# Introducción a los modelos mixtos (SIA 3011003)

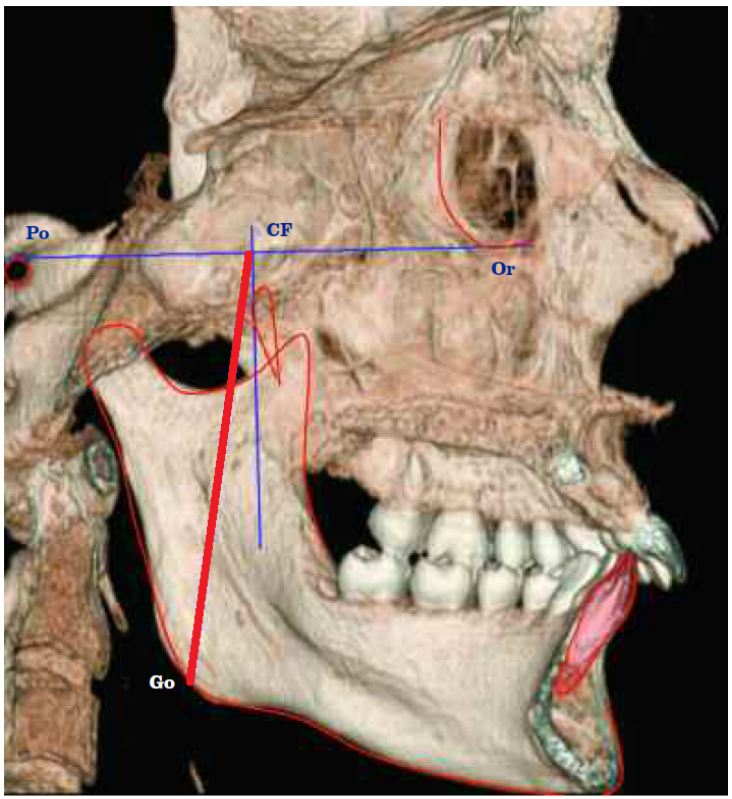
Profesor Juan Carlos Salazar-Uribe
jcsalaza@unal.edu.co

¿Qué son medidas repetidas? Las medidas repetidas se obtienen cuando una respuesta se mide repetidamente en un conjunto de unidades. Por ejemplo, usando un espectrofotómetro[1], se pueden hacer varias mediciones en el capó de un carro (es decir, sobre la tapa del motor) para evaluar si la pintura ha sido o no aplicada uniformemente o si ha sido repintado.

[1] El espectrofotómetro es un instrumento con el que se apoya la espectrofotometría para medir la cantidad de intensidad de luz absorbida. Sin duda, el espectrofotómetro es la herramienta que aporta una mayor fiabilidad a la hora de obtener información precisa sobre el código de color de un carro.

¿Qué son datos longitudinales? Datos longitudinales, a veces llamados datos de panel, son datos que se recopilan a través de una serie de observaciones repetidas de los mismos sujetos durante un cierto período de tiempo, y son útiles para evaluar estadísticamente el cambio. Los datos longitudinales siguen efectivamente la misma muestra a lo largo del tiempo, lo que difiere fundamentalmente de los datos de corte transversal (cross-sectional data) porque siguen a los mismos sujetos durante algún tiempo (dinámico), mientras que los datos de corte transversal muestrean diferentes sujetos (ya sean individuos, empresas, países o regiones) en un solo punto en el tiempo (estático).

Por ejemplo, se mide la evolución de la denominada altura facial posterior (ver gráfico abajo[2]) a un grupo de niñas y niños de una cierta ciudad, cada año y por un espacio de 10 desde los 6 años. (desbalanceo):




▲ **Figura 13.24.** La distancia lineal desde los puntos Gonion (Go) y centro de la cara (CF) representan la magnitud Altura facial posterior.

Figura 1: Altura Facial Posterior AFP

[2]Fuente: http://ortoface.com/wp-content/uploads/2016/12/ANALISIS-DE-RICKETTS.pdf

- Los datos longitudinales son un caso especial de medidas repetidas.
- Los datos longitudinales y las medidas repetidas se utilizan ampliamente en las ciencias sociales, ciencias de la salud, incluso entre economistas, politólogos y sociólogos.
- Los datos longitudinales son datos que se recopilan secuencialmente a los mismos sujetos o clusters a lo largo del tiempo. Esta es una característica distintiva de este tipo de datos.

Datos longitudinales y de panel. (Frees, 2004[3]) La estadística se trata de datos. Es la disciplina que se ocupa de la recopilación, resumen y análisis de datos para hacer aseveraciones sobre los fenómenos del mundo. Cuando los analistas recolectan datos, en realidad están recolectando información cuantificada, es decir, transformada a una escala numérica. Hay muchas reglas bien entendidas para reducir datos, utilizando medidas de resumen numéricas o gráficas. Estas medidas de resumen se pueden vincular a una representación teórica, o modelo, de los datos. Con un modelo que está calibrado por datos, se pueden hacer afirmaciones sobre sobre los fenómenos del mundo que nos rodea.

[3]Frees, W. (2004). *Longitudinal and Panel Data Analysis and Applications in the Social Sciences* Cambridge University Press

Los datos transversales, o una sección transversal de una población de estudio, en estadística y econometría son un tipo de datos recopilados mediante la observación de muchos sujetos (como individuos, empresas, países o regiones) en el mismo momento, sin tener en cuenta las diferencias en el tiempo. El análisis de datos transversales o de corte transversal generalmente consiste en comparar las diferencias entre los sujetos.

Por ejemplo, si queremos medir los niveles actuales de obesidad en una población, podemos tomar una muestra aleatoria de 1000 personas de esa población (también conocida como sección transversal de esa población), medir su peso y estatura (para calcular, por ejemplo, el Índice de Masa Corporal IMC) y calcular qué porcentaje de esa muestra se categoriza como obesa. Esta muestra transversal nos proporciona una "foto instantánea" de esa población, en ese momento único. Tenga en cuenta que no sabemos, con base en una muestra transversal, si la obesidad está aumentando o disminuyendo; solo podemos describir la proporción actual[4].

[4] http://en.wikipedia.org/wiki/Cross-sectional_data

Como usuarios de métodos estadísticos, identificamos una entidad básica que medimos recopilando información en una escala numérica. Esta entidad básica se conoce como *unidad de investigación* o *unidad de observación*, o *sujeto* o *individuo*.

El análisis de regresión y el análisis de series temporales son dos métodos estadísticos aplicados que se utilizan para analizar datos. Con el análisis de regresión, analizamos datos de una muestra representativa de sujetos (generalmente recopilamos la variable de interés, una respuesta, y el objetivo es evaluar cómo algunos factores afectan esta respuesta); este tipo de análisis es estático.

En contraste, con el análisis de series temporales, identificamos uno o más sujetos y los observamos a lo largo del tiempo. Esto nos permite estudiar las relaciones a lo largo del tiempo, el aspecto dinámico de un problema. Cuando utilizamos el análisis de series de tiempo, nos restringimos a un número limitado de sujetos (al menos uno) que tiene muchas observaciones a lo largo del tiempo.

El análisis de datos longitudinales representa una combinación de regresión y análisis de series de tiempo. Al igual que con muchos conjuntos de datos de regresión, los datos longitudinales se componen de una sección transversal de sujetos. A diferencia de los datos de regresión, con los datos longitudinales observamos a los sujetos a lo largo del tiempo. A diferencia de los datos de series temporales, con los datos longitudinales observamos muchos sujetos. La observación de una amplia sección transversal de sujetos a lo largo del tiempo nos permite estudiar aspectos dinámicos, así como transversales, de un problema.

Por otro lado, **los datos de panel**, provienen de encuestas a individuos. En este contexto, un 'panel' es un grupo de individuos encuestados repetidamente a lo largo del tiempo. Los datos transversales difieren de los datos de series temporales, en los que se observa la misma entidad agregada o de pequeña escala en varios puntos en el tiempo, por ejemplo, los datos longitudinales, que siguen los cambios de un sujeto a lo largo del tiempo. Otra variante, **datos de panel** (o datos transversales de series temporales (TSCS)), combina ambos y analiza múltiples sujetos y cómo ellos(as) cambian con el transcurso del tiempo. El análisis de panel utiliza datos de panel para examinar los cambios en las variables a lo largo del tiempo y las diferencias en las variables entre sujetos.

En este contexto, los modelos para datos longitudinales a veces se diferencian de los datos de regresión y de series de tiempo a través de sus subíndices dobles. Con esta notación, podemos distinguir entre respuestas por sujeto y ocasión de medición.

$$Y_{ij}$$

donde $i$ representa el $i$-ésimo sujeto y $j$ la $j$-ésima ocasión de medición. El subínce $j$ algunas veces se puede reemplazar por el subíndice $t$ sin pérdida de generalidad.

Para ello, defina $y_{it}$ como la respuesta para el sujeto $i$ durante el período de tiempo $t$. Un conjunto de datos longitudinales consta de observaciones del $i$ésimo sujeto durante $t = 1, 2, \ldots, T$ períodos de tiempo, para cada uno de $i = 1, 2, \ldots, n$ sujetos. Así, observamos:

$$\begin{aligned} \text{Primer sujeto} &- \{y_{11}, y_{12}, \ldots, y_{1T_1}\} \\ \text{Segundo sujeto} &- \{y_{21}, y_{22}, \ldots, y_{2T_2}\} \\ &\vdots \\ n\textit{th} \text{ sujeto} &- \{y_{n1}, y_{n2}, \ldots, y_{nT_n}\} \end{aligned}$$

Esta forma de recolectar información de datos longitudinales se conoce comunmente como "Formato Ancho" (Wide Format en inglés, WF) y es una forma frecuente de registrar este tipo de datos.

Un ejemplo genérico de este tipo datos:

| A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|
| **ID** | **y1** | **y2** | **y3** | **y4** | **y5** | **y6** |
| 1 | y11 | y12 | y13 | y14 | y15 | y16 |
| 2 | y21 | y22 | y23 | y24 | y25 | y26 |
| 3 | y31 | y32 | y33 | y34 | y35 | y36 |
| 4 | y41 | y42 | y43 | y44 | y45 | y46 |
| 5 | y51 | y52 | y53 | y54 | y55 | y56 |

Figura 2: Ilustración datos longitudinales en formato ancho

No es necesario que todos los sujetos tengan completas todas las mediciones (datos pérdidos o missing) ni que tengan el mismo número de mediciones (desbalanceo):

| A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|
| **ID** | **y1** | **y2** | **y3** | **y4** | **y5** | **y6** |
| 1 | y11 | y12 | y13 | y14 | y15 | y16 |
| 2 | y21 | y22 | y23 | | y25 | y26 |
| 3 | y31 | | y33 | y34 | y35 | y36 |
| 4 | y41 | y42 | y43 | y44 | y45 | |
| 5 | y51 | y52 | | y54 | y55 | y56 |

Figura 3: Ilustración datos longitudinales en formato ancho con datos missing y desbalanceo

Un ejemplo concreto de formato ancho (datos completos y balanceados):

| Obs | Group | Id | $y_{152}$ | $y_{174}$ | $y_{201}$ | $y_{227}$ |
|---|---|---|---|---|---|---|
| 1 | Control | 1 | 2.79 | 3.10 | 3.30 | 3.38 |
| 2 | Control | 2 | 3.30 | 3.90 | 4.34 | 4.96 |
| 3 | Control | 3 | 3.98 | 4.36 | 4.79 | 4.99 |
| 4 | Control | 4 | 4.36 | 4.77 | 5.10 | 5.30 |
| 5 | Control | 5 | 4.34 | 4.95 | 5.42 | 5.97 |
| 6 | Control | 6 | 4.59 | 5.08 | 5.36 | 5.76 |
| 7 | Control | 7 | 4.41 | 4.56 | 4.95 | 5.23 |
| 8 | Control | 8 | 4.24 | 4.64 | 4.95 | 5.38 |
| 9 | Control | 9 | 4.82 | 5.17 | 5.76 | 6.12 |
| 10 | Control | 10 | 3.84 | 4.17 | 4.67 | 4.67 |
| 11 | Control | 11 | 4.07 | 4.31 | 4.90 | 5.10 |
| 12 | Control | 12 | 4.28 | 4.80 | 5.27 | 5.55 |

Sin embargo, a fin de implementar procedimientos de visualización y análisis estadístico (clásico y moderno), generalmente en la práctica, se trabaja con un formato conocido como "Formato Largo" (Long Format en inglés, LF):

| ID | y |
|---|---|
| 1 | y11 |
| 1 | y12 |
| 1 | y13 |
| 1 | y14 |
| 1 | y15 |
| 1 | y16 |
| : | |
| 6 | y61 |
| 6 | y62 |
| 6 | y63 |
| 6 | y64 |
| 6 | y65 |
| 6 | y66 |

Figura 4: Ilustración datos longitudinales en formato largo

Un ejemplo concreto de formato largo (datos completos y balanceados):

| ID | Group | $Y$ |
|---|---|---|
| 1 | Control | 2.79 |
| 1 | Control | 3.10 |
| 1 | Control | 3.30 |
| 1 | Control | 3.38 |
| 2 | Control | 3.30 |
| 2 | Control | 3.90 |
| 2 | Control | 4.34 |
| 2 | Control | 4.96 |
| ⋮ | ⋮ | ⋮ |
| 12 | Control | 4.28 |
| 12 | Control | 4.80 |
| 12 | Control | 5.27 |
| 12 | Control | 5.55 |

(Littell et al. 2000) Los modelos estadísticos para datos son descripciones de cómo se podrían producir los datos y básicamente constan de dos partes: (1) una fórmula que relaciona la respuesta a todas las variables explicativas (por ejemplo, efectos), y (2) una descripción de la distribución de probabilidad asumida para caracterizar la variación aleatoria que afecta la respuesta observada.

**Aspectos del Modelo Líneal General GLM**. Considere un experimento con 5 fármacos A, B, C, D y E, aplicados a sujetos para controlar la presión arterial. Sea $\mu_i$ la presión arterial media de los sujetos tratados con el fármaco $i$ ($i = A,\ B,\ C,\ D,\ E$). El modelo más simple para describir cómo se produjeron las observaciones de este experimento para el fármaco A es

$$Y_A = \mu_A + e$$

Es decir, una observación de la presión arterial $Y_A$ en un sujeto dado tratado con el fármaco A es igual a la media del fármaco A más la variación aleatoria resultante de lo que sea particular de un sujeto dado que no sea el fármaco A. La variación aleatoria, denotada por el término $e$, se llama el error en $Y$. De ello se deduce que $e$ es una variable aleatoria con media cero y varianza $\sigma^2$.

Esta es la versión más simple de un modelo estadístico lineal, es decir, un modelo en el que la observación es la suma de los términos del lado derecho del modelo que surgen del tratamiento u otros factores explicativos más el error aleatorio. El modelo $Y_A = \mu_A + e$ se denomina modelo de medias porque el único término en el lado derecho del modelo, además de la variación aleatoria, es una media de tratamiento. Tenga en cuenta que la media es también el valor esperado de $Y_A$. La media se puede modelar de varias maneras. El primer enfoque conduce a un modelo de efectos. Usted puede definir el efecto de la droga A como $\alpha_A$ tal que $\mu_A = \mu + \alpha_A$, donde $\mu$ se define como el intercepto. Esto conduce al llamado modelo de análisis de varianza de una vía (ONE WAY ANOVA).

$$Y_A = \mu + \alpha_A + e$$

que es la forma más simple de un modelo de efectos.

Tenga en cuenta que el modelo de efectos tiene más parámetros, en este caso 6 ($\mu$, $\alpha_A$, $\alpha_B$, $\alpha_C$, $\alpha_D$, $\alpha_E$) y 5 niveles de factores (en este caso: A, B, C, D y E). Se dice que estos modelos están sobreparametrizados porque hay más parámetros para estimar que elementos únicos de información. Dichos modelos requieren alguna restricción en la solución para estimar los parámetros. A menudo, en este tipo de modelos, la restricción implica definir $\mu$ como la media general, lo que implica $\alpha_A = \mu_A - \mu$ y, por lo tanto,

$$\sum_{i=A}^{E} \alpha_i = 0$$

Esto se llama una restricción de suma igual a cero o suma a cero.

Su ventaja es que si el número de observaciones por tratamiento es igual, es fácil de interpretar. Sin embargo, para diseños complejos con observaciones desiguales por tratamiento, la restricción de suma a cero se vuelve intratable, mientras que las restricciones alternativas son de aplicación más general. Los procedimientos SAS utilizan la restricción de que el último nivel de factor, en este caso $\alpha_E$, se establece en cero. En general, para el modelo de efectos, la estimación de la media $\mu_A = \mu + \alpha_A$ es única e interpretable, pero los componentes individuales $\mu$ y los $\alpha_i$ ($i=A,\ldots,E$) puede no serlo.

Otro enfoque para modelar $\mu_A$, que sería apropiado si los niveles A a E representaran dosis o cantidades de un fármaco administrado a los pacientes, es utilizar la regresión lineal. Específicamente, sea $X_A$ la dosis de fármaco correspondiente al tratamiento A, $X_B$ la dosis de fármaco correspondiente al tratamiento B, y así sucesivamente. Luego, el modelo de regresión, $\mu_A = \beta_0 + \beta_1 X_A$, podría usarse para describir un aumento (o disminución) lineal en la presión arterial media como una función del cambio de dosis. Esto da lugar al modelo de regresión lineal estadístico

$$Y_A = \beta_0 + \beta_1 X_A + e$$

Supongamos ahora que cada fármaco (o dosis de fármaco) se aplica a varios sujetos, digamos, $n$ de ellos para cada fármaco. Además, suponga que los sujetos se asignan a cada fármaco completamente al azar. Entonces el experimento es un diseño completamente al azar. Las presiones sanguíneas se determinan para cada sujeto. Entonces $Y_{A1}$ representa la presión sanguínea observada en el primer sujeto tratado con el fármaco A. En general, $Y_{ij}$ representa la observación en el $j-$ésimo sujeto tratado con el medicamento $i$. Luego se puede escribir la ecuación del modelo $Y_{ij} = \mu + e_{ij}$, donde $e_{ij}$ es una variable aleatoria con media cero y varianza $\sigma^2$. Esto significa que las presiones arteriales de diferentes sujetos que reciben el mismo tratamiento no son todas iguales.

El error $e_{ij}$ representa esta variación. Observe que este modelo utiliza el supuesto de que la varianza de $e_{ij}$ es la misma, $\sigma^2$, para cada fármaco. Este supuesto puede o no ser válido en una cierta situación dada; los modelos más complejos permiten variaciones distintas entre las observaciones dentro de diferentes tratamientos. Además, tenga en cuenta que el modelo se puede elaborar mediante una descripción adicional de $\mu_i$ (por ejemplo, como un modelo de efectos $\mu_i = \mu + \alpha_i$ o como un modelo de regresión $\mu_i = \beta_0 + \beta_1 X_i$).

Una forma alternativa de representar los modelos anteriores los describe a través de una distribución de probabilidad asumida. Por ejemplo, el modelo estadístico lineal habitual para datos que surgen de diseños completamente aleatorios, asume que los errores tienen una distribución normal. Por lo tanto, se puede escribir el modelo $Y_{ij} = \mu_i + e_{ij}$ equivalente a $Y_{ij} \sim N(\mu_i,\ \sigma^2)$ si el $e_{ij}$ se asume iid $N(0,\ \sigma^2)$

De manera similar, el modelo ANOVA de una vía, se puede escribir como $Y_{ij} \sim N(\mu + \alpha_i,\ \sigma^2)$ y el modelo de regresión lineal como $Y_{ij} \sim N(\beta_0 + \beta_1 X_i,\ \sigma^2)$

En forma matricial:

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$$

donde

$$\varepsilon \sim MN\left(E(\boldsymbol{\varepsilon}) = \mathbf{O},\ \Sigma_{\varepsilon} = \sigma^2\mathbf{I}\right)$$

El estimador de mínimos cuadrados OLS para $\beta$ es

$$\widehat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$$

Si $\Sigma_{\varepsilon} = \sigma^2\mathbf{V}$ donde $\mathbf{V}$ se conoce.

$$\widehat{\beta} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$$

(Estimador de mínimos cuadrados generalizados)

El **modelo lineal generalizado (GLM)** es una generalización flexible de la regresión lineal ordinaria que permite variables de respuesta que tienen modelos de distribución de errores distintos de una distribución normal. El GLM generaliza la regresión lineal al permitir que el modelo lineal se relacione con la variable de respuesta a través de una **función de enlace** y al permitir que la magnitud de la varianza de cada medición sea una función de su valor predicho. El modelo de regresión clásico es un caso particular de un GLM, con respuesta normal. Esta respuesta se discute en detalle en gran parte de este curso.

Los modelos lineales generalizados fueron formulados por John Nelder y Robert Wedderburn (1972[5]) como una forma de unificar varios otros modelos estadísticos, incluida la regresión lineal, la regresión logística y la regresión de Poisson. Propusieron un método iterativo de mínimos cuadrados ponderados para la estimación de máxima verosimilitud de los parámetros del modelo. La estimación de máxima verosimilitud sigue siendo popular y es el método predeterminado en muchos paquetes de computación estadística. Se han desarrollado otros enfoques, incluidos los enfoques bayesianos y los ajustes de mínimos cuadrados para estabilizar varianza. Este tipo de modelos se discuten más adelante en el curso, desde una perspectiva adaptada a datos longitudinales.

[5] https://docs.ufpr.br/~taconeli/CE225/Artigo.pdf

Cuando la variable de respuesta es continua,los modelos de regresión lineal clásica se pueden extender para manejar resultados correlacionados que son típicos en los datos longitudinales. Esto se puede lograr a través de la introducción de efectos aleatorios en el modelo que tratan de explicar la heterogeneidad y variabilidad extra que el término de error por si solo no puede explicar. Este enfoque produce una clase versátil de modelos de regresión para datos longitudinales conocidos como modelos lineales de efectos mixtos, LMM o MLM (Fitzmaurice y Laird, 2014).

los MLM son una poderosa herramienta estadística en investigación aplicada ya que no solo está ampliamente disponible en software estándar, sino que también tiene una sólida base teórica. El término modelo mixto se refiere a la naturaleza de las partes del modelo que explica la media de un modelo estadístico.

Ya se resaltó que cuando se trabaja con datos longitudinales que estan en format ancho, estos se deben transformar a formato largo a fin de poder usar funciones relacionadas con visualización y modelamiento. El R ofrece alternativas para transformar datos longitudianles de formato ancho a formato largo. Librerías reshape y plyr

```
#EJEMPLO CLASE 2 INTROMLM. DE FORMATO ANCHO A FORMATO LARGO
library(reshape2)
library(plyr)

abeto_ancho<-read.csv(file="Sitka_Spruce_ancho.csv",
                      header=T,sep=',',dec='.')
head(abeto_ancho)

df<-melt(abeto_ancho,id.vars = c("ID", "Group"),
         value.name="log_growth")
abeto_largo<-arrange(df, ID)
head(abeto_largo)
```

## PASANDO DE FORMATO ANCHO A FORMATO LARGO.

```
#EJEMPLO CLASE 2 INTROMLM. FORMATO ANCHO A FORMATO LARGO
library(reshape2)
library(plyr)

abeto_ancho<-read.csv(file="Sitka_Spruce_ancho.csv",
                      header=T,sep=',',dec='.')
head(abeto_ancho)
```

```
##    ID   Group y152 y174 y201 y227
## 1   1 Control 2.79 3.10 3.30 3.38
## 2   2 Control 3.30 3.90 4.34 4.96
## 3   3 Control 3.98 4.36 4.79 4.99
## 4   4 Control 4.36 4.77 5.10 5.30
## 5   5 Control 4.34 4.95 5.42 5.97
## 6   6 Control 4.59 5.08 5.36 5.76
```

## PASANDO DE FORMATO ANCHO A FORMATO LARGO.

```
df<-melt(abeto_ancho,id.vars = c("ID", "Group"),
         value.name="log_growth")
abeto_largo<-arrange(df, ID)
head(abeto_largo)
```

```
##    ID   Group variable log_growth
## 1   1 Control     y152       2.79
## 2   1 Control     y174       3.10
## 3   1 Control     y201       3.30
## 4   1 Control     y227       3.38
## 5   2 Control     y152       3.30
## 6   2 Control     y174       3.90
```

## PASANDO DE FORMATO ANCHO A FORMATO LARGO.

```
df<-melt(abeto_ancho,id.vars = c("ID", "Group"),
         value.name="log_growth")
abeto_largo<-arrange(df, ID)
tail(abeto_largo)
```

```
##     ID Group variable log_growth
## 91 23 Ozone     y201        6.12
## 92 23 Ozone     y227        6.41
## 93 24 Ozone     y152        4.52
## 94 24 Ozone     y174        4.91
## 95 24 Ozone     y201        5.04
## 96 24 Ozone     y227        5.71
```

**PASANDO DE FORMATO LARGO A FORMATO ANCHO.** También se requiere algunas veces pasar de fomato largo a formato ancho.

```
#Ejemplos clase 1 INTROMLM
library(lattice)
library(ggplot2)
library(gridExtra)
library(Rcpp)
library(MASS)
library(lattice)

data<-read.csv(file="SPRUCE1.csv",header=T,sep=',',dec='.')
data_ancho<-reshape(data[ , c("ID","Group", "log_growth","time1")],
                    direction="wide", v.names="log_growth",
                    idvar="ID",timevar = "time1")
head(data_ancho)
```

## PASANDO DE FORMATO LARGO A FORMATO ANCHO.

```
##    ID   Group log_growth.152 log_growth.174 log_growth.201 log_growth.227
## 1   1 Control           2.79           3.10           3.30           3.38
## 5   2 Control           3.30           3.90           4.34           4.96
## 9   3 Control           3.98           4.36           4.79           4.99
## 13  4 Control           4.36           4.77           5.10           5.30
## 17  5 Control           4.34           4.95           5.42           5.97
## 21  6 Control           4.59           5.08           5.36           5.76
```