Trabajo práctico N°2

Mariana Vargas Vieyra

August 8, 2017

Abstract

En este trabajo abordamos dos problemas distintos. En primer lugar estudiamos el efecto de dos tratamientos en una muestra de niveles de glucemia. Para ello usamos un modelo mixto generalizado. En segundo lugar, exploramos y modelamos la producción de leche de un grupo de vacas distribuidas en diferentes tambos. Para este último usamos un modelo no lineal mixto.

1 Problema 1

Nuestro objetivo para este primer problema es evaluar el efecto de dos programas de alimentación distintos, llámense A y B, en los niveles de glucemia hallados en medidas repetidas en adultos mayores.

1.1 Análisis exploratorio

El comando `summary` de `R` expone la estructura de los datos y las variables:

```
Programa.Alimentacion    Persona   Toma    GlucemiaAlta       Total
 A:50                   1      :10   1:20   Min.   : 2.00   Min.   :50
 B:50                   2      :10   2:20   1st Qu.: 8.00   1st Qu.:50
                        3      :10   3:20   Median :12.00   Median :50
                        4      :10   4:20   Mean   :12.99   Mean   :50
                        5      :10   5:20   3rd Qu.:16.00   3rd Qu.:50
                        6      :10          Max.   :40.00   Max.   :50
                        (Other):40

    prop
Min.   :0.0400
1st Qu.:0.1600
Median :0.2400
Mean   :0.2598
3rd Qu.:0.3200
Max.   :0.8000
```

Añadimos una variable `prop` que guarda las proporciones de los conteos por conveniencia. Esto nos permitirá mayor rapidez al graficar y agrega información importante a los datos. En la figura 1 podemos ver el comportamiento de dichas proporciones de las tomas para cada una de las personas, distinguidas

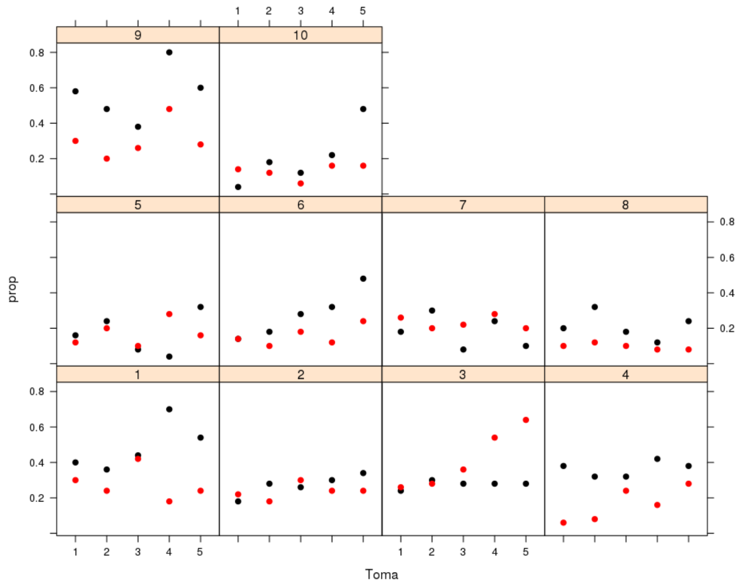

Figure 1: Gráfico del comportamiento de las proporciones en función de las tomas para cada persona.

por tratamiento. Notar que no hay un comportamiento cuadrático marcado con respecto a la variable `toma`.

Un gráfico de mosaico (figura 2) nos muestra que los datos están balanceados.

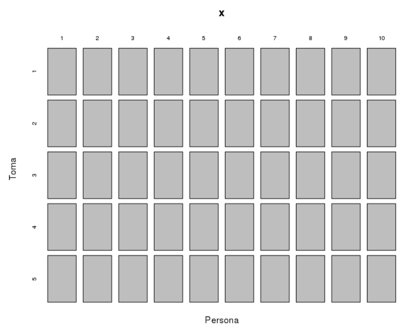

Figure 2: Gráfico de cruzamientos entre persona y toma.

1.2 Modelos

Hay una decisión de diseño respecto de qué variables serán tomadas como efecto fijo o aleatorio. Dado que hay sólo dos programas de alimentación siendo evaluados la variable `Programa.Alimentación` será un efecto fijo. Estamos interesadas en el efecto de estos sobre la población de adultos mayores, por lo tanto es razonable asumir nuestra variable `persona` como un efecto aleatorio cuyas observaciones representan una muestra de una población con distribución normal. Algunas observaciones al respecto:

1. Ignorar el efecto `Persona` implicaría una partición del error poco apropiada: parte del error podría ser considerado como aleatorio, los niveles de significatividad no serían confiables.

2. Asumir `Persona` como efecto fijo sería erróneo: estaríamos modelando cada una de estas personas y no podríamos extrapolar las conclusiones a la población de adultos mayores, que es el objetivo de nuestro estudio.

Ajustamos los siguientes modelos:

- El primer modelo evaluado es tal que

$$y_{i,j,k} = \beta_{0,i} + \beta_{1,i}\tau_i + \gamma_j + \varepsilon_{i,j,k} \tag{1}$$

  en donde $i = A, B$, $j = 1, ..., 10$, y $k = 1, ..., 5$, $y_{i,j,k}$ es la respuesta de la j-ésima persona bajo el k-ésimo tratamiento, τ_i es el efecto del i-ésimo tratamiento, que se asume efecto fijo, γ_j es el efecto aleatorio de la j-ésima persona y es tal que $\gamma_j \sim N(0, \sigma_p^2)$, y $\varepsilon_{i,j,k} \sim N(0, \sigma^2)$ el error aleatorio.

- A continuación agregamos `toma` como efecto fijo:

$$y_{i,j,k} = \beta_{0,i} + \beta_{1,i}\tau_i + \beta_{2,k}t_k + \gamma_j + \varepsilon_{i,j,k} \tag{2}$$

  en donde ahora t_k representa el efecto de la toma k-ésima.

Para comparar los modelos realizamos un análisis de la varianza con el comando `anova`, que arroja como resultado que el segundo modelo, el menos parcimonioso, ajusta mejor los datos.

```
    Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
mo1  3 727.74 735.56 -360.87   721.74
mo2  7 695.69 713.93 -340.85   681.69 40.048      4   4.23e-08 ***
```

Analizando los resultados del modelo seleccionado observamos que el tratamiento es significativo, así también como las tomas 1 (incluida en el intercepto), 4, y 5. La varianza estimada para `persona` es $\hat{\sigma_p}^2 = 0.22$.

Recordar que dado que la respuesta consiste en conteos usamos un modelo mixto generalizado con una función link logit, es decir,

$$\begin{aligned} X\beta &= \eta(\mu) \\ &= ln(\frac{\mu}{1-\mu}) \end{aligned}$$

en donde $\mu = \beta_0 + \beta_1\tau + \beta_2 T$ para el modelo de la ecuación (2) en forma matricial. Por lo tanto debemos tomar la inversa de esta función en los resultados arrojados por `R` para obtener las estimaciones.

2 Problema 2

En esta sección estudiamos la producción de leche de una muestra de vacas distribuidas en varios tambos. La figura 3 nos muestra la tendencia de la cantidad de litros de leche producidos en función de los días de lactancia para algunas vacas de entre la muestra con la que contamos. Podemos ver que, si bien no es muy marcada, hay un comportamiento exponencial negativo en los datos.

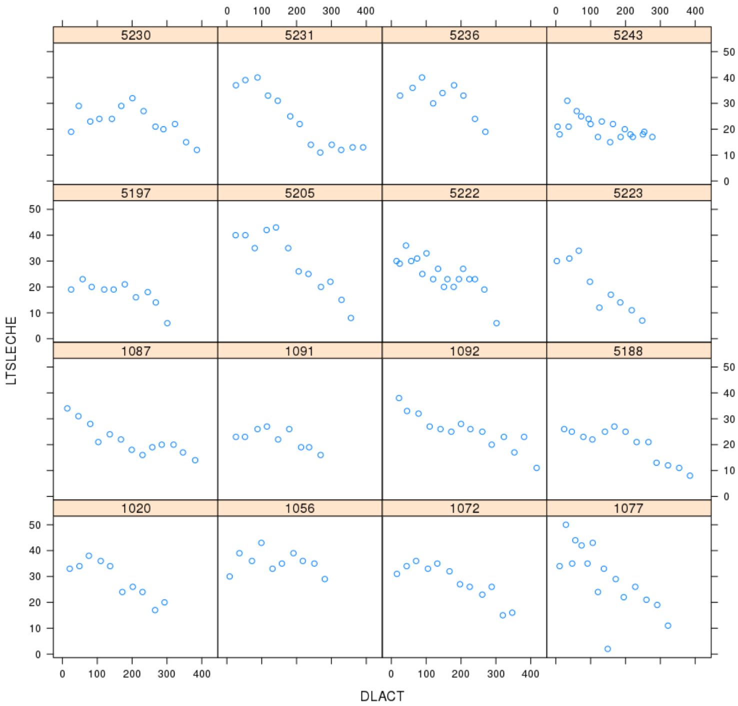

Figure 3: Tendencia de litros de leche producidos en función de los días de lactancia.

Un modelo apropiado para evaluar niveles de lactancia es el de Wood, definido de la siguiente manera:

$$y = ax^b e^{-cx} + \varepsilon$$

Si ajustamos un modelo de efectos fijos las estimaciones que arroja `R` son $\hat{a} = 19.55$, $\hat{b} = 0.13$, y $\hat{c} = 0.002$. El problema con ajustar una curva para todos los sujetos es que las tendencias individuales (que podemos ver en la figura 3) no son debidamente modeladas y se incorporan al modelo en forma de error aleatorio. El gráfico separado por vaca para residuos vs. predichos de la figura 4 muestra un patrón en su distribución [1].

Algo análogo ocurre con los tambos (figura 5).

Podemos ver en la figura 6 cómo ajusta la curva de este primer modelo a las primeras mil observaciones de nuestros datos.

[1] Recordemos que estamos graficando una submuestra de vacas.

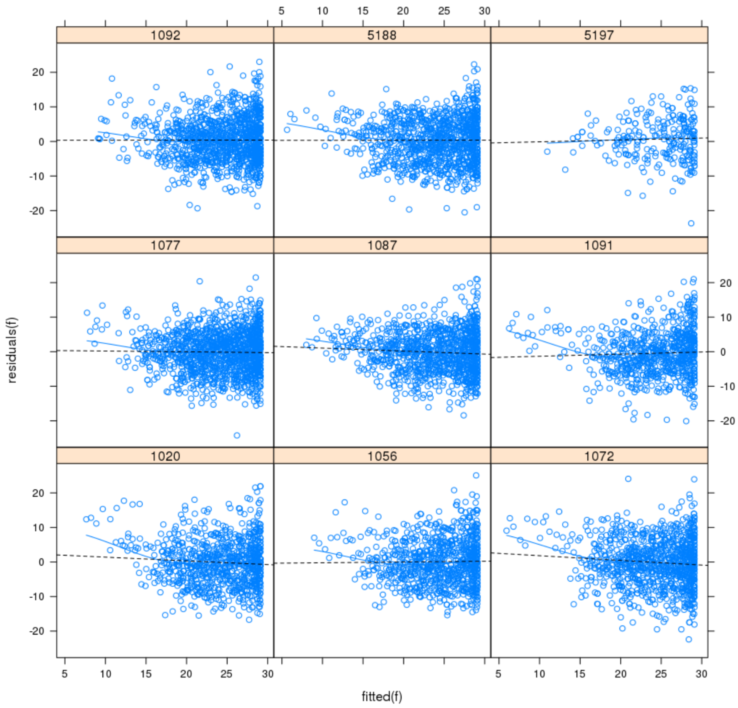

Figure 4: Residuos vs. predichos por vaca.

2.1 Comparación con modelos mixtos

Para resolver los problemas que acarrea el modelo anterior exploraremos algunas alternativas:

- Un modelo que agrupe por tambo, es decir, introduciremos un efecto aleatorio para el tambo. Más específicamente, asumiremos una alteración introducida por el factor `TAMBO` en el coeficiente multiplicativo a:

  $$y_{i,j} = (a + \beta_i)x^b e^{-cx} + \varepsilon_{i,j} \tag{3}$$

  en donde $y_{i,j}$ es la cantidad de litros de leche producido por la vaca j en el tambo i en función de los días de lactancia transcurridos, x. Comparamos este modelo con el de efectos fijos mediante un anova, y concluimos que es necesario modelar el efecto del tambo:

```
> anova.lme(f, frand)
      Model df      AIC      BIC    logLik   Test  L.Ratio p-value
f         1  4 55020.56 55048.72 -27506.28
frand     2  5 54796.17 54831.36 -27393.08 1 vs 2 226.3922  <.0001
```

- Un modelo que agrupe por tambo y a su vez por vaca. Esto es:

  $$y_{i,j} = (a + \beta_i + \gamma_j)x^b e^{-cx} + \varepsilon i, j \tag{4}$$

  Un anova para comparar entre los últimos dos modelos nos permite concluir que este último modelo es el que mejor ajusta los datos.

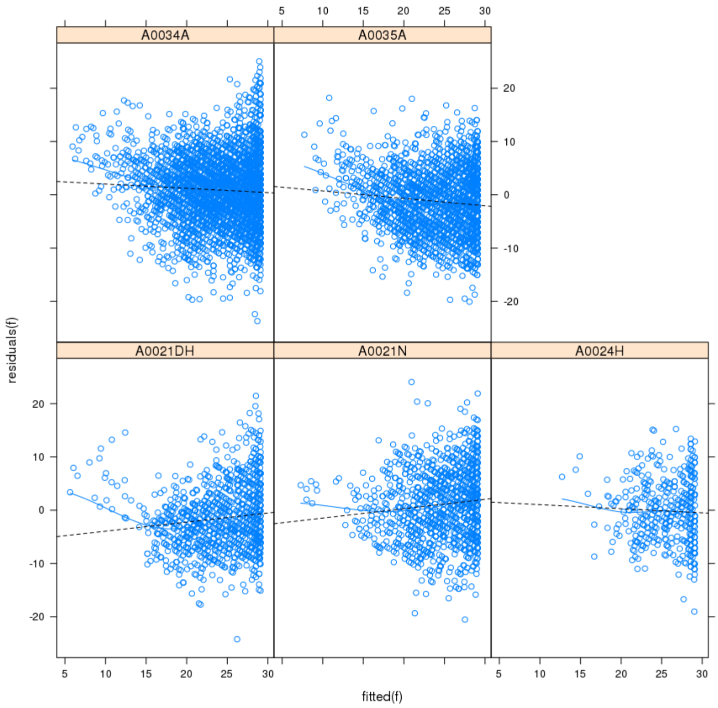

Figure 5: Residuos vs. predichos por tambo.

```
> anova.lme(frand, frand_cow)
          Model df      AIC      BIC    logLik   Test  L.Ratio p-value
frand         1  5 54796.17 54831.36 -27393.08
frand_cow     2  6 51695.21 51737.45 -25841.61 1 vs 2 3102.955  <.0001
```

Notar además que el patrón en los residuos vs. predichos ha mejorado, como se puede ver en la figura 7.

2.2 Predicción para 305 días de lactancia

Comenzamos por ajustar un modelo con un intercepto aleatorio para cada vaca, esto es:

$$y_{i,j} = (a + \gamma_j)x^b e^{-cx} + \varepsilon_{i,j} \tag{5}$$

en donde γ_j es el efecto aleatorio de la j-ésima vaca. En R,

```
frand_cow <- nlme(model = LTSLECHE ~ a * DLACT^b * exp(-c * DLACT),
                  fixed = list(a ~ 1, b~1, c~1),
                  random = a ~ 1|VACANUMERO,
                  data = df,
                  start = c(a=10, b=1, c=-0.001))
```

Usamos la ecuación 5 para calcular la cantidad de litros de leche esperada para cada día, para cada vaca. Las estimaciones de los parámetros arrojadas

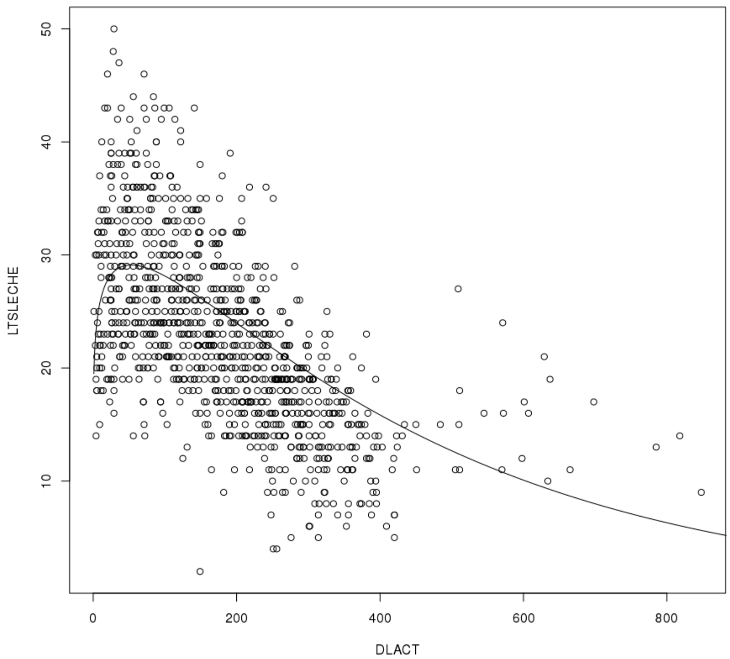

Figure 6: Curva del modelo de efectos fijos.

son $a = 14.52$, $b = 0.219$, y $c = 0.003$. El vector $\gamma = (\gamma_1, ...\gamma_v)$, $v =$ número de vacas, puede encontrarse con el comando

```
re <- random.effects(frand_cow)$a
```

Para obtener los totales por vaca agregamos la matriz de litros por día por vaca, `predicted`, sumando sobre el eje de los días:

```
accumulated = apply(predicted, 1, sum)
```

lo que nos da como resultado la estimación de cantidad de litros total después de 305 días de lactancia para cada vaca.

2.3 Correlación entre variables observadas y predichas

Luego de calcular los valores predichos que corresponden a cada uno de los percentiles solicitados, calculamos la correlación entre estos y las variables observadas. Obtuvimos los siguientes resultados:

- **Percentil 2.5,** correlación de 0.42,
- **Percentil 50,** correlación de 0.73,
- **Percentil 97.5,** correlación de 0.86.

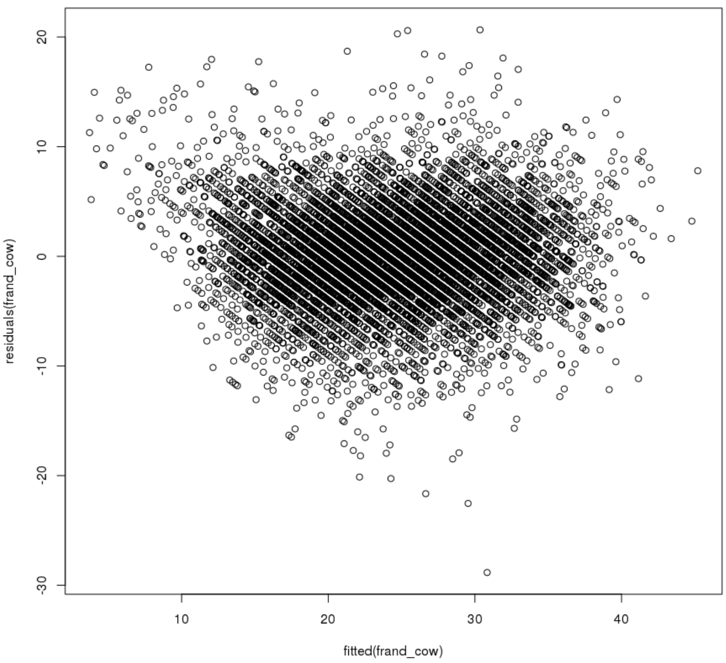

Figure 7: Residuos vs. predichos para el modelo (4).