Datos correlacionados

Outline

Motivación

- En varios estudios se observa la variable respuesta para cada **unidad de investigación** mas de una vez
  - Se mide una variable respuesta varias veces en una misma persona
  - Se mide una variable respuesta anualmente en regiones seleccionadas al azar (Garcia et al., 2012)
- Medidas tomadas en una misma unidad de investigación se denominan **cluster** y tienden a estar mas correlacionadas entre si que con el resto de datos

Capacidad pulmonar

- Estudio longitudinal en niños y adolecentes en seis ciudades de EEUU
- Niños en edades de 6-7 años fueron enrolados y observados una vez al año
- El objetivo del estudio era caracterizar los cambios en su función pulmonar
    - FEV: Volumen espiratorio forzado definido como el volumen espirado (litros) por segundo.

Capacidad pulmonar

```
> head(fev1,n=20)
   id   ht     age baseht baseage logfev1     loght    logbht
1   1 1.20  9.3415   1.20  9.3415 0.21511 0.1823216 0.1823216
2   1 1.28 10.3929   1.20  9.3415 0.37156 0.2468601 0.1823216
3   1 1.33 11.4524   1.20  9.3415 0.48858 0.2851790 0.1823216
4   1 1.42 12.4600   1.20  9.3415 0.75142 0.3506568 0.1823216
5   1 1.48 13.4182   1.20  9.3415 0.83291 0.3920421 0.1823216
6   1 1.50 15.4743   1.20  9.3415 0.89200 0.4054651 0.1823216
7   1 1.52 16.3723   1.20  9.3415 0.87129 0.4187103 0.1823216
8   2 1.13  6.5873   1.13  6.5873 0.30748 0.1222176 0.1222176
9   2 1.19  7.6496   1.13  6.5873 0.35066 0.1739534 0.1222176
10  2 1.49 12.7392   1.13  6.5873 0.75612 0.3987761 0.1222176
11  2 1.53 13.7741   1.13  6.5873 0.86710 0.4252677 0.1222176
12  2 1.55 14.6940   1.13  6.5873 1.04732 0.4382549 0.1222176
13  2 1.56 15.8220   1.13  6.5873 1.15373 0.4446858 0.1222176
14  2 1.57 16.6680   1.13  6.5873 0.92426 0.4510757 0.1222176
15  2 1.57 17.6318   1.13  6.5873 1.13462 0.4510757 0.1222176
16  3 1.18  6.9131   1.18  6.9131 0.43178 0.1655144 0.1655144
17  3 1.23  7.9754   1.18  6.9131 0.38526 0.2070142 0.1655144
18  3 1.30  8.9665   1.18  6.9131 0.59884 0.2623642 0.1655144
19  3 1.35  9.9877   1.18  6.9131 0.75142 0.3001046 0.1655144
20  3 1.47 11.0773   1.18  6.9131 0.96698 0.3852624 0.1655144
```

- **ht**: Altura (pulgadas)
- **age**: Edad (años)
- **baseht**: Altura en el enrolamiento

Capacidad pulmonar

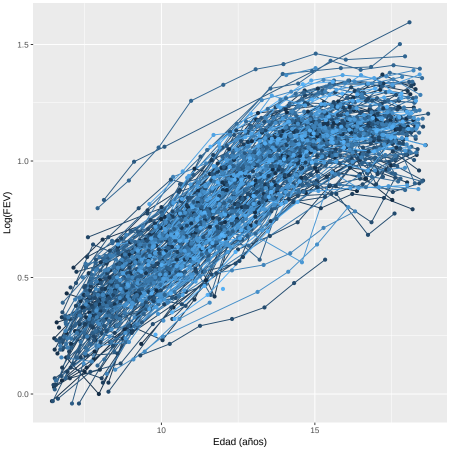


Figura 1: Evolución de capacidad pulmonar vs. edad

Capacidad pulmonar

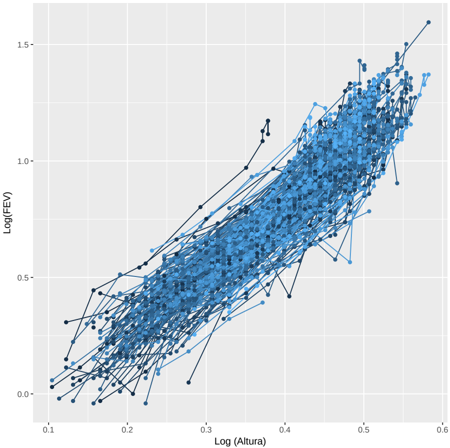


Figura 2: Evolución de capacidad pulmonar vs. altura

Capacidad pulmonar

- Si asumimos que cada observación vienen de una persona distinta: Regresión lineal
- Codigo de R

```
> model0 <- lm(logfev1 ~ age + loght + baseage + logbht,fev1)
> summary(model0)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.332894   0.020712 -16.073  < 2e-16 ***
age          0.028683   0.002077  13.810  < 2e-16 ***
loght        2.043430   0.068801  29.700  < 2e-16 ***
baseage     -0.014981   0.003960  -3.783 0.000159 ***
logbht       0.392232   0.082635   4.747 2.22e-06 ***

Residual standard error: 0.1141 on 1988 degrees of freedom
Multiple R-squared:  0.8799,    Adjusted R-squared:  0.8797
F-statistic:  3641 on 4 and 1988 DF,  p-value: < 2.2e-16
```

- Interpretación:
  - Por cada año adicional en la edad esperamos un aumento de 0.03 en capacidad pulmonar ($pvalor < 10^{-16}$)

Capacidad pulmonar

- El modelo asume que tenemos 1989 personas cuando en realidad tenemos solo 299

```
> length(unique(fev1$id))
[1] 299
```

- A mayor cantidad de datos, mayor poder a detectar diferencias
- ¿Cómo afecta el hecho que inflamos los datos?
    - Estimación
    - Incertidumbre

Estudio: Depresión

- Estudio de 340 personas afectadas por depresión (Agresti, 2007)
- Sujetos fueron aleatorizados a un nuevo tratamiento o uno estandar
- Se evaluo los niveles de depresión despues de una, dos y cuatro semanas de suministrado el tratamiento
- Respueta: Comportamiento normal (1) y anormal (0)

```
> head(depre)
  case severity treat time outcome
1    1        0     0    0       1
2    1        0     0    1       1
3    1        0     0    2       1
4    2        0     0    0       1
5    2        0     0    1       1
6    2        0     0    2       1
```

Estudio: Depresión

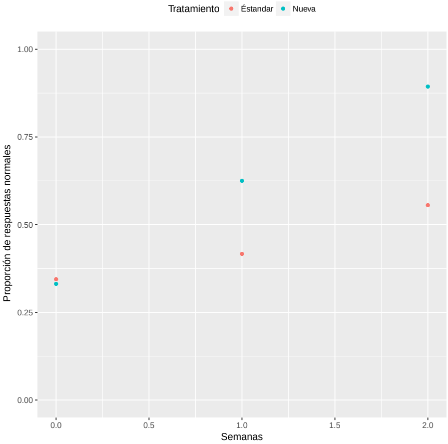


Figura 3: Proporción de respuesta favorable al tratamiento

Estudio: Depresión

- **Modelo A**

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 Severidad + \beta_2 Trt + \beta_3 Tiempo$$

  - Asume que la tendencia lineal en el tiempo es la misma para cada grupo
  - Me genera 12 probabilidades marginales $(2 \times 2 \times 3 = 12)$

- **Modelo B**

$$\begin{aligned}\log\left(\frac{\pi}{1-\pi}\right) &= \beta_0 + \beta_1 Severidad + \beta_2 Trt + \beta_3 Tiempo \\ &+ \beta_4 Trt \times Tiempo\end{aligned}$$

  - Asume que la tendencia lineal en el tiempo es diferente por tratamiento

Estudio Depresión: Modelo A

```
> model1 = glm(outcome~severity+treat+time,data=depre,family=binomial(link="logit"))
> summary(model1)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.4815     0.1463  -3.292 0.000996 ***
severity     -1.2867     0.1452  -8.859  < 2e-16 ***
treat         0.8833     0.1424   6.202 5.56e-10 ***
time          0.9013     0.0901  10.003  < 2e-16 ***

    Null deviance: 1411.9  on 1019  degrees of freedom
Residual deviance: 1192.7  on 1016  degrees of freedom
AIC: 1200.7
```

- El odds de tener una respuesta normal fue 2.4 veces ($p - value < 0{,}001$) en aquellos que recibieron la intervención en comparación con aquellos que no la recibieron.
- Controlando por el efecto del tratamiento, pacientes tuvieron una tendencia a tener mejor respuesta con el tiempo (OR$=$ $2{,}46$)

Estudio Depresión: Modelo A

```
> depre$pred = predict.glm(model1,new.data=depre,type="response")
>
> depre %>%
+   group_by(severity,treat,time)  %>%
+   summarize(oprob = mean(outcome,na.rm=T),
+             prpob = mean(pred,na.rm=T))
Source: local data frame [12 x 5]
Groups: severity, treat [?]

   severity treat  time     oprob     prpob
      <int> <int> <int>     <dbl>     <dbl>
1         0     0     0 0.5125000 0.3818970
2         0     0     1 0.5875000 0.6034246
3         0     0     2 0.6750000 0.7893504
4         0     1     0 0.5285714 0.5991119
5         0     1     1 0.7857143 0.7863437
6         0     1     2 0.9714286 0.9006336
7         1     0     0 0.2100000 0.1457719
8         1     0     1 0.2800000 0.2959006
9         1     0     2 0.4600000 0.5085900
10        1     1     0 0.1777778 0.2921665
11        1     1     1 0.5000000 0.5040934
12        1     1     2 0.8333333 0.7145597
```

Estudio Depresión: Modelo B

- Modelo

```
> model2 = glm(outcome~severity+treat*time,data=depre,family=binomial(link="logit"
> summary(model2)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.02799    0.16391  -0.171    0.864
severity    -1.31391    0.14641  -8.974  < 2e-16 ***
treat       -0.05960    0.22221  -0.268    0.789
time         0.48241    0.11476   4.204 2.63e-05 ***
treat:time   1.01744    0.18879   5.389 7.08e-08 ***
AIC: 1171.9
```

- A la semana el efecto es $e^{-0,05} = 0,95$ (o sea nulo)
- A las dos semanas el efecto del tratamiento es

$$e^{-0,05+1,02} = 2,64$$

- A las cuatro semanas el efecto del tratamiento es

$$e^{-0,05+2\times 1,02} = 7,32$$

Depresión: Modelo B vs. Modelo A

```
> depre$pred2 = predict.glm(model2,new.data=depre,type="response")
>
> depre  %>%
+   group_by(severity,treat,time)   %>%
+   summarize(oprob = mean(outcome,na.rm=T),
+             prpob1 = mean(pred1,na.rm=T),
+             prpob2 = mean(pred2,na.rm=T))
Source: local data frame [12 x 6]
Groups: severity, treat [?]

   severity treat  time     oprob    prpob1    prpob2
      <int> <int> <int>     <dbl>     <dbl>     <dbl>
1         0     0     0 0.5125000 0.3818970 0.4930033
2         0     0     1 0.5875000 0.6034246 0.6116905
3         0     0     2 0.6750000 0.7893504 0.7184601
4         0     1     0 0.5285714 0.5991119 0.4781159
5         0     1     1 0.7857143 0.7863437 0.8041229
6         0     1     2 0.9714286 0.9006336 0.9484424
7         1     0     0 0.2100000 0.1457719 0.2071979
8         1     0     1 0.2800000 0.2959006 0.2974465
9         1     0     2 0.4600000 0.5085900 0.4068325
10        1     1     0 0.1777778 0.2921665 0.1975777
11        1     1     1 0.5000000 0.5040934 0.5245687
12        1     1     2 0.8333333 0.7145597 0.8317682
```

- **Interpretación**: El modelo B presenta una mejor predicción de la probabilidad de una respuesta normal.

Modelo

- Para la persona *ith* se mide un vector de respuestas

$$\underline{Y}_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix}$$

  donde Y_{ij} es la respuesta medida para la persona i-ésima en el tiempo j-ésimo

- Para cada Y_{ij} se miden un conjunto de p covariables (que podrian cambiar o no en el tiempo)

$$\underline{X}_{ij} = \begin{pmatrix} X_{ij1} \\ X_{ij2} \\ \vdots \\ X_{ijp} \end{pmatrix}$$

- En general, tenemos las siguientes medidas de covariables de la persona i-ésima

$$X_i = \begin{pmatrix} X_{i1}^t \\ X_{i2}^t \\ \vdots \\ X_{in_i}^t \end{pmatrix} = \begin{pmatrix} X_{i11} & X_{i12} & \dots & X_{i1p} \\ \vdots & & & \\ X_{in_i1} & X_{in_i2} & \dots & X_{in_ip} \end{pmatrix}$$

Modelo

- Supuestos
  - Un subconjunto de β varia aleatoriamente de un individuo a otro
  - Individuos tienen su propia trayectoria en el tiempo
- Modelo

$$Y_{ij} = X_{ij}^t\beta + b_i + \epsilon_{ij}$$

donde

  - β: efectos que no varian de persona a persona (**Efectos fijos**)
  - b_i: Variabilidad asociada al individuo i-ésimo (**Efecto aleatorio**)
  - ϵ_{ij}: Error de muestreo de las observaciones

Modelo

- Supuestos
  - $b_i \sim N(0, \sigma_b^2)$
  - $\epsilon_{ij} \sim_{iid} N(0, \sigma^2)$

$$\epsilon_i = Normal\ \ Multivariada_{n_i}(0, \sigma^2 I_{n_i})$$

  - b_i y ϵ_{ij} son independientes

- Media condicional: Trayectoria media de un individuo

$$E[Y_{ij} \mid b_i] = X_{ij}^t \beta + b_i$$

- Media marginal: Respuesta media de la población

$$E[Y_{ij}] = X_{ij}^t \beta$$

Modelo

- Modelo

$$
\begin{aligned}
Y_{ij} &= X_{ij}^t\beta + b_i + \epsilon_{ij} \\
&= (\beta_1 + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp}) + b_i + \epsilon_{ij} \\
&= (\beta_1 + b_i) + (\beta_2 X_{ij2} + \cdots + \beta_p X_{ijp}) + \epsilon_{ij}
\end{aligned}
$$

donde

- β_1: Intercepto poblacional
- b_i: Desviación del individuo i-ésimo del intercepto poblacional

Modelo

- Covarianza Marginal

$$\begin{aligned} Var(Y_{ij}) &= Var(X_{ij}^t\beta + b_i + \epsilon_{ij}) \\ &= Var(b_i + \epsilon_{ij}) \\ &= Var(b_i) + Var(\epsilon_{ij}) \\ &= \sigma_b^2 + \sigma^2 \end{aligned}$$

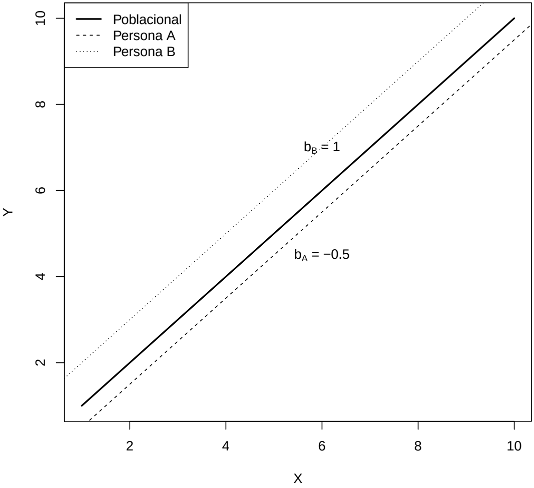
- Convarianza entre observaciones del individuo i-ésimo

$$Cov(Y_{ij}, Y_{ik}) = \sigma_b^2$$

- Correlación entre observaciones del individuo i-ésimo

$$Cor(Y_{ij}, Y_{ik}) = \frac{Cov(Y_{ij}, Y_{ik})}{\sqrt{Var(Y_{ij})}\sqrt{Var(Y_{ik})}} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2}$$

Modelo lineal con intercepto aleatorio



Estudio: Capacidad pulmonar

- Codigo de R

```
> fev1$loght  <- log(fev1$ht)
> fev1$logbht <- log(fev1$baseht)
> model1<- lme(logfev1 ~ age + loght + baseage + logbht,
+              random= ~ 1 | id,fev1)
> summary(model1)
Linear mixed-effects model fit by REML
 Random effects:
 Formula: ~1 | id
        (Intercept)   Residual
StdDev:  0.09551044 0.06428759

Fixed effects: logfev1 ~ age + loght + baseage + logbht
                 Value  Std.Error   DF  t-value p-value
(Intercept) -0.2981618 0.03919683 1692 -7.60678  0.0000
age          0.0243361 0.00129469 1692 18.79679  0.0000
loght        2.1964847 0.04327137 1692 50.76069  0.0000
baseage     -0.0172798 0.00754613  296 -2.28989  0.0227
logbht       0.3078644 0.14681355  296  2.09698  0.0368

Number of Observations: 1993
Number of Groups: 299
```

- Estimaciones
  - $\beta = (-0{,}30, 0{,}02, 2{,}20, -0{,}02, 0{,}31)$
  - $\sigma_b = 0{,}096$ y $\sigma = 0{,}06$

Estudio: Capacidad pulmonar

- Media condicional de una persona i-ésima

$$\begin{aligned} E[Y_{i.} \mid b_i] &= -0{,}3 + b_i + 0{,}02 Edad + 2{,}20 \log(Talla) \\ &\quad -0{,}02 Edad_0 + 0{,}31 \log(Talla_0) \end{aligned}$$

donde $b_i \sim N(0, 0{,}01)$

- Media poblacional

$$\begin{aligned} E[Y_{i.}] &= -0{,}3 + 0{,}02 Edad + 2{,}20 \log(Talla) \\ &\quad -0{,}02 Edad_0 + 0{,}31 \log(Talla_0) \end{aligned}$$

Estudio: Capacidad Pulmonar

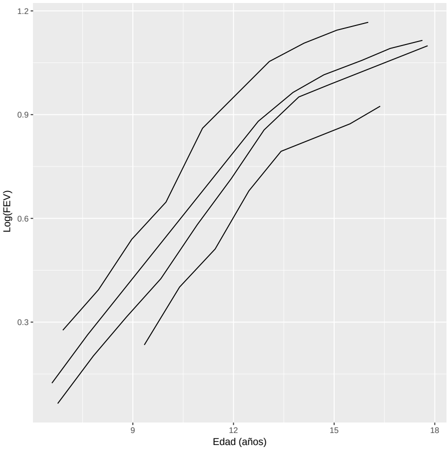


Figura 4: Capacidad pulmonar por edad

Estudio: Capacidad Pulmonar

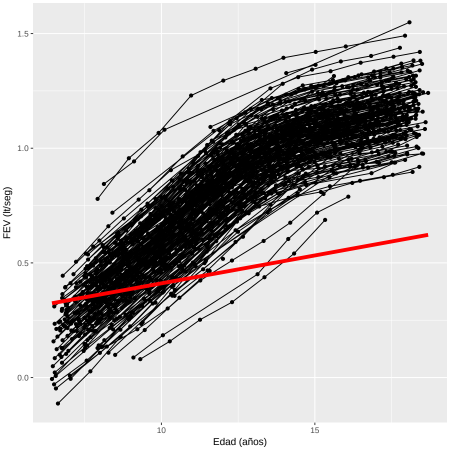


Figura 5: Capacidad pulmonar por edad

Modelo

- Supongamos que el efecto de la edad tambien es un valor aleatorio (depende del individuo)
- Modelo
$$Y_{ij} = X_{ij}^t \beta + b_{1i} + b_{2i} Z_{ij} + \epsilon_{ij}$$
donde Z puede o no estar como un efecto fijo.
- Componentes
  - Fijos: β
  - Aleatorios: b_{1i}, b_{2i} y ϵ_{ij}

Modelo

- Supuesto 1: ϵ es independiente de (b_{1i}, b_{2i})
- Supuesto 2: Distribuciones

$$\begin{pmatrix} b_{1i} \\ b_{2i} \end{pmatrix} = Normal_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12}^2 & \sigma_2^2 \end{pmatrix} \right)$$

y $\epsilon_{ij} \sim N(0, \sigma^2)$. Donde

- σ_j^2 es la varianza de los efectos aleatorios b_{ji} para $j = 1, 2$
- σ_{12} es la covarianza entre los efectos aelatorios

Modelo

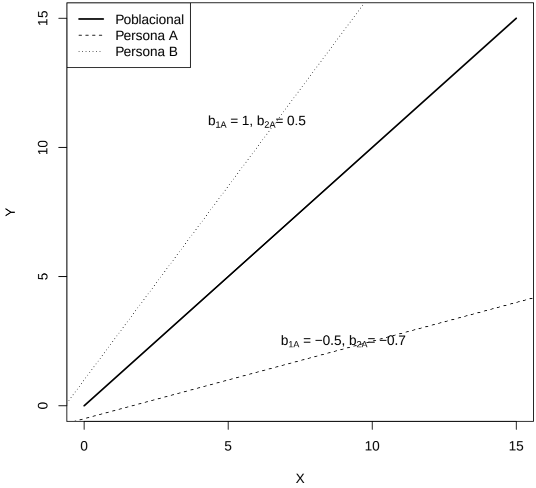
- Modelo

$$\begin{aligned} Y_{ij} &= X_{ij}^t\beta + b_{1i} + b_{2i}X_{ij2} + \epsilon_{ij} \\ &= (\beta_1 + \beta_2 X_{ij2} + \cdots + \beta_p X_{ij2}) + b_{1i} + b_{2i}Z_{ijk} + \epsilon_{ij} \\ &= (\beta_1 + b_{1i}) + (\beta_2 + b_{2i})X_{ij2} + \sum_{k\neq 2}\beta_k X_{ijk} + \epsilon_{ij} \end{aligned}$$

donde

- β_1: Intercepto poblacional
- b_{1i}: Desviación del individuo i-ésimo del intercepto poblacional
- β_2: Efecto poblacional de la variable edad
- b_{2i}: Desviación del individuo i-ésimo de la pendiente poblacional

Modelo lineal con intercepto y pendiente aleatoria



Estudio: Capacidad pulmonar

- Código de R

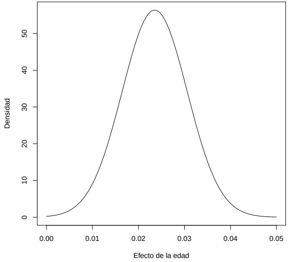
```
> model2 <- lme(logfev1 ~ age + loght + baseage + logbht,random= ~ age|id,fev1)
> summary(model2)
Linear mixed-effects model fit by REML

Random effects:
 Formula: ~age | id
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev      Corr
(Intercept) 0.110485541 (Intr)
age         0.007078381 -0.553
Residual    0.060237881

Fixed effects: logfev1 ~ age + loght + baseage + logbht
                 Value  Std.Error   DF   t-value p-value
(Intercept) -0.2883233 0.03871675 1692 -7.44699  0.0000
age          0.0235286 0.00139534 1692 16.86231  0.0000
loght        2.2371984 0.04353724 1692 51.38585  0.0000
baseage     -0.0165088 0.00745785  296 -2.21362  0.0276
logbht       0.2182148 0.14552087  296  1.49954  0.1348
```

- Efectos aleatorios

$$\begin{pmatrix} b_{1i} \\ b_{2i} \end{pmatrix} = Normal\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0{,}11^2 & -0{,}55(0{,}11)(0{,}01) \\ -0{,}001 & 0{,}01^2 \end{pmatrix}\right)$$

Modelo: Capacidad Pulmonar

- A nivel poblacional, por cada año adicional la capacidad pulmonar (en logaritmo) aumenta, en promedio, en 0.02.
- La variabilidad, individual, asociada a el efecto de la edad



donde

$$\beta_2 + b_{2i} \sim N(0{,}024, 0{,}01^2)$$

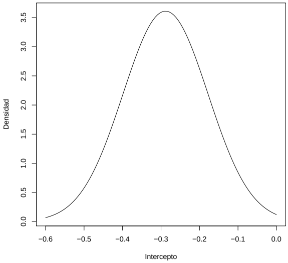
Modelo: Capacidad Pulmonar

- Correlación negativa

$$Corr(b_{1i}, b_{2i}) = -0{,}56$$

entonces mientras mas variabilidad hay en el punto de intercepción (punto basal) menor habra en el efecto de edad sobre capacidad pulmonar

- La variabilidad asociada con el individuo es alta



Estudio Capacidad Pulmonar

- Efectos fijos vs. aleatorios

```
> summary(model2)$tTable
                  Value   Std.Error   DF   t-value      p-value
(Intercept) -0.28832334 0.038716747 1692 -7.446993 1.513766e-13
age          0.02352863 0.001395339 1692 16.862308 4.319067e-59
loght        2.23719842 0.043537243 1692 51.385854 0.000000e+00
baseage     -0.01650884 0.007457846  296 -2.213620 2.761722e-02
logbht       0.21821482 0.145520868  296  1.499543 1.347986e-01
> summary(model0)$coef
               Estimate  Std. Error    t value      Pr(>|t|)
(Intercept) -0.33289431 0.020711666 -16.072792  9.570442e-55
age          0.02868270 0.002076943  13.810054  1.711923e-41
loght        2.04343017 0.068801382  29.700423 9.601334e-161
baseage     -0.01498081 0.003959672  -3.783345  1.593115e-04
logbht       0.39223217 0.082634781   4.746575  2.216469e-06
```

- Las estimaciones son afectadas: Sesgo
- Los errores éstandar son afectados: Pobre inferencia

Comparación de modelos

- La libreria **lme4** al igual que la libreria **lme** sirven para implementar modelos lineales de efectos mixtos
- La libreria **lmerTest** complementa la libreria **lme4** para poder comparar modelos
- Ejemplo

```
> model1a <- lmer(logfev1 ~ age + loght + baseage + logbht + (1|id),data=fev1)
> model2a <- lmer(logfev1 ~ age + loght + baseage + logbht + (1+age |id),data=fev1
> anova(model1a,model2a)
refitting model(s) with ML (instead of REML)
Data: fev1
Models:
object: logfev1 ~ age + loght + baseage + logbht + (1 | id)
..1: logfev1 ~ age + loght + baseage + logbht + (1 + age | id)
       Df     AIC     BIC logLik deviance Chisq Chi Df Pr(>Chisq)
object  7 -4517.6 -4478.4 2265.8  -4531.6
..1     9 -4586.7 -4536.4 2302.4  -4604.7 73.11      2  < 2.2e-16 ***
```

por lo tanto se prefiere el modelo con pendiente e intercepto aleatorio

Comparación de modelos

- Podemos explorar añadir un efecto aleatorio asociado a la relación entre altura y capacidad pulmonar.

- Al comparar modelos

```
> model3a <- lmer(logfev1 ~ age + loght + baseage + logbht + (1+age +loght|id),dat

> anova(model2a,model3a)
refitting model(s) with ML (instead of REML)
Data: fev1
Models:
object: logfev1 ~ age + loght + baseage + logbht + (1 + age | id)
..1: logfev1 ~ age + loght + baseage + logbht + (1 + age + loght |
..1:      id)
       Df     AIC     BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
object  9 -4586.7 -4536.4 2302.4  -4604.7
..1    12 -4602.6 -4535.5 2313.3  -4626.6 21.907      3  6.821e-05 ***
```

- El modelo con dos pendientes aleatorias (edad y altura) es el mejor modelo que describe los datos

Estudio: Capacidad Pulmonar

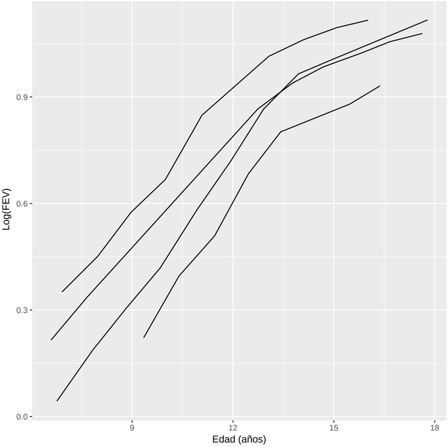


Figura 6: Capacidad pulmonar por edad: Modelo con pendiente aleatoria para edad y altura

Estudio: Acumulación de grasa corporal

- Estudio longitudinal en 162 niñas para evaluar su incremento en grasa corporal
- Estudios cientificos han encontrado que este proceso se inicia justo antes o alrededor de la primera mestruación
- El objetivo es evaluar el efecto del tiempo en este proceso
- Datos

```
> head(fat,n=12)
   id   age agemen  time   pbf
1   1  9.32  13.19 -3.87  7.94
2   1 10.33  13.19 -2.86 15.65
3   1 11.24  13.19 -1.95 13.51
4   1 12.19  13.19 -1.00 23.23
5   1 13.24  13.19  0.05 10.52
6   1 14.24  13.19  1.05 20.45
7   2  8.84  13.28 -4.44 16.17
8   2 10.08  13.28 -3.20 13.34
9   2 11.03  13.28 -2.25 16.05
10  2 12.77  13.28 -0.51 15.26
11  2 13.51  13.28  0.23 22.53
12  2 14.03  13.28  0.75 18.24
```

Estudio: Acumulación de grasa corporal

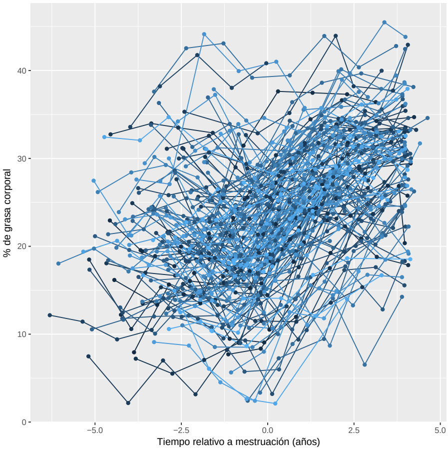


Figura 7: Evolución en el tiempo del porcentaje de grasa corporal

Estudio: Acumulación de grasa corporal

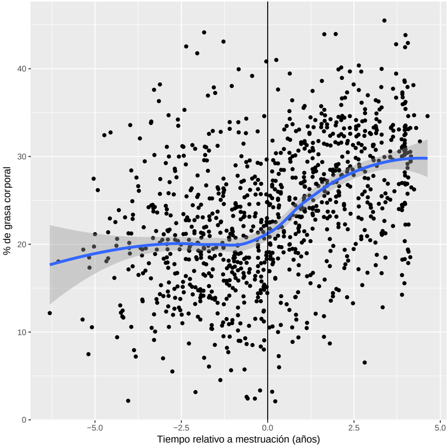


Figura 8: Evolución en el tiempo del porcentaje de grasa corporal: Curva suave

Estudio: Acumulación de grasa corporal

- Codigo de R

```
> fat$time0    <- fat$time*I(fat$time>=0)
> model1b      <- lmer(pbf ~ time + time0 + (1+ time + time0 | id),fat)
> summary(model1b)
Linear mixed model fit by REML

Random effects:
 Groups    Name        Variance Std.Dev. Corr
 id        (Intercept) 45.941   6.778
           time         1.631   1.277     0.29
           time0        2.750   1.658    -0.54  -0.83
 Residual               9.473   3.078
Number of obs: 1049, groups:  id, 162

Fixed effects:
            Estimate Std. Error       df t value Pr(>|t|)
(Intercept)  21.3614     0.5646 161.5600  37.838  < 2e-16 ***
time          0.4171     0.1572 108.4600   2.654  0.00915 **
time0         2.0471     0.2280 132.6800   8.980 2.22e-15 ***
```

- Existe un efecto mas pronunciado post inicio de la mestruación

Normalidad

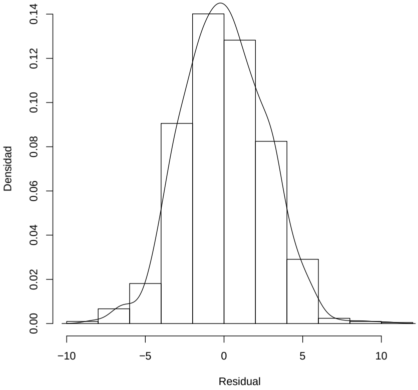


Figura 9: Normalidad de los residuales

Linealidad

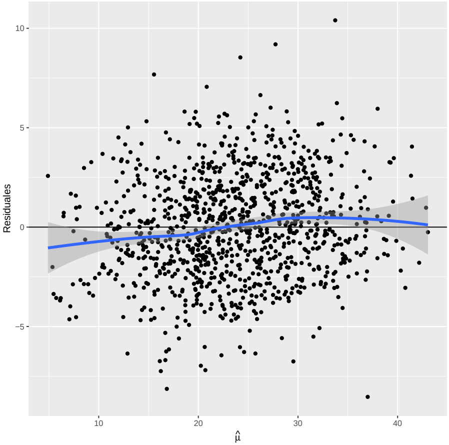


Figura 10: Predictor lineal vs. residuales: Evaluación de la función de enlace

Linealidad

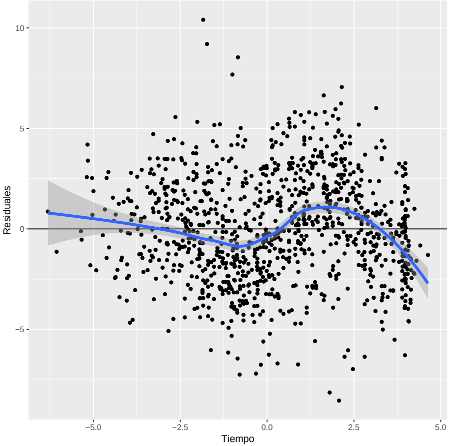


Figura 11: Covariable vs. residuales: Evaluación de la función de enlace

Normalidad

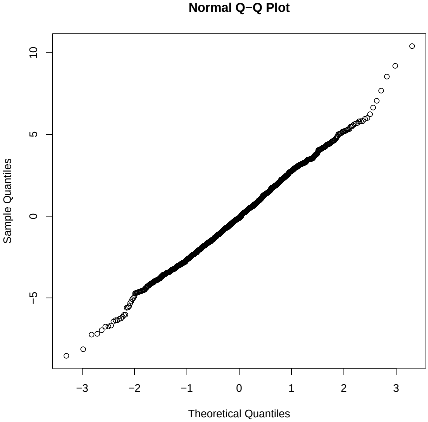


Figura 12: Normalidad de los residuales

Modelo

- Para la persona *ith* se mide un vector de respuestas

$$\underline{Y}_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix}$$

donde Y_{ij} es la respuesta medida para la persona i-ésima en el tiempo j-ésimo

- Para cada Y_{ij} se miden un conjunto de p covariables (que podrian cambiar o no en el tiempo)

$$\underline{X}_{ij} = \begin{pmatrix} X_{ij1} \\ X_{ij2} \\ \vdots \\ X_{ijp} \end{pmatrix}$$

- En general, tenemos las siguientes medidas de covariables de la persona *i*-ésima

$$X_i = \begin{pmatrix} X_{i1}^t \\ X_{i2}^t \\ \vdots \\ X_{in_i}^t \end{pmatrix} = \begin{pmatrix} X_{i11} & X_{i12} & \dots & X_{i1p} \\ & & & \\ \vdots & & & \\ X_{in_i1} & X_{in_i2} & \dots & X_{in_ip} \end{pmatrix}$$

Modelo

- Supuestos
  - Un subconjunto de β varia aleatoriamente de un individuo a otro
  - Individuos tienen su propia trayectoria en el tiempo
- Modelo

$$g(E[Y_{ij} \mid X_{ij}, b_i]) = X_{ij}^t\beta + b_i$$

donde

  - $E[Y_{ij}]$ es el valor esperado (media) de la variable a estudiar
  - $g(\dot{})$ es una función que enlaza la media con las covariables
  - β: efectos que <u>no</u> varian de persona a persona (**Efectos fijos**)
  - b_i: Variabilidad asociada al individuo i-ésimo (**Efecto aleatorio**)

Modelo

- Modelo

$$\log\left[\frac{P(Y_{ij}=1\mid b_i)}{P(Y_{ij}=0\mid b_i)}\right]=\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right)=X_{ij}^t\beta+b_i$$

donde

- μ_{ij} es la probabilidad de observar el evento en en la persona i-ésima en el tiempo j

$$\mu_{ij}=E[Y_{ij}\mid X_{ij}]$$

- β son los efectos fijos de las covariables
- b_i son los efectos aleatorios del cluster i

$$b_i\sim N(0,\sigma^2)$$

Estudio: Depresión

- Codigo de R

```
> model4 = glmer(outcome ~ severity + treat*time + (1|case), data=depre,
+               family=binomial(link="logit"))
> summary(model4)
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) [
 Family: binomial  ( logit )
Formula: outcome ~ severity + treat * time + (1 | case)

Random effects:
 Groups Name        Variance Std.Dev.
 case   (Intercept) 0.003231 0.05684
Number of obs: 1020, groups:  case, 340

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.02797    0.16406  -0.170    0.865
severity    -1.31488    0.15261  -8.616  < 2e-16 ***
treat       -0.05967    0.22239  -0.268    0.788
time         0.48274    0.11566   4.174 3.00e-05 ***
treat:time   1.01817    0.19150   5.317 1.06e-07 ***
```

- Despues de controlar por el tiempo, existe aun un pequeño efecto asociado al individuo

$$b_i \sim N(0, 0{,}003)$$

Modelo: Especificación

- Media

$$E[Y_{ij} \mid X_{ij}] = \mu_{ij}$$

y $g(\mu_{ij}) = X_{ij}^t \beta$

- Varianza

$$Var[Y_{ij} \mid X_{ij}] = \phi v(\mu ij)$$

donde ϕ es un parámetro de escala y $v(\dots)$ es una función de varianza que depende de la media

- Asociación entre medidas repetidas

$$Cov(Y_{ij}, Y_{ik}) = f(\alpha, \mu_{ij}, \mu_{ik})$$

Modelo

- Supuesto: Dado X_{ij}, Y_{ij} es independiente del respto de observaciones de esa covariable

$$
\begin{aligned}
E[Y_{ij} \mid X_i] &= E[Y_{ij} \mid X_{i1}, \ldots, X_{in_i}] \\
&= E[Y_{ij} \mid X_{ij}]
\end{aligned}
$$

- Precaución: No funciona cuando (Y_{ij}, X_{ij}) predicen $X_{i,j+1}$

Modelo: Especificación

- Media
$$E[Y_{ij} \mid X_{ij}] = \mu_{ij}$$
y $g(\mu_{ij}) = X_{ij}^t \beta$
  - Lineal: $g(\mu_{ij}) = \mu_{ij}$
  - Logistico: $g(\mu_{ij}) = \log\left[\mu_{ij}/(1-\mu_{ij})\right]$
- Varianza
$$Var[Y_{ij} \mid X_{ij}] = \phi v(\mu_{ij})$$
donde ϕ es un parámetro de escala y $v(\cdot)$ es una función de varianza que depende de la media.
  - Lineal: $v(\mu_{ij}) = 1$
  - Logistico: $v(\mu_{ij}) = \mu_{ij}(1-\mu_{ij})$

Modelo

- Asociación entre medidas repetidas

$$V_i = A_i^{1/2} Corr(Y_i) A_i^{1/2}$$

donde V_i es la **covarianza de trabajo** y

- $Corr(Y_i)$ **nuestro modelo** para la correlación entre datos
- A_i es una matriz diagonal cuyo componente j-ésimo es

$$\phi v(\mu_{ij})$$

donde

- $v(\mu_{ij})$ es **nuestro modelo** sobre el comportamiento de la media
- ϕ es estimado tipicamente de los datos

Estructura de correlación

- Independencia

$$Corr(Y_{is}, Y_{it}) = 0 \ \forall s, t$$

entonces

$$Cor(Y_i \mid \alpha) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

- Características
  - Observaciones sin correlación

Estructuras de correlación

- Intercambiable

$$Cor(Y_{is}, Y_{it}) = \alpha \;, \quad s \neq t \in \{1, \ldots, T\}$$

entonces

$$Corr(Y_i \mid \alpha) = \begin{pmatrix} 1 & \alpha & \alpha & \alpha \\ \alpha & 1 & \alpha & \alpha \\ \alpha & \alpha & 1 & \alpha \\ \alpha & \alpha & \alpha & 1 \end{pmatrix}$$

- Caracteristicas
  - La correlación es la misma para todo par de observaciones de un mismo individuo
  - No depende de cual alejadas estan entre ellas
  - Facilita la estimación

Estructuras de correlación

- Autoregresiva

$$Cor(Y_{is}, Y_{it}) = \alpha^{|s-t|} \quad s, t \in \{1, \ldots, T\}$$

entonces

$$Corr(Y_i \mid \alpha) = \begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^3 \\ \alpha & 1 & \alpha & \alpha^2 \\ \alpha^2 & \alpha & 1 & \alpha \\ \alpha^3 & \alpha^2 & \alpha & 1 \end{pmatrix}$$

- Caracteristica
  - Se tiene un solo parámetro a estimar
  - Asume que la correlación se debilita a medida que las mediciones se alejan

Estructura de correlación

- Sin extructura

$$Cor(Y_{is}, Y_{it}) = \alpha_{s,t} \;\; s, t \in \{1, \ldots, T\}$$

entonces

$$Corr(Y_i \mid \alpha) = \begin{pmatrix} 1 & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{12} & 1 & \alpha_{23} & \alpha_{24} \\ \alpha_{13} & \alpha_{32} & 1 & \alpha_{34} \\ \alpha_{14} & \alpha_{42} & \alpha_{43} & 1 \end{pmatrix}$$

- Caracteristica
  - La mas compleja de todas: No introduce ningun supuesto sobre los datos
  - Considera un parámetro para cada par de observaciones (conservando la simetria)
  - Requiere considerable cantidad de observaciones dentro de cada individuo (cluster)

Estudio: Depresión

```
> model5 = geem(outcome ~ severity + treat*time, id=case, data=depre,
+               family=binomial(link="logit"), corstr="independence" )
> summary(model5)
             Estimates Model SE Robust SE    wald         p
(Intercept)   -0.02799   0.1616    0.1742 -0.1607 8.723e-01
severity      -1.31400   0.1443    0.1460 -9.0000 0.000e+00
treat         -0.05959   0.2191    0.2285 -0.2608 7.943e-01
time           0.48240   0.1131    0.1199  4.0220 5.764e-05
treat:time     1.01700   0.1861    0.1877  5.4210 6.000e-08

 Estimated Correlation Parameter:  0
 Correlation Structure:  independence
 Est. Scale Parameter:  0.9859

 Number of GEE iterations: 2
 Number of Clusters:  340    Maximum Cluster Size:  3
 Number of observations with nonzero weight:  1020
```

- Model SE y Robust SE son el estimador de la desviación éstandar asumiendo el modelo de independencia y usando el estimador del sandwich, respectivamente
- Casi siempre el estimador del sandwich es mas amplio

Estudio: Depresión

```
> model6 = geem(outcome ~ severity + treat*time, id=case, data=depre,
+               family=binomial(link="logit"), corstr="exchangeable" )
> summary(model6)
            Estimates Model SE Robust SE    wald          p
(Intercept)  -0.02810   0.1614    0.1742 -0.1613 8.719e-01
severity     -1.31400   0.1439    0.1460 -9.0020 0.000e+00
treat        -0.05926   0.2190    0.2286 -0.2593 7.954e-01
time          0.48250   0.1133    0.1199  4.0230 5.757e-05
treat:time    1.01700   0.1864    0.1877  5.4190 6.000e-08

 Estimated Correlation Parameter:  -0.00337
 Correlation Structure:  exchangeable
 Est. Scale Parameter:  0.9859

 Number of GEE iterations: 2
 Number of Clusters:  340    Maximum Cluster Size:  3
 Number of observations with nonzero weight:  1020
>
```

- La estimación de la correlación es $-0,003$ (casi nula)
- **Interpretación**: Pudimos haber ignorado la correlación y asumir que todas las obervaciones son independientes

Referencias I

Agresti, A. (2007). *Introduction to Categorical Data Analysis*. Wiley, New Jersey.

Garcia, P., Holmes, K., Carcamo, C., Garnett, G., Hughes, J., Campos, P., and Whittington, W. (2012). Prevention of sexually transmitted infections in urban communities (Peru PREVEN): a multicomponent community-randomised controled trial. *Lancet*, 379(11):1120–1128.

Wedderburn, A. (1974). Quasi-Likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61(3):439–447.