Medidas Repetidas

Este es un problema que puede ser bastante complejo y es importante distinguir el diseño con el que estemos trabajando para tratarlo adecuadamente.

Diferentes modelos surgen de acuerdo a los distintos supuestos y perspectivas que se pueden realizar llevando a distintas preguntas posibles, caracterizaciones y por ende a distintos métodos para tratar cada caso.

Algunos de estos modelos reciben distintos nombres según la bibliografía: modelos longitudinales, efectos aleatorios, medidas repetidas, etc. y son casos particulares del modelo lineal con efectos mixtos. La terminología no es universal.

Vamos a ver un ejemplo muy sencillo usando un modelo lineal y combinándolo con medidas repetidas y efectos aleatorios.

Toy Example

Las medidas repetidas se refieren a que a un individuo o unidad experimental se lo mide varias veces.

Supongamos que en un estudio participan 3 individuos y que a cada participante se le realizan 4 preguntas.

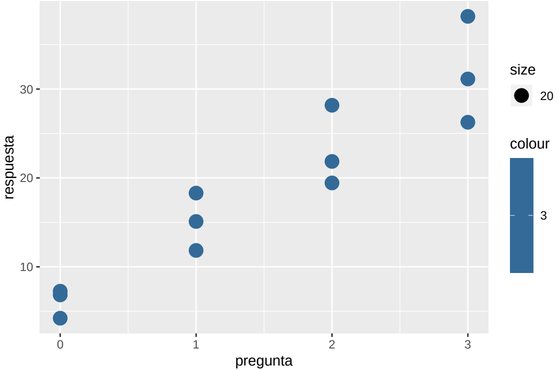
1. Asumimos que cada participante es elegido al azar de una población.
2. Cada participante responde las preguntas de acuerdo a su puntaje, que seguramente tiene una escala personal.

Toy Example

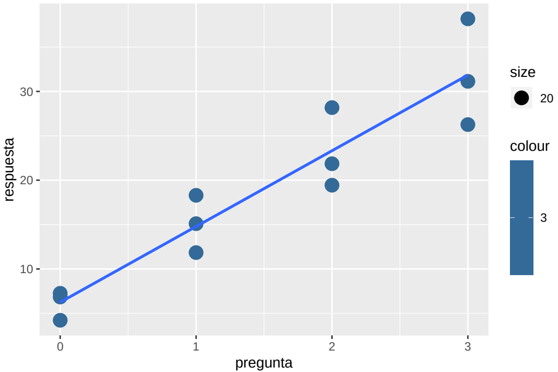
```
> setwd("C:\\Users\\Ana\\Dropbox\\Ana\\GLM\\2019\\Doctex")
> ejemplo<- read.table("toyexample.txt",header=T)
> head(ejemplo)
```

```
  participante edad pregunta respuesta
1           p1   32        0  6.840733
2           p1   32        1 15.103398
3           p1   32        2 21.854398
4           p1   32        3 31.133745
5           p2   38        0  7.258445
6           p2   38        1 18.296092
```

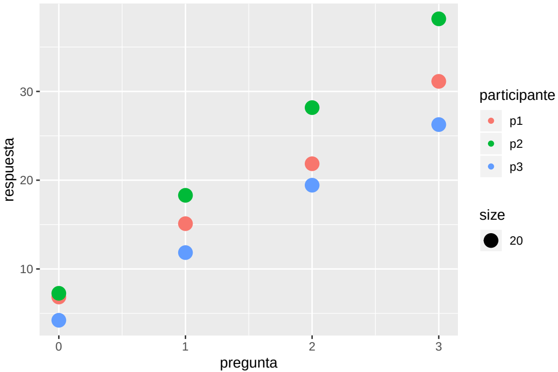
Puntos



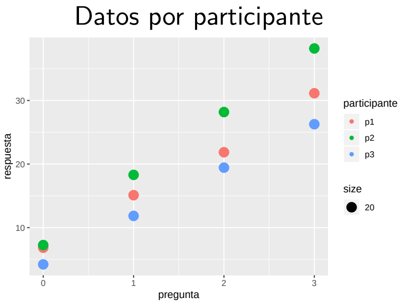
Ajuste Común: olvidando las repeticiones!!!

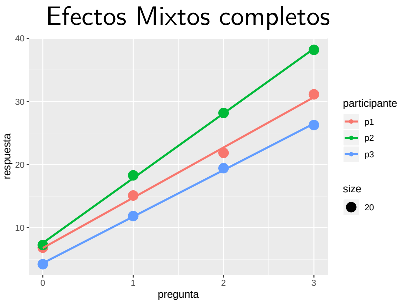


Distinguiendo a los participantes...

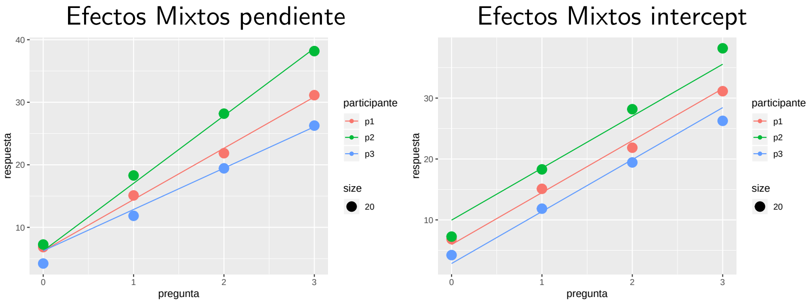


Distinguiendo a los participantes...





Distinguiendo a los participantes...



Modelo Simple: inapropiado

No se considera ningún agrupamineto en los datos:

$$\begin{aligned} r_{si} &= \beta_0 + \beta_1 q_i + e_{si} \\ e_{si} &\sim N\left(0, \sigma^2\right) \end{aligned}$$

En general los datos con estructura como los que tenemos suelen tener para cada individuo una latencia que rompe con el supuesto de independencia.

Modelo con *offset*

Podemos expandir el modelo teniendo en cuenta esto e incorporando un *offset* para cada individuo:

$$
\begin{aligned}
r_{si} &= \beta_0 + S_{0s} + \beta_1 q_i + e_{si} \\
S_{0s} &\sim N\left(0, \tau_{00}^2\right) \\
e_{si} &\sim N\left(0, \sigma^2\right)
\end{aligned}
$$

β_0 y β_1: efectos fijos, se asumen constantes de un experimento a otro

S_{0s}: efectos aleatorios, en otro experimento tendríamos otra muestra de sujetos y por lo tanto otra realización de S_{0s}.

En este caso particular: **intercepts aleatorias**.

τ_{00}^2: parámetro del efecto aleatorio

Modelo Mixto con intercept y pendiente aleatorias

Podemos expandir aún más el modelo de manera de permitir que cada participante tenga su pendiente y su intercept:

$$
\begin{aligned}
r_{si} &= \beta_0 + S_{0s} + (\beta_1 + S_{1s})\, q_i + e_{si} \\
(S_{0s}, S_{1s}) &\sim N\left(0, \begin{bmatrix} \tau_{00}^2 & \rho\tau_{00}\tau_{11} \\ \rho\tau_{00}\tau_{11} & \tau_{11}^2 \end{bmatrix}\right) \\
e_{si} &\sim N\left(0, \sigma^2\right)
\end{aligned}
$$

Modelo Mixto con intercept y pendiente aleatorias

Podemos expandir aún más el modelo de manera de permitir que cada participante tenga su pendiente y su intercept:

$$
\begin{aligned}
r_{si} &= \beta_0 + S_{0s} + (\beta_1 + S_{1s})\, q_i + e_{si} \\
(S_{0s}, S_{1s}) &\sim N\left(0, \begin{bmatrix} \tau_{00}^2 & \rho\tau_{00}\tau_{11} \\ \rho\tau_{00}\tau_{11} & \tau_{11}^2 \end{bmatrix}\right) \\
e_{si} &\sim N\left(0, \sigma^2\right)
\end{aligned}
$$

¿Cómo sería el modelo para el modelo mixto con pendiente aleatoria?

Toy Example

```
> attach(ejemplo)
> library(lme4)
> model_in <- lmer(respuesta ~ pregunta + (1 | participante), data=ejemplo)
> summary(model_in)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: respuesta ~ pregunta + (1 | participante)
   Data: ejemplo

REML criterion at convergence: 50.8

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.53166 -0.36381  0.07729  0.54039  1.48095

Random effects:
 Groups       Name        Variance Std.Dev.
 participante (Intercept) 13.491   3.673
 Residual                  3.118   1.766
Number of obs: 12, groups:  participante, 3

Fixed effects:
            Estimate Std. Error t value
(Intercept)   6.2510     2.2857   2.735
pregunta      8.5330     0.4559  18.717

Correlation of Fixed Effects:
         (Intr)
pregunta -0.299
```

Toy Example

```
> model_sl <- lmer(respuesta ~ pregunta + (pregunta - 1 | participante), data=ejemplo)
> summary(model_sl)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: respuesta ~ pregunta + (pregunta - 1 | participante)
   Data: ejemplo

REML criterion at convergence: 43.3

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.8598 -0.4903  0.2170  0.5545  1.1481

Random effects:
 Groups       Name     Variance Std.Dev.
 participante pregunta 4.542    2.131
 Residual              1.192    1.092
Number of obs: 12, groups:  participante, 3

Fixed effects:
            Estimate Std. Error t value
(Intercept)   6.2510     0.5274  11.852
pregunta      8.5330     1.2624   6.759

Correlation of Fixed Effects:
         (Intr)
pregunta -0.179
```

Toy Example

```
> model_insl <- lmer(respuesta ~ pregunta + (1 + pregunta | participante), data=ejemplo)
> summary(model_insl)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: respuesta ~ pregunta + (1 + pregunta | participante)
   Data: ejemplo

REML criterion at convergence: 33.2

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.66688 -0.40683  0.06808  0.60233  0.98101

Random effects:
 Groups       Name        Variance Std.Dev. Corr
 participante (Intercept) 2.5870   1.608
              pregunta    2.2788   1.510    0.89
 Residual                 0.2683   0.518
Number of obs: 12, groups:  participante, 3

Fixed effects:
            Estimate Std. Error t value
(Intercept)   6.2510     0.9617   6.500
pregunta      8.5330     0.8818   9.677

Correlation of Fixed Effects:
         (Intr)
pregunta 0.818
```

Toy Example

```
> anova(model_in, model_sl, model_insl)


Data: ejemplo
Models:
model_in: respuesta ~ pregunta + (1 | participante)
model_sl: respuesta ~ pregunta + (pregunta - 1 | participante)
model_insl: respuesta ~ pregunta + (1 + pregunta | participante)
           Df    AIC    BIC  logLik deviance   Chisq Chi Df Pr(>Chisq)
model_in    4 62.146 64.085 -27.073   54.146
model_sl    4 53.874 55.813 -22.937   45.874  8.2719      0  < 2.2e-16 ***
model_insl  6 46.997 49.906 -17.498   34.997 10.8773      2   0.004345 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Penalización en Regresión Logística

Cuando el número de covariables (p) es grande relativamente respecto del tamaño muestral (n) pueden presentarse algunos problemas como los siguientes:

1. los estimadores de los coeficientes pueden tener un incremento en la varianza.
2. Tiende a haber sobreajuste.
3. Si $p > n$ (por ejemplo en microarrays), el EMV no existe.

¿Qué podemos hacer?

Cuando la relación p/n es grande,
una estrategia es apostar a un modelo ralo o esparso
esto es asumir que solo unas pocas k covariables son relevantes

Betting on sparcity!

Penalización en Regresión Logística

El EMV minimiza la deviance:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} d\left(y_i, \mathbf{x}_i^t\boldsymbol{\beta}\right)$$

Cuando la relación p/n es grande, una forma popular de reducir el efecto del sobreajuste y la variabilidad es agregando un término de penalización:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} d\left(y_i, \mathbf{x}_i^t\boldsymbol{\beta}\right) + I_\lambda(\boldsymbol{\beta})$$

donde I_λ es una función no negativa que depende de un vector de parámetros de ajuste.

Esencialmente, I_λ restringe los valores de los estimadores, de manera que estos se muevan en un rango más aceptable.

Penalizaciones

Existen distintas opciones para I_λ. Entre las más usadas figuran:

- Penalización Ridge o ℓ_2: $I_\lambda(\boldsymbol{\beta}) = (\lambda/2)\|\boldsymbol{\beta}\|_2^2 = (\lambda/2)\sum_{i=1}^{p} \beta_i^2$

- Penalización LASSO o ℓ_1: $I_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1 = \lambda\sum_{i=1}^{p} |\beta_i|$

- Penalización Elastic Net: $I_\lambda(\boldsymbol{\beta}) = \lambda\left\{\alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)/2\|\boldsymbol{\beta}\|_2^2\right\}$

donde $(\lambda, \alpha) \in \mathbb{R}_{\geq 0} \times [0, 1]$

Observaciones

- λ controla el impacto del término de regularización. Si $\lambda = 0$, el estimador coincide con el EMV, es decir la penalización no tiene efecto. Si $\lambda \to \infty$ el impacto de $I_\lambda(\boldsymbol{\beta})$ aumenta, forzando a los coeficientes a ser cada vez más pequeños.
- La selección de λ es en ese sentido crítica y se realiza por convalidación cruzada.
- No se penaliza la intercept.
- Es importante que cada una de las vaiables esté estandarizada para que todas tengan promedio 0 y la misma escala 1.
- Penalización Ridge reduce el sobreajuste y es adecuada cuando hay colinealidad entre las covariables, pero no *selecciona variables*, es decir con alta probabilidad las estimaciones de todas las coordenadas son no nulas.
- Penalización LASSO sí selecciona variables con alta probabilidad.
- Penalización Elastic Net: permite realizar selección de variables si $\alpha > 0$ y arroja mejores resultados que la penalización Lasso cuando hay un alto grado de colinealidad entre las variables.

Penalización en Regresión Logística

Veamos en R algunos ejemplos