1. Medidas Repetidas

Propuesta

El propósito será el estudiar y proponer métodos más flexibles a través de estructuras no lineales, usando los GAM y VGAM Hastie and Tibshirani (1990); Yee (2015). Se desarrollarán modelos de clasificación y/o de progresión basados en replicaciones en la variable respuesta y/o variable explicativa.

Desventajas a considerar: Trabajar con la estructura de las repeticiones implica mayor esfuerzo, y No necesariamente se obtienen mejores resultados. Aumenta la varianza.

1.1. Opciones para el desarrollo del modelo

Sea Y_i variable respuesta binaria:
Ej: $Y_i = 0$ sujeto sano y $Y_i = 1$ sujeto con Parkinson, $i = 1, \ldots, N$, $N = 80$ sujetos (40 sanos y 40 Parkinson)

$$\begin{aligned} Y_i &\sim Bernoulli(\theta_i) \\ \log \frac{\theta_i}{1-\theta_i} &= \eta_i \qquad logit \end{aligned}$$

Análogo para el caso de una variable respuesta continua:

$$\begin{aligned} Y_i &\sim Normal(\theta_i, \sigma_i^2) \\ \theta_i &= \eta_i \end{aligned}$$

Sean $z_{i1}, \ldots, z_{ip}$ variables explicativas para el sujeto i.
Ej: $p = 2$, edad y sexo.

Sean $x_{i1j}, \ldots, x_{iqj}$ variables explicativas para el sujeto i con repeticiones j.
Ej: características de la voz, $j = 1, 2, 3$ repeticiones y $q = 44$ características.

Predictor aditivo para usar GAM:

$$\begin{aligned} \eta_i &= f_1(z_{i1}) + f_2(z_{i2}) + \cdots + f_p(z_{ip}) \\ &+ g_1(x_{i11}, \ldots, x_{i1J}) + g_2(x_{i21}, \ldots, x_{i2J}) + \cdots + g_q(x_{iq1}, \ldots, x_{iqJ}) \end{aligned}$$

donde las funciones $f_1(z_{i1}), f_2(z_{i2}), \ldots, f_p(z_{ip})$ pueden ser de suavizado *smooth* (regresión polinomial, B-splines, etc.).

¿Cómo trabajar con las funciones $g_1(x_{i11}, \ldots, x_{i1J}), g_2(x_{i21}, \ldots, x_{i2J}), \ldots, g_q(x_{iq1}, \ldots, x_{iqJ})$**?**

Alternativas para el cálculo de $g_k(x_{ik1}, \ldots, x_{ikJ})$, $k = 1, \ldots, q$

1. **Usar las repeticiones como si fueran obtenidas de sujetos independientes:**

$$\begin{aligned} Y_{ij} &\sim Bernoulli(\theta_{ij}) \\ \log \frac{\theta_{ij}}{1-\theta_{ij}} &= \eta_{ij} \qquad logit \\ \eta_{ij} &= f_1(z_{i1}) + f_2(z_{i2}) + \cdots + f_p(z_{ip}) \\ &+ g_1(x_{i1j}) + g_2(x_{i2j}) + \cdots + g_q(x_{iqj}) \end{aligned}$$

esto implica que $Y_{i1} = Y_{i2} = \cdots = Y_{iJ}$ son la respuesta observada del mismo sujeto i, pero se consideran como si fueran independientes.

Y_1	z_{11}	z_{12}	x_{111}	x_{121}
			x_{112}	x_{122}
			x_{113}	x_{123}
Y_2	z_{21}	z_{22}	x_{211}	x_{221}
			x_{212}	x_{222}
			x_{213}	x_{223}

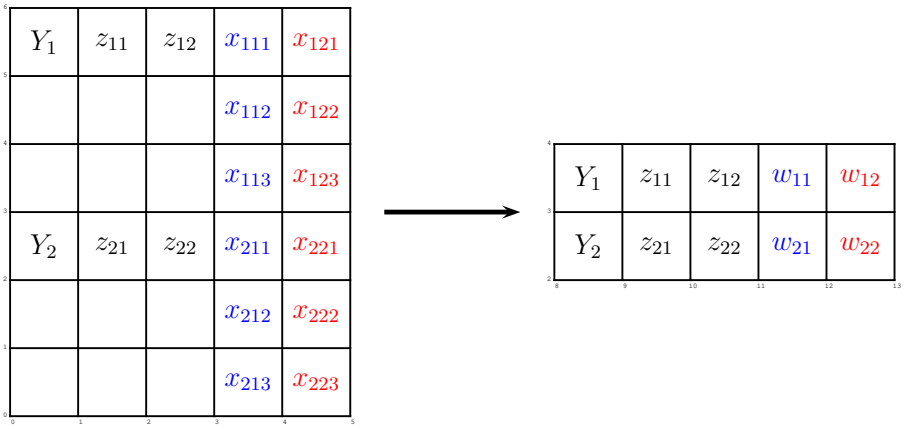
Y_1	z_{11}	z_{12}	x_{111}	x_{121}
Y_1	z_{11}	z_{12}	x_{112}	x_{122}
Y_1	z_{11}	z_{12}	x_{113}	x_{123}
Y_2	z_{21}	z_{22}	x_{211}	x_{221}
Y_2	z_{21}	z_{22}	x_{212}	x_{222}
Y_2	z_{21}	z_{22}	x_{213}	x_{223}

2. **Usar la media (o la mediana) de las repeticiones:**

$$w_{ik} = \frac{1}{J}\sum_{j=1}^{J} x_{ikj} \quad o \quad w_{ik} = mediana\{x_{ik1}, \ldots, x_{ikJ}\}$$
$$g_k(x_{ik1}, \ldots, x_{ikJ}) \;=\; g_k(w_{ik})$$

elegir algún método *smoothing* para $g_k(w_{ik})$.

$$\begin{aligned} Y_i &\sim Bernoulli(\theta_i) \\ \log\frac{\theta_i}{1-\theta_i} &= \eta_i \qquad logit \\ \eta_i &= f_1(z_{i1}) + f_2(z_{i2}) + \cdots + f_p(z_{ip}) \\ &+ g_1(w_{i1}) + g_2(w_{i2}) + \cdots + g_q(w_{iq}) \end{aligned}$$



3. **Usar las repeticiones como si fueran diferentes variables:**

$$g_k(x_{ik1}, \ldots, x_{ikJ}) \quad = \quad g_{k1}(x_{ik1}) + \cdots + g_{kJ}(x_{ikJ})$$

donde cada *smoothing* $g_{kj}(x_{ikj})$ puede variar para $k = 1, \ldots, q$ y $j = 1, \ldots, J$.

$$\begin{aligned} Y_i &\sim Bernoulli(\theta_i) \\ \log \frac{\theta_i}{1-\theta_i} &= \eta_i \qquad logit \\ \eta_i &= f_1(z_{i1}) + f_2(z_{i2}) + \cdots + f_p(z_{ip}) \\ &+ g_{11}(x_{i11}) + \cdots + g_{1J}(x_{i1J}) \\ &+ g_{21}(x_{i21}) + \cdots + g_{2J}(x_{i2J}) \\ &+ \cdots + g_{q1}(x_{iq1}) + \cdots + g_{qJ}(x_{iqJ}) \end{aligned}$$

Y_1	z_{11}	z_{12}	x_{111}	x_{121}
			x_{112}	x_{122}
			x_{113}	x_{123}
Y_2	z_{21}	z_{22}	x_{211}	x_{221}
			x_{212}	x_{222}
			x_{213}	x_{223}

Y_1	z_{11}	z_{12}	x_{111}	x_{112}	x_{113}	x_{121}	x_{122}	x_{123}
Y_2	z_{21}	z_{22}	x_{211}	x_{212}	x_{213}	x_{221}	x_{222}	x_{223}

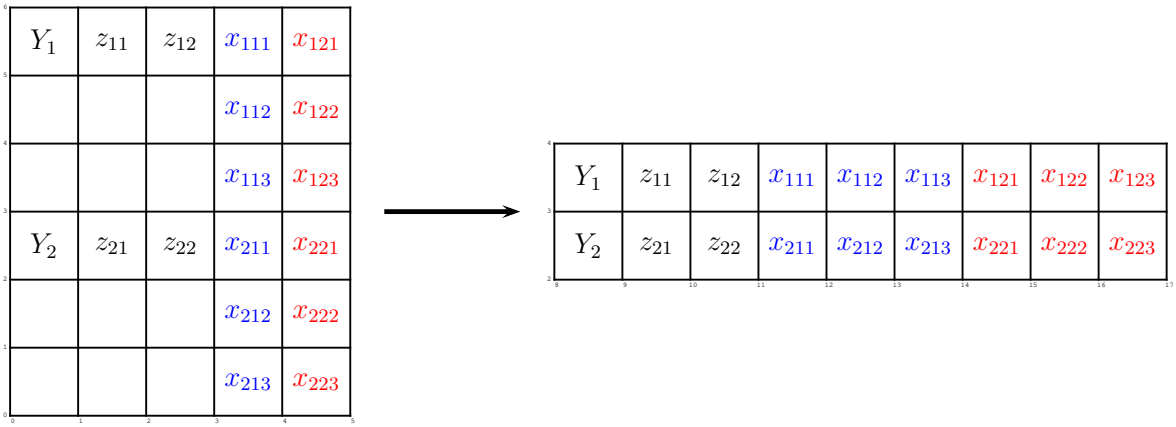
```
mod1 <- gam(Estado ~ s(HNR05.1) + s(HNR05.2) + s(HNR05.3)
                   + s(MFCC0.1) + s(MFCC0.2) + s(MFCC0.3) ,
          family = binomial(link=logit), data=Park_merge)

mod2 <- gam(Estado ~ s(HNR05.1,bs="cr") + s(HNR05.2,bs="cr") + s(HNR05.3,bs="cr")
                   + s(MFCC0.1,bs="cr") + s(MFCC0.2,bs="cr") + s(MFCC0.3,bs="cr") ,
          family = binomial(link=logit), gamma=1.4, data=Park_merge)
```

4. **Usar funciones suavizadas (smooths) de varias variables:**

$$g_k(x_{ik1}, \ldots, x_{ikJ}) \quad = \quad g_k(x_{ik1}, \ldots, x_{ikJ})$$

donde $g_k(x_{ik1}, \ldots, x_{ikJ})$ es el smooth de las repeticiones conjuntamente.

$$\begin{aligned} Y_i &\sim Bernoulli(\theta_i) \\ \log \frac{\theta_i}{1-\theta_i} &= \eta_i \qquad logit \\ \eta_i &= f_1(z_{i1}) + f_2(z_{i2}) + \cdots + f_p(z_{ip}) \\ &+ g_1(x_{i11}, \ldots, x_{i1J}) + g_2(x_{i21}, \ldots, x_{i2J}) + \cdots + g_q(x_{iq1}, \ldots, x_{iqJ}) \end{aligned}$$




```
mod4 <- gam(Estado ~ s(HNR05.1,MFCC0.1 k=5) + s(HNR05.2,MFCC0.2,k=5)
                   + s(HNR05.3,MFCC0.3,k=5) ,
            family = binomial(link=logit), gamma=1.4, data=Park_merge)

mod8 <- gam(Estado ~ te(HNR05.1,MFCC0.1,k=3) + te(HNR05.2,MFCC0.2,k=3)
                   + te(HNR05.3,MFCC0.3,k=3) ,
            family = binomial(link=logit), gamma=1.4, data=Park_merge)
```

`mgcv:: s(..., k=-1,fx=FALSE,bs="tp")` isotropic thin plate regression spline

`mgcv:: te( )` tensor product smoothing

`s(x1,x2)` isotropic thin plate regression spline

`te(x1,x2,k)` tensor product smooth, with dimension k for each marginal basis

`te(x1,x2,k=3)` $= \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$

2. Errores de Medición

Los errores de medición ocurren cuando una o más variables del modelo de interés no pueden observarse exactamente (Buonaccorsi, 2010; Carroll et al., 2006). Hay muchas razones por las cuales pueden ocurrir estos errores, dos de los errores más comunes son errores de instrumento (por la calibración de los instrumentos de medida) y de muestreo (en la recolección de los datos). Cuando ocurre un error de medición, en lugar de la variable de interés se observa una variable sustituta (*surrogate*) sujeta a error. Cuando los valores de las variables verdaderas y las observadas son categóricos, el error de medición se conoce como error de clasificación o mala clasificación (*misclassification*).

Buonaccorsi (2010) menciona que hay tres ingredientes principales en un problema de error de medición:

1. Un modelo para el verdadero valor (*modelo estructural*), que puede ser esencialmente cualquier modelo estadístico.
2. Un modelo de error de medición, que involucra la especificación de la relación que existe entre el valor verdadero (sin error) y el valor observado (sujeto a error).
3. Datos, información o supuestos adicionales, que puedan ser requeridos para corregir el error de medición. Esto no siempre puede obtenerse, en cuyo caso sólo podrá hacerse una evaluación del error de medición. Esta información adicional puede ser diversa, por ejemplo, conocimiento sobre algunos de los parámetros de error de medición o sus funciones, valores replicados (mediciones repetidas), errores estándar estimados asociados a las variables propensas a error, validación (interna o externa) en la que se obtienen valores verdaderos y mal medidos sobre un conjunto de sujetos, o variables instrumentales.

Buonaccorsi (2010) también menciona que existen dos objetivos generales en un problema de error de medición:

- Determinar cuáles son las consecuencias de ignorar los errores de medición.
- Determinar cómo corregir el error de medición.

Los efectos de los errores de medición son diversos, pueden resultar en una estimación más débil acerca de la relación entre las variables explicativas y la variable respuesta; y principalmente causan que las estimaciones de los coeficientes de medición estén sesgadas y por tanto es necesario corregir el error para tales efectos.

Se han propuesto modelos lineales jerárquicos generalizados para el diagnóstico y seguimiento de algunas enfermedades, con datos sujetos a con errores de medición (respuesta continua) o de clasificación (respuesta binaria u ordinal) (Naranjo et al., 2019a,b, 2020a,b).

2.1. Propuesta

El propósito será estudiar y proponer métodos que consideren el supuesto de error de medición o de clasificación, usando los GAM y VGAM Hastie and Tibshirani (1990); Yee (2015). Se desarrollarán modelos de clasificación y/o de progresión basados en errores de medición en la variable respuesta y/o variable explicativa.

Suponga que la variable respuesta observada Y_i^* está sujeta a un error de medición (error aditivo) Buonaccorsi (2010); Carroll et al. (2006),

$$Y_i^* \sim \mathrm{N}(W_i, \sigma^2), \tag{1}$$

donde la variable latente W_i representa la respuesta no observada, relacionada a un conjunto de variables explicativas $x_{i1}, \ldots, x_{ip}$ a través de un GAM,

$$\begin{aligned} \mathbb{E}[W_i] &= \mu(\mathbf{x}_i), \\ g(\mu(\mathbf{x}_i)) &= f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}). \end{aligned} \tag{2}$$

Para el proceso de estimación se requiere que en (1) se introduzcan datos, información o supuestos adicionales, que puedan ser requeridos para corregir el error de medición. Esta información adicional puede ser diversa, por ejemplo, con métodos Bayesianos se introducen distribuciones iniciales informativas acerca de los parámetros, también se pueden introducir restricciones monótonas $W_{i1} \leq W_{i2} \leq \cdots \leq W_{iT}$ si se tienen observaciones a lo largo del tiempo de cada sujeto.

2.2. Caso de Aplicación

$Y_{i1}, \ldots, Y_{iT}$	variable respuesta continua longitudinal, restricción de monotonía $Y_{i1} \leq Y_{i2} \leq \cdots \leq Y_{iT}$ sujeto con Parkinson.
$z_{i1}, \ldots, z_{ip}$	variables explicativas, ej. edad y sexo.
$x_{i1j}, \ldots, x_{iqj}$	variables explicativas con repeticiones j características de la voz, $j = 1, 2, 3$ repeticiones y $q = 44$ características

Alternativas para el cálculo

1. Distribución Truncada dependiente del tiempo anterior, por ejemplo la Normal

$$Y_{it}|Y_{i,t-1} \sim \text{Normal}(\eta_{it}, \sigma^2)T[Y_{it} \geq y_{i,t-1}]$$

Bibliografía

Buonaccorsi, J. P. (2010). *Measurement Error*. Chapman & Hall/CRC, London.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman & Hall/CRC, Boca Raton, Florida, second edition.

Hastie, T. J. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall.

Naranjo, L., Esparza, L. J. R., and Pérez, C. J. (2020a). A hidden Markov model to address measurement errors in ordinal response scale and non-decreasing process. *Mathematics*, 8(4):622.

Naranjo, L., Lesaffre, E., and Pérez, C. J. (2020b). A mixed hidden Markov model for multivariate monotone disease processes in the presence of measurement errors. *Statistical Modelling*, 0(0):1471082X20973473.

Naranjo, L., Pérez, C. J., Fuentes-García, R., and Martín, J. (2019a). A hidden Markov model addressing measurement errors in the response and replicated covariates for continuous nondecreasing processes. *Biostatistics*, kxz004:1–15.

Naranjo, L., Pérez, C. J., Martín, J. R., Mutsvari, T., and Lesaffre, E. (2019b). A Bayesian approach for misclassified ordinal response data. *Journal of Applied Statistics*, 46(12):2198–2215.

Yee, T. W. (2015). *Vector Generalized Linear and Additive Models. With an Implementation in R (Vol. Springer Series in Statistics)*. Springer.