y de pendiente positiva en donde a medida que aumenta la altura de los pacientes va aumentando el volumen espiratorio algo consecuente ya que se supone que su capacidad pulmonar se va desarrollando siempre y cuando no hallan factores externos que afecten el desarrollo pulmonar.

FIGURA B



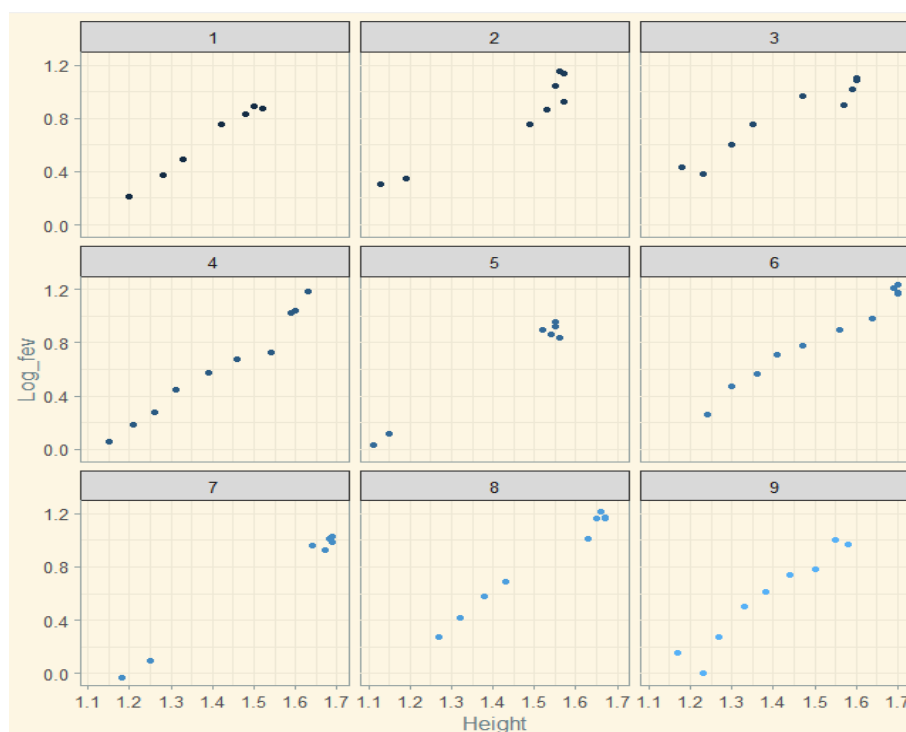
Se prestará particular atención a la variable edad y altura En la **figura c** se observa un patrón claro, en donde las 9 nubes de puntos (9 pacientes escogidos al azar) tienen un comportamiento diferente despues de la edad de 12 o 15 años en donde para algunos el volumen espiratorio empieza a ser constante o decreciente.

Figura c Age vs Log_fev



En la **figura D** se observa una tendencia creciente en algunos pacientes de lo que se infiere es que aumenta la altura y también el volumen espiratorio.

Figura D Height vs log_fev



A continuación, se presentan las correlaciones entre las distintas variables y se denota como hay una correlación lineal alta y positiva entre las variables aquellas que tienen una alta correlación son las variables Age y Height con el volumen espiratorio.

CORRELACIONES

	Sujeto	Height	Age	Init_Height	Init_Age	Log_fev
Sujeto	1.0000000	0.0923640	0.0748050	0.5209193	0.6757846	0.0928586
Height	0.0923640	1.0000000	0.8883297	0.3221311	0.1563213	0.9310110
Age	0.0748050	0.8883297	1.0000000	0.1305732	0.1560883	0.8715364
Init_Height	0.5209193	0.3221311	0.1305732	1.0000000	0.7822197	0.2862395
Init_Age	0.6757846	0.1563213	0.1560883	0.7822197	1.0000000	0.1547218
Log_fev	0.0928586	0.9310110	0.8715364	0.2862395	0.1547218	1.0000000

Acorde a este analisis descriptivo se infiere que si ha habido una afectación en los pulmones de los sujetos.

Modelo LMM

b) Fit a LMM to this data set and **interpret** the outputs.

Acontinuacion se efectua un modelo lineal mixto con intercepto aleatorio en el software R studio en los variables significativas son Age y Height.

```
resu<-suppressMessages(lmer(Log_fev~ Height+Age+(1|Sujeto),data=Pollution))
summary(resu)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: Log_fev ~ Height + Age + (1 | Sujeto)
## Data: Pollution
##
## REML criterion at convergence: -4482.3
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -5.8718 -0.5199  0.0706  0.5987  2.8289
##
## Random effects:
## Groups Name Variance Std.Dev.
## Sujeto (Intercept) 0.011155 0.10562
## Residual 0.004056 0.06369
## Number of obs: 1994, groups: Sujeto, 300
##
## Fixed effects:
##              Estimate Std. Error      df t value Pr(>|t|)
## (Intercept) -1.858e+00  3.072e-02 1.966e+03  -60.49  <2e-16 ***
## Height      1.619e+00  3.013e-02 1.894e+03   53.72  <2e-16 ***
## Age         1.977e-02  1.312e-03 1.865e+03    15.07  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##      (Intr) Height
## Height -0.961
## Age     0.834 -0.935
```

En los efectos fijo se deducen que las dos variables Age y Height junto con la pendiente son significativas

A continuación, se presentan los valores estimados de los Efectos fijos (beta gorros) y efectos aleatorios (desviación estimada de (y) Y Desviación estimada del Intercepto aleatorio.

$$\hat{\Theta} = (\hat{\beta}_0 = -1.86, \hat{\beta}_1 = 1.62, \hat{\beta}_2 = 0.029, \hat{\sigma}_y = 0.064, \hat{\sigma}_{bo} = 0.106)^\top$$

el modelo ajustado

$$\text{Logfev}_{ij} \sim N(\hat{\mu}, 0.064^2)$$

$$\hat{\mu}_{ij} = -1.86 + 1.62\text{Height}_{ij} + 0.029\text{Age} + b_{0i} + b_{1i}\text{Height}_{ij} + b_{2i}\text{Age}_{ij}$$

CONCLUSIONES

Se concluye que Los diseños longitudinales sirven para estudiar los procesos de cambio directamente asociados con el paso del tiempo.

De las bases de datos analizadas se analizaron no sólo los cambios o perfiles individuales, sino que también se determinó si el cambio era significativo y si se daban diferencias entre los distintos sujetos de la muestra. Es importante destacar que en la práctica los estudios longitudinales suelen tener datos no balanceados e incompletos caso que se vio en el segundo punto del trabajo y para estos casos se descarta el ANOVA y MANOVA y se emplea un modelo lineal mixto.

BIBLIOGRAFIA

[1] Notas de clase analisis de datos longitudinales y medidas repetidas, profesor Juan Carlos Salazar, Universidad nacional de Colombia sede medellin, 2021.

[2] Analisis de datos Longitudinales, Carlos Javier rincón, 2019, link: <https://bookdown.org/cjrinconr/adl2019/ADL2019.pdf>, fecha de consulta: 07/11/2021.

[3] Estudios longitudinales de medidas repetidas. Modelos de diseño y análisis, Jaume Arnau y Roser Bono, Escritos de Psicología vol.2 no.1 Málaga dic. 2008, link https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1989-38092008000300005, fecha de consulta: 07/11/2021.