Introducción a los modelos mixtos (SIA 3011003)

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

El poder de la estadística. La estadística se utiliza, entre otras muchas cosas, para informar las políticas que afectan a millones de personas. Los métodos estadísticos se aplican en todos los campos que involucran la toma de decisiones, utilizando un cuerpo de datos para hacer conjeturas informadas a pesar de la falta de conocimiento definitivo.

La estadística se han convertido en una herramienta clave de las ciencias sociales y se aplica en una variedad de disciplinas que incluyen economía, psicología, ciencias políticas, sociología y antropología. Utilizando fuentes como encuestas, censos y registros administrativos, los estadísticos recopilan y analizan datos para darnos impresiones a gran y pequeña escala de nuestra compleja sociedad, construyendo una imagen más clara de dónde estamos y hacia dónde vamos[1].

[1] Traducido al español de https://esrc.ukri.org/about-us/50-years-of-esrc/50-achievements/the-power-of-statistics/

Se verá material relacionado con el modelo lineal mixto (MLM). De una manera muy intuitiva inicialmente, el MLM (o LMM por sus siglas en inglés) constituye una extensión del modelo de regresión múltiple de manera que se puedan modelar y extraer información de datos longitudinales.

En el modelo de regresíon lineal múltiple clásico, a cada sujeto se le registra una medida o respuesta de interés, que se denota y_i donde el subíndice indica que se registra una y solo una respuesta al sujeto.

y_i: Altura de un niño (covariable x_1: Género) i a una cierta edad (covariable x_2: Edad):

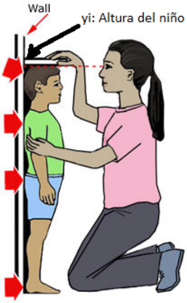

Figura 1: Medición de la altura de un niño de una cierta edad

En el modelo de regresíon lineal múltiple mixto, a cada sujeto se le registra una medida o respuesta de interés, varias veces en el tiempo, que se denota y_{ij} donde el subíndice indica que se registra dicha medida al sujeto i y el subíndice j indica que dicha medición se hizo al tiempo j.

y_{ij}: Altura de un niño i al tiempo j (niño es una covariable x_1: Género) a una cierta edad (covariable x_2: Edad):

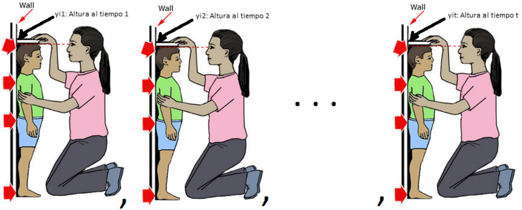

Figura 2: Medición de la altura de un niño de una cierta edad en varias ocasiones

Los datos longitudinales, a veces denominados datos de panel, usualmente, están conformados por una misma muestra en diferentes momentos en el tiempo. La muestra puede estar formada por personas, animales, plantas, hogares, establecimientos, etc. Los datos longitudinales son ahora una parte integral de los estudios experimentales y empíricos en una variedad de disciplinas, desde las ciencias médicas hasta las sociales y empresariales.

En algunos casos, los investigadores preguntan cómo maduran las personas y cómo se desarrollan a tiempo sus capacidades cognitivas; en otros, preguntamos si ocurre un evento de interés y cuándo, o cómo cambian en el tiempo medidas craneofaciales de un grupo de niños y niñas bajo estudio, entre otras situaciones interesantes.

El análisis de datos longitudinales enfrenta dos problemas principales: primero, la separación del cambio evolutivo (edad) e histórico (período) y su posible interacción, y segundo, la interdependencia entre observaciones de la misma variable para el mismo individuo en diferentes momentos.

Los datos de panel longitudinal o datos longitudinales, por definición, involucran dos o más observaciones sobre los mismos individuos (aunque en ocasiones se podría disponer de solo una observación de uno o varios sujetos). Muchas técnicas estadísticas asumen que las observaciones o mediciones de variables son mutuamente independientes, pero esto es claramente improbable cuando se mide al mismo individuo en la misma variable más de una vez, especialmente para intervalos de medición relativamente más cortos en lugar de más largos.

El análisis de datos longitudinales consiste en las herramientas y métodos estadísticos utilizados para analizar los datos recopilados sobre el mismo grupo de personas en múltiples ocasiones a lo largo del tiempo. Existen diferentes enfoques para manejar datos longitudinales. Uno de las más populares es la técnica de regresión basada en modelos lineales mixtos (MLM o LMM) (que se definen más adelante).

LMM es una poderosa herramienta estadística en la investigación aplicada, ya que no solo está ampliamente disponible en el 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.

- Una característica definitoria de los estudios longitudinales es que las mediciones de los mismos individuos se toman repetidamente a lo largo del tiempo.
- Los estudios longitudinales permiten el estudio directo del cambio a lo largo del tiempo.
- El objetivo principal, es caracterizar los cambios en la respuesta a lo largo del tiempo y los factores que influyen en el cambio.
- Con medidas repetidas, se pueden capturar cambios intraindividuales.

Ilustración del proceso de envejecimiento de una persona.

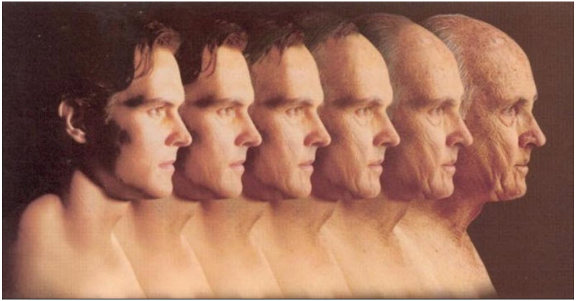

Figura 3: envejecimiento de una persona

Los datos longitudinales y los experimentos de mediciones repetidas deben tratarse con modelos que incorporan estructuras de covarianza que surge cuando la observaciones se toman sobre el mismo sujeto o unidad. Esto se puede hacer usando MLM. Pero, ¿qué es un MLM? Una posible respuesta podría proporcionarse en el siguiente ejemplo acerca de datos de crecimiento de árboles de abeto.

Figura 4: Evolución del crecimiento de un árbol

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

(Continued)

Obs	Group	Id	y_{152}	y_{174}	y_{201}	y_{227}
13	Ozone	13	4.53	5.05	5.18	5.41
14	Ozone	14	4.97	5.32	5.83	6.29
15	Ozone	15	4.37	4.81	5.03	5.19
16	Ozone	16	4.58	4.99	5.37	5.68
17	Ozone	17	4.00	4.50	4.92	5.44
18	Ozone	18	4.73	5.05	5.33	5.92
19	Ozone	19	5.15	5.63	6.11	6.39
20	Ozone	20	4.10	4.46	4.84	5.29
21	Ozone	21	4.56	5.12	5.40	5.69
22	Ozone	22	5.16	5.49	5.74	6.05
23	Ozone	23	5.46	5.79	6.12	6.41
24	Ozone	24	4.52	4.91	5.04	5.71

Observe que este conjunto de datos tiene el mismo número de árboles por grupo (12) y se tomaron 4 medidas de crecimiento en los días 152, 174, 201 y 227, respectivamente. Es decir, cada árbol aporta 4 medidas. El objetivo principal de este estudio es comparar el crecimiento de árboles sembrados en dos ambientes diferentes: control y ambiente enriquecido con ozono.

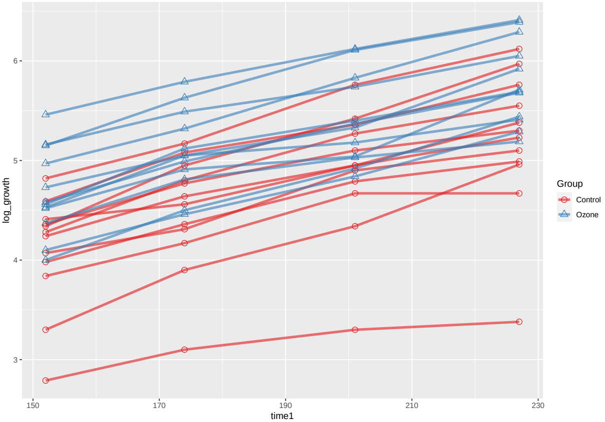

Figura 5: Perfiles ambos grupos. Abetos

MODELO LINEAL MIXTO: ESTADÍSTICOS DE RESUMEN

En este gráfico se observa un crecimiento lineal en el tiempo en ambos grupos (específicamente, en el intervalo de 152-227 días) con un intersecto aleatorio (parece que cada árbol en cada grupo tiene su propio intersecto pero las pendientes son similares). Para analizar este tipo de datos existen algunas técnicas. Una técnica muy utilizada, pero no necesariamente buena, consiste en resumir o comprimir los datos mediante estadísticos de resumen. Algunas de estas técnicas también pueden ser útiles para imputar datos faltantes.

ESTADÍSTICOS DE RESUMEN Y EL MODELO DE DOS ETAPAS (TWO STAGES MODEL)

Procedimiento general:

- Resuma los datos de cada sujeto usando alguna estadística (por ejemplo, la media muestral)
- Realice un análisis con esos estadísticos de resumen

De esta forma, el análisis de datos longitudinales, se reduce al análisis de observaciones independientes y podría realizarse utilizando métodos estadísticos clásicos.

ESTADÍSTICOS DE RESUMEN Y EL MODELO DE DOS ETAPAS (TWO STAGES MODEL)

Entre los métodos más comunes encontramos:

- Análisis de cada punto por separado. Por ejemplo, se puede usar una prueba t de Student para datos a los 152 días para comparar los entornos de control y ozono.
- Área bajo la curva AUC. A menudo, el AUC se utiliza como medida de la eficacia de un fármaco o tratamiento específico. El problema con este método es que dos curvas diferentes pueden producir áreas iguales bajo la curva.
- Última observación realizada (Last Observation Carried Forward, LOCF). Se recomienda en situaciones donde faltan datos.

ESTADÍSTICOS DE RESUMEN Y EL MODELO DE DOS ETAPAS (TWO STAGES MODEL)

Entre los métodos más comunes encontramos:

- Análisis de incrementos.
- Regresión lineal simple. Para cada sujeto estimamos una línea de regresión usando MCO y luego se comparan los intersectos y las pendientes para cada grupo.

Ventajas y desventajas de los estadísticos de resumen:

- Al usarlos se podría perder mucha información.
- No dicen nada sobre la tendencia desde el primero hasta el último punto.
- Es difícil ajustar los datos faltantes.
- Ignoran la posible correlación entre mediciones repetidas tomadas sobre el mismo sujeto.

Por estas razones, no se recomienda su uso.

ESTADÍSTICOS DE RESUMEN Y EL MODELO DE DOS ETAPAS (TWO STAGES MODEL)

Ejemplo de estadístico de resumen. Considere los datos de Sitka Spruce. ¿Qué pasa si resumimos la información de cada árbol mediante un modelo de regresión lineal?

id	grupo	intercepto	Pendiente	id	grupo	intercepto	Pendiente
1	Control	1,68505	0,007732	13	Ozono	3,00017	0,010835
2	Control	0.08301	0.021443	14	Ozono	2.25471	0.017760
3	Control	1.95227	0.013675	15	Ozono	2.86841	0.010512
4	Control	2,54210	0,012416	16	Ozono	2,41254	0,014549
5	Control	1.17541	0.021191	17	Ozono	1.17788	0.018765
6	Control	2.37889	0.014953	18	Ozono	2.37311	0.015302
7	Control	2,64562	0,011363	19	Ozono	2,69236	0,016592
8	Control	2.01944	0.014764	20	Ozono	1.71878	0.015670
9	Control	2.10608	0.017832	21	Ozono	2.47344	0.014425
10	Control	2.11265	0.011803	22	Ozono	3.43345	0.011547
11	Control	1,83654	0,014634	23	Ozono	3,57231	0,012587
12	Control	1.79106	0.016891	24	Ozono	2.28002	0.014668

Ejemplo de estadístico de resumen. Considere los datos de Sitka Spruce.

Group	Means	Std	t-test	P-value
Intercepts:				
Control	1.8607	0.1979	-2.39	0.0259 **
Ozone	2.5214	0.1933		
Slopes:				
Control	0.0149	0.00400	0.33	0.7444
Ozone	0.0144	0.00264		

De acuerdo con estos resultados, existen diferencias entre los intersectos pero no entre las pendientes. **Hay un gran problema: ignoramos la dependencia en las observaciones (las tratamos como independientes)**

Programa SAS para ajustar el análisis estadístico de resumen basado en regresión lineal:

```
data spruce;
input  ID Group$  y152   y174  y201 y227 @@;
cards;
 1     Control  2.79    3.10    3.30    3.38    2     Control  3.30    3.90    4.34    4.96
 3     Control  3.98    4.36    4.79    4.99    4     Control  4.36    4.77    5.10    5.30
 5     Control  4.34    4.95    5.42    5.97    6     Control  4.59    5.08    5.36    5.76
 7     Control  4.41    4.56    4.95    5.23    8     Control  4.24    4.64    4.95    5.38
 9     Control  4.82    5.17    5.76    6.12    10    Control  3.84    4.17    4.67    4.67
 11    Control  4.07    4.31    4.90    5.10    12    Control  4.28    4.80    5.27    5.55
 13    Ozone    4.53    5.05    5.18    5.41    14    Ozone    4.97    5.32    5.83    6.29
 15    Ozone    4.37    4.81    5.03    5.19    16    Ozone    4.58    4.99    5.37    5.68
 17    Ozone    4.00    4.50    4.92    5.44    18    Ozone    4.73    5.05    5.33    5.92
 19    Ozone    5.15    5.63    6.11    6.39    20    Ozone    4.10    4.46    4.84    5.29
 21    Ozone    4.56    5.12    5.40    5.69    22    Ozone    5.16    5.49    5.74    6.05
 23    Ozone    5.46    5.79    6.12    6.41    24    Ozone    4.52    4.91    5.04    5.71
 ;
 run;
```

Programa SAS para ajustar el análisis estadístico de resumen basado en regresión lineal:

```
data spruce1;
set spruce;
array xx(4)  y152   y174  y201 y227;
 do i=1 to 4;
 log_growth=xx{i};
 if i=1 then time1=152;
 if i=2 then time1=174;
 if i=3 then time1=201;
 if i=4 then time1=227;
 output;
end;
run;
proc sort data=spruce1;
by id;
run;
```

ESTADÍSTICOS DE RESUMEN Y EL MODELO DE DOS ETAPAS (TWO STAGES MODEL)

Programa SAS para ajustar el análisis estadístico de resumen basado en regresión lineal:

```
proc reg data=spruce1 noprint outest=one;
by id;
model log_growth=time1;
run;
data slopes;
retain id group intercept time1;
set one;
keep id group intercept time1;
if id<=12 then group='Control';else group='Ozone';
run;
proc print data=slopes noobs;
run;

proc ttest;
class group;
var intercept;
title 'T-test for intercepts';
run;
```

ESTADÍSTICOS DE RESUMEN Y EL MODELO DE DOS ETAPAS (TWO STAGES MODEL)

Sin embargo, una modificación del método de estadísticos de resumen basado en regresión lineal simple, permite aproximar bien los perfiles asociados a cada sujeto. Esta modificación se conoce como el modelo de dos etapas.

Modelo de dos etapas:

1. **Etapa 1**: Ajuste un modelo de regresión separado para cada sujeto.
2. **Etapa 2**: Explique la variabilidad de los coeficientes de regresión específicos del sujeto, usando variables conocidas (efectos fijos).

Procederemos a explicar mejor estos dos puntos.

ESTADÍSTICOS DE RESUMEN Y EL MODELO DE DOS ETAPAS (TWO STAGES MODEL)

Formulación general. **Etapa 1**. Suponga que Y_{ij} es la respuesta del i-ésimo sujeto en el momento X_{ij}, $i = 1, 2, \ldots, m$ y $j = 1, 2, \ldots, n_i$. Por lo tanto, tenemos que $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})^T$ es el vector de respuestas para el i-ésimo sujeto. Así, el modelo

$$\mathbf{Y}_i = \mathbf{Z}_i \boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i,$$

describe la variabilidad intra-sujeto, donde $\mathbf{Z}_i$ es una matriz de orden $n_i \times q$ de covariables conocidas, $\boldsymbol{\beta}_i$ es un vector q-dimensional con los coeficientes de regresión específicos del sujeto y $\boldsymbol{\epsilon}_i \sim N(0, \Sigma_i)$ aquí Σ_i es la matriz de covarianza. En resumen, la etapa 1 describe la variabilidad intra-sujeto.

Etapa 2. En esta etapa se modela la variabilidad entre sujetos. Esta variabilidad puede ser modelada, si los β_i están relacionados con variables conocidas. En otras palabras, si hacemos

$\boldsymbol{\beta}_i = \mathbf{K}_i\boldsymbol{\beta} + \mathbf{b}_i$, donde

$$
\begin{array}{ccl}
\mathbf{K}_i & : & \text{es una matriz } q \times p \text{ de covariables conocidas} \\
\boldsymbol{\beta} & : & p\text{-vector de parámetros de regresión desconocidos} \\
\mathbf{b}_i & \sim & N(0, \mathbf{D})
\end{array}
$$

Aquí $\mathbf{D}$ es una matriz de covarianza $q \times q$.

ESTADÍSTICOS DE RESUMEN Y EL MODELO DE DOS ETAPAS (TWO STAGES MODEL)

Este MLM se puede escribir de una forma más precisa como:

$$\begin{aligned}
\mathbf{Y_i} &= \mathbf{X_i}\beta + \mathbf{Z_i}\mathbf{b_i} + \epsilon_i \quad (\mathbf{Z_i}\mathbf{K}_i = \mathbf{X_i}) \\
\mathbf{b_i} &\sim N(\mathbf{0}, \mathbf{D}) \\
\epsilon_i &\sim N(\mathbf{0}, \Sigma_i) \\
\mathbf{b}_1, \ldots, \mathbf{b}_m, \epsilon_1, \ldots, \epsilon_m &: \text{mutuamente independientes}
\end{aligned}$$

Terminología:

- β: *Efectos Fijos*.
- $\mathbf{b}_i$ and ϵ_i: *Efectos aleatorios*.
- *Componentes de Varianza*: Elementos de las matrices $\mathbf{D}$ y $\mathbf{\Sigma}_i$

Note que la matriz de varianzas y covarianzas del vector aleatorio $\mathbf{Y_i}$ es

$$Cov(\mathbf{Y_i}) \;=\; \mathbf{Z_i D Z_i}' + \Sigma_i$$

(Ejercicio)

Forma matricial general del MLM. Sea

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y_1} \\ \mathbf{Y_2} \\ \vdots \\ \mathbf{Y_m} \end{pmatrix}$$

(donde el vector para el sujeto i, $\mathbf{Y_i}$, tiene n_i componentes)

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_m \end{pmatrix}$$

$$\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \end{pmatrix}$$

ESTADÍSTICOS DE RESUMEN Y EL MODELO DE DOS ETAPAS (TWO STAGES MODEL)

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \\ \vdots \\ \mathbf{Z}_m \end{pmatrix}$$

$$\mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_m \end{pmatrix}$$

Así, la forma matricial general del MLM está definida como[2]:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{b} + \epsilon$$

donde $E(\epsilon) = \mathbf{0}$ y

$$\mathbf{\Sigma}_\epsilon = \begin{pmatrix} \mathbf{\Sigma}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{\Sigma}_2 & \cdots & 0 \\ \vdots & \ldots & \ldots & \vdots \\ 0 & 0 & \cdots & \mathbf{\Sigma}_m \end{pmatrix}$$

[2]Comparela con la forma matricial del MLG:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

donde $E(\epsilon) = \mathbf{0}$ y $\mathbf{\Sigma}_\epsilon = \sigma^2\mathbf{I}$

$$\mathbf{\Sigma_b} = \begin{pmatrix} \mathbf{D} & 0 & \cdots & 0 \\ 0 & \mathbf{D} & \cdots & 0 \\ \vdots & \ldots & \ldots & \vdots \\ 0 & 0 & \cdots & \mathbf{D} \end{pmatrix} = \mathbf{I}_m \otimes \mathbf{D}$$

($\otimes$ **Producto Kronecker**)

Ejemplo: datos de Sitka Spruce Como acabamos de ver, para cada grupo de árboles, un modelo razonable incluye intersectos aleatorios y una pendiente común. Entonces, si queremos utilizar el método de las dos etapas, en la primera etapa se debe formular el siguiente modelo:

$$\mathbf{Y}_i = \mathbf{Z}_i \boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i$$

donde

$$\mathbf{Z}_i = \begin{pmatrix} 1 & 152 \\ 1 & 174 \\ 1 & 201 \\ 1 & 227 \end{pmatrix}$$

y

$$\boldsymbol{\beta}_i^T = \begin{pmatrix} \beta_{1i}, & \beta_{2i} \end{pmatrix}$$

Aquí, β_{1i} es el intersecto desconocido para el sujeto i y β_{2i} es la pendiente desconocida para el sujeto i

En la segunda etapa, se relacionan los intersectos sujeto-específicos y el efecto del tiempo con los tratamientos ambientales (*Control* u *Ozono*) mediante:

$$\beta_{1i} = \beta_0 + b_{1i}$$

donde b_{1i} es un efecto aleatorio que modifica el intercepto y

$$\beta_{2i} = \beta_1 C_i + \beta_2 O_i$$

(efectos fijos por tratamiento). Además,

$$C_i = \begin{cases} 1 & \text{Si } \textit{Control} \\ 0 & \text{En otro caso} \end{cases}$$

y

$$O_i = \begin{cases} 1 & \text{Si } \textit{Ozone} \\ 0 & \text{En otro caso} \end{cases}$$

C_i y O_i son variables indicadoras.

ESTADÍSTICOS DE RESUMEN Y EL MODELO DE DOS ETAPAS (TWO STAGES MODEL)

La interpretación de los parámetros es la siguiente:

1. β_0: La respuesta media al comienzo del tratamiento.
2. β_1 y β_2: efectos medios del tiempo para los grupos *Control* y *Ozono*, respectivamente.

ESTADÍSTICOS DE RESUMEN Y EL MODELO DE DOS ETAPAS (TWO STAGES MODEL)

De acuerdo con la forma en que se formulan los modelos, las elecciones apropiadas para $\boldsymbol{\beta}$, $\mathbf{K}_i$ y $\mathbf{b}_i$ son:

$$\boldsymbol{\beta}^T = \begin{pmatrix} \beta_0, & \beta_1, & \beta_2 \end{pmatrix}$$

$$\mathbf{K}_i = \begin{pmatrix} 1 & 0 & 0 \\ 0 & C_i & O_i \end{pmatrix}$$

$$\mathbf{b}_i^T = \begin{pmatrix} \beta_i, & 0 \end{pmatrix}$$

ESTADÍSTICOS DE RESUMEN Y EL MODELO DE DOS ETAPAS (TWO STAGES MODEL)

En el análisis se puede realizar explícitamente una aproximación en dos etapas. Sin embargo, este es solo otro ejemplo del uso de estadísticos de resumen, ya que $\mathbf{Y_i}$ se resume mediante $\boldsymbol{\beta}_i$ que es un estadístico de resumen que se analiza en la segunda etapa. Los problemas asociados con esta estrategia de estadísticos de resumen se pueden evitar combinando las dos etapas:

$$\begin{aligned} \mathbf{Y_i} &= \mathbf{Z_i}\beta_i + \epsilon_i \\ \beta_i &= \mathbf{K_i}\beta + \mathbf{b_i} \end{aligned}$$

en un solo modelo:

$$\begin{aligned} \mathbf{Y_i} &= \mathbf{Z_i}\mathbf{K}_i\beta + \mathbf{Z_i}\mathbf{b_i} + \epsilon_i \\ &= \mathbf{X_i}\beta + \mathbf{Z_i}\mathbf{b_i} + \epsilon_i \end{aligned}$$

Example: Sitka Spruce data (Continuación) En este ejemplo, la formulación del MLM es:

$$y_{ij} = \beta_o + \beta_1 C_i x_j + \beta_2 O_i x_j + b_i + \epsilon_{ij}$$

y, puesto que $C_i = 1 - O_i$, se tiene que

$$y_{ij} = \beta_o + \beta_1 x_j + (\beta_2 - \beta_1) O_i x_j + b_i + \epsilon_{ij}$$

donde $b_i \sim N(0, \sigma^2)$ y $\boldsymbol{\epsilon}_i \sim N(0,\ \boldsymbol{\Sigma}_i)$ (mutuamente independientes). Este modelo se conoce como Modelo de interceptos aleatorios.