UNIVERSIDAD NACIONAL DE COLOMBIA

Facultad de ciencias

Departamento de estadística

Tercer caso de estudio Estadística Bayesiana

2023-2

Autor

Jherson Guzman R., Sebastián A. Lozano R.

Docente

Juan Camilo Sosa Martínez

Noviembre 2023

1. Alcaldía de Bogotá 2023

Preguntas

Ajustar el modelo propuesto usando un muestreador de Gibbs con $a = b = 1$ (incluir un anexo con todos los detalles). Reportar visual y tabularmente las estimaciones puntuales, los intervalos de credibilidad al 95 % y los resultados oficiales de la Registraduría Nacional del Estado Civil para Galán, Bolivar y Oviedo, expresando todas las cifras en puntos porcentuales. Interpretar los resultados obtenidos (máximo 500 palabras).

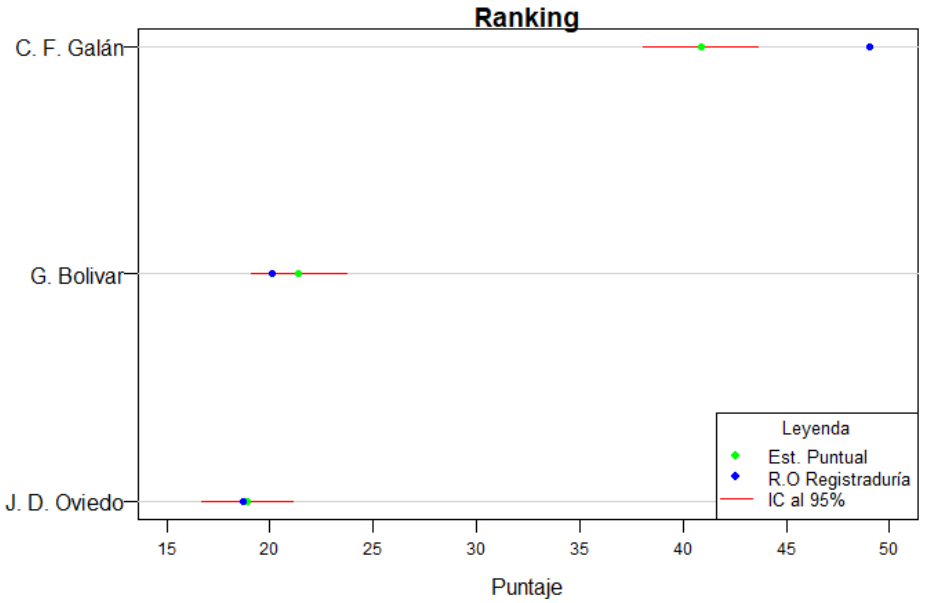

Figura 1: Resultados gráficos de la alcaldía

Candidato	Est.Puntual	Li IC95 %	Ls IC95 %	R.O registraduría
C. F. Galán	40.90 %	38.10 %	43.60 %	49.02 %
G. Bolivar	21.40 %	19.10 %	23.70 %	20.11 %
J. D. Oviedo	18.90 %	16.70 %	21.10 %	18.71 %

Tabla 1: Resumen de Resultados

Nota: Las columnas en la tabla anterior son: Est.Puntual = Estimación Puntual, Li IC95 % = Límite inferior del intervalo de credibilidad al 95 %, Ls IC95 % = Límite superior del intervalo de credibilidad al 95 % y R.O registraduría = resultados oficiales de la Registraduría Nacional del Estado Civil.

Interpretación:

De los resultados anteriores, observamos que el orden de nuestra estimación puntual se mantiene en los resultados de la Registraduría (*1.* C. F. Galán, *2.* G. Bolivar, *3.* J. D. Oviedo). Sin embargo, resulta llamativo el hecho de que para el candidato C.F. Galán, la estimación puntual (40.90 %) está considerablemente distante del valor real proporcionado por la Registraduría (49.02 %). De hecho, este valor no cae dentro del intervalo de credibilidad al 95 %, una situación que no ocurre con los otros dos candidatos, donde tanto la estimación como el valor real son muy cercanos. En ambos casos, el valor dado por la Registraduría se encuentra dentro del intervalo de credibilidad.

Otro aspecto que destaca de los resultados anteriores es que, para el primer candidato, el valor proporcionado por la Registraduría es superior al obtenido mediante la estimación (es decir que el modelo propuesto subestimo la intención de voto). Esta situación contrasta con los otros dos candidatos, donde el valor dado por la Registraduría resulta ser menor que la estimación (en este caso hubo una leve sobreestimación por parte del modelo).

2. Datos Diabetes

Modelo 1: Regresión clásica previa unitaria

Distribución previa: Previa unitaria (*unit information prior*; Kass y Wasserman, 1995).

Modelo 2: Regresión clásica previa g

Distribución previa: Previa g (*g-prior*; Zellner, 1986).

Modelo 3: Regresión rígida

Distribución previa:

$$p(\boldsymbol{\beta}, \sigma^2, \lambda) = \mathsf{N}(\boldsymbol{\beta} \mid \mathbf{0}_p, \tfrac{\sigma^2}{\lambda}\mathbf{I}_p) \cdot \mathsf{GI}(\sigma^2 \mid \nu_0/2, \nu_0\,\sigma_0^2/2) \cdot \mathsf{G}(\lambda \mid a_\lambda, b_\lambda)\,,$$

con $\nu_0 = 1$, $\sigma_0^2 = \hat{\sigma}^2_{\text{OLS}}$, $a_\lambda = 1$ y $b_\lambda = 2$.

Modelo 4: Regresión con errores correlacionados

Distribución muestral:

$$\boldsymbol{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2, \rho \sim \mathsf{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{C}_\rho)\,,$$

donde $\mathbf{C}_\rho$ es una matriz con estructura autoregresiva de primer orden de la forma

$$\mathbf{C}_\rho = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}.$$

Distribución previa:

$$p(\boldsymbol{\beta}, \sigma^2, \rho) = \prod_{j=1}^{p} \mathsf{N}(\beta_j \mid 0, \tau_0^2) \cdot \mathsf{GI}(\sigma^2 \mid \nu_0/2, \nu_0\, \sigma_0^2/2) \cdot \mathsf{U}(\rho \mid a_\rho, b_\rho)$$

con $\tau_0^2 = 50$, $\nu_0 = 1$, $\sigma_0^2 = \hat{\sigma}^2_{\text{OLS}}$, $a_\rho = 0$ y $b_\rho = 1$.

Preguntas

Ajustar cada modelo utilizando los datos de entrenamiento $(\boldsymbol{y}_{\text{train}}, \mathbf{X}_{\text{train}})$ (incluir un anexo con todos los detalles).

1. Para cada modelo, generar $\hat{\boldsymbol{y}}_{\text{test}} = \mathbf{X}_{\text{test}}\hat{\boldsymbol{\beta}}$ usando los coeficientes de regresión estimados $\hat{\boldsymbol{\beta}} = \mathsf{E}(\boldsymbol{\beta} \mid \boldsymbol{y}_{\text{train}})$. Graficar $\hat{y}_{\text{test}}$ frente y_{test} y calcular el error absoluto medio $\frac{1}{n}\sum_i |y_{\text{test},i} - \hat{y}_{\text{test},i}|$ correspondiente.

   Solución

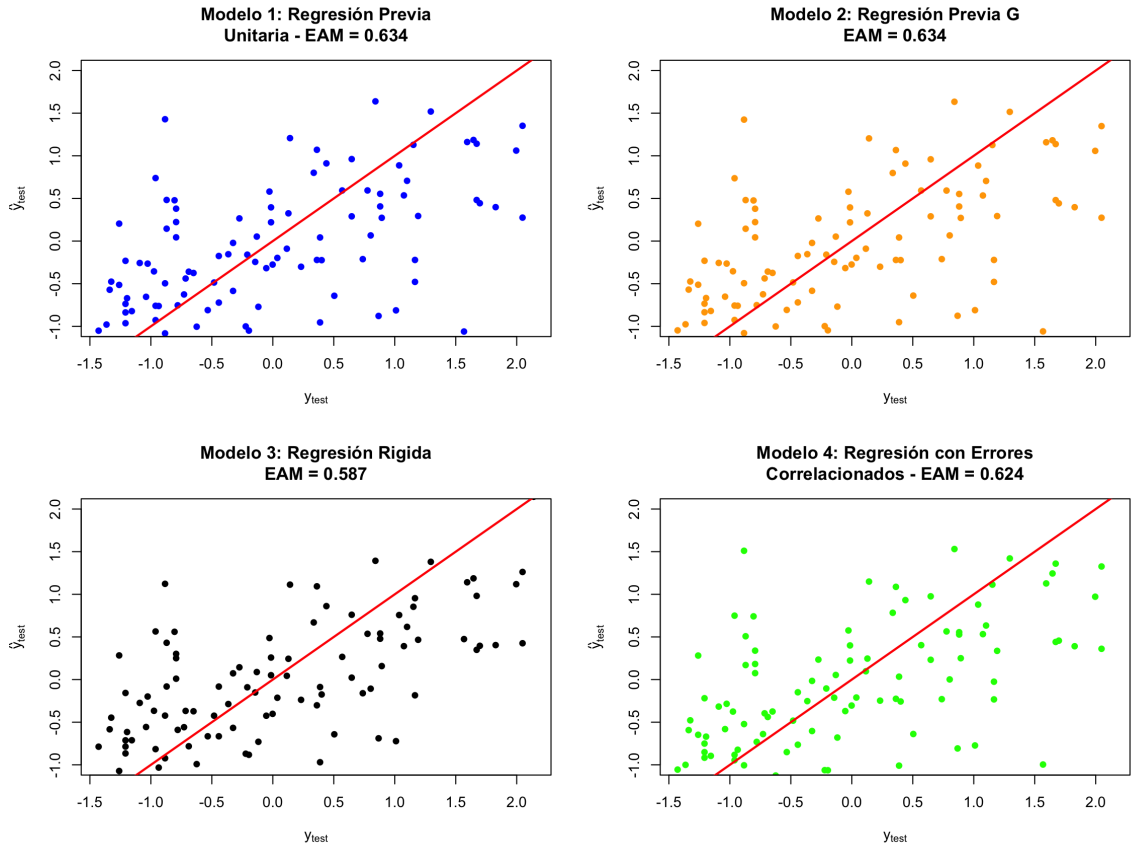

Figura 2: Estimación de Y-test para los modelos

2. Para cada modelo, chequear la bondad de ajuste usando la media como estadístico de prueba. Graficar la distribución predictiva posterior por medio de un histograma. **Solución**

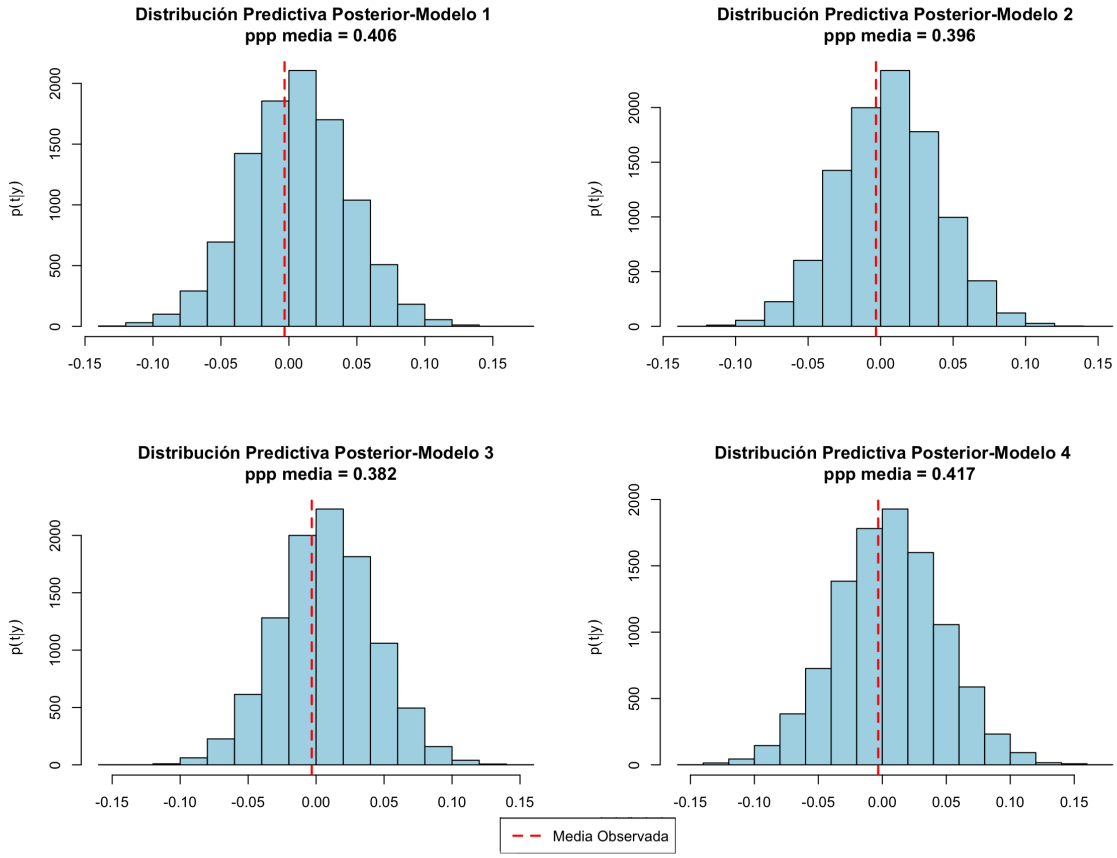

Figura 3: Histograma de distribución predictiva posterior por modelo

3. Para cada modelo, calcular el DIC. Presentar los resultados tabularmente usando tres (3) cifras decimales.

Solución

Modelo	DIC
Modelo 1 - Regresión Clásica Previa Unitaria	1440.190
Modelo 2 - Regresión Clásica Previa g	1426.252
Modelo 3 - Regresión Rigida	1425.188
Modelo 4 - Regresión con errores correlacionados	1438.010

Tabla 2: Tabla de comparación de modelos: DIC

4. Interpretar los resultados obtenidos en los numerales anteriores (máximo 500 palabras).

   Solución

   En este ejercicio se plantearon cuatro modelos distintos para un mismo conjunto de datos, posterior a su modelamiento se realizaron varios métodos de validación. A continuación se explicarán los resultados obtenidos. En primera instancia evaluamos la capacidad predictiva de los modelos usando su **Error absoluto medio (EAM)**, de acuerdo esta función de perdida el mejor modelo resulto ser la *Regresión rigida* con un EAM = 0,587, seguida por la *Regresión con errores correlacionados* y en ultimo lugar los modelos *Regresión Previa Unitaria* y *Regresión previa G*, de las gráficas vemos que en ningún caso hay demasiados puntos sobre o debajo de la recta $y_{test} = \hat{y}_{test}$, lo que indica buena capacidad predictiva de todos los modelos.

   Posteriormente, verificamos la bondad de ajuste del modelo usando la media como estadística de prueba, en este caso los modelos que mejor se comportaron fueron *Regresión Previa Unitaria* y *Regresión con errores correlacionados*, pues fueron en los que obtuvimos un *ppp* más cercano a 0,5, sin embargo cabe destacar que los otros modelos no tuvieron un valor *ppp* muy lejano a estos, la diferencia entre el peor modelo (Rígida con un *ppp* = 0,382) y mejor modelo fue tan solo de 0,3 aproximadamente, lo que indicaría una buena bondad de ajuste para todos los modelos.

   Por último, se hizo la validación interna de los modelos calculando su **DIC**, en este caso el modelo que resulto mejor fue el modelo *Regresión Rigida*, pues era el modelo que presentaba el DIC más bajo.

   Después de analizar los resultados obtenidos, concluimos que el mejor modelo parece ser la Regresión Rigida pues fue el modelo con mejor habilidad predictiva y mejor comportamiento evidencia en su validación interna (calculando el DIC), y a pesar de que fue el

peor modelo de acuerdo a la bondad de ajuste, como ya se dijo anteriormente no fue un valor muy lejano a 0,5 ni muy lejano a los valores *ppp* de los otros modelos.

Anexo

A continuación se van a mostrar las distribuciones condicionales completas para los modelos trabajados en el caso.

Modelo para encuesta alcaldía Bogotá

$$\theta|\text{resto} \sim \text{Dirichlet}(\mathbf{n}+\alpha)$$

$$p(\alpha|\text{resto}) \propto \frac{\Gamma(k\alpha)}{\Gamma(\alpha)} \prod_{j=1}^{k} \theta_j^{\alpha-1} \frac{b^a}{\Gamma(a)} \alpha^{a-1} exp\{b\alpha\}$$

El parámetro α no cuenta con una distribución condicional completa cerrada. De modo que para el modelamiento se uso el algoritmo de Metropolis. En nuestro ejercico usando un valor $\delta = 1{,}15$ se obtuvo una tasa de aceptación del 38,64 %

Log.verosimilitud:

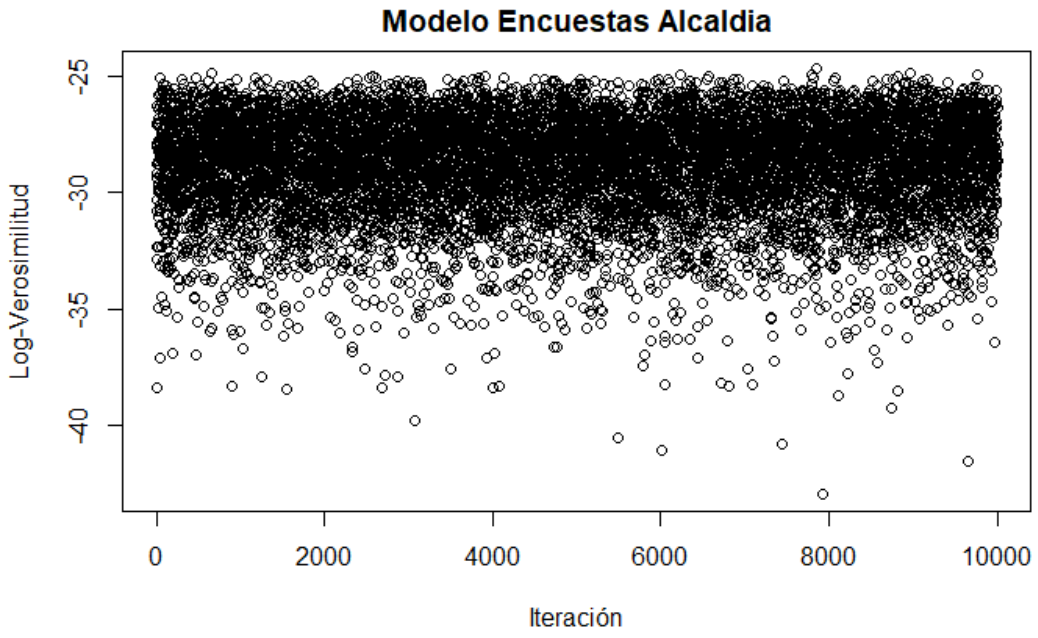

Figura 4: Histograma de distribución predictiva posterior por modelo

Modelo 1 - Regresión clásica previa unitaria:

$$\boldsymbol{\beta} \mid \text{resto} \sim \mathsf{N}_p\left((\boldsymbol{\Sigma}_0^{-1} + \tfrac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0 + \tfrac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\boldsymbol{y}), (\boldsymbol{\Sigma}_0^{-1} + \tfrac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right)$$
$$\sigma^2 \mid \text{resto} \sim \mathsf{GI}\left(\frac{\nu_0 + n}{2}, \frac{\nu_0\sigma_0^2 + \mathsf{SSR}(\boldsymbol{\beta})}{2}\right)$$

Para este modelo usamos un muestreador de Gibbs con un periodo de calentamiento de 11000 y donde 1000 hacían parte del periodo de calentamiento.

Modelo 2 - Regresión clásica previa g:

$$\boldsymbol{\beta} \mid \text{resto} \sim \mathsf{N}_p\left(\frac{g}{g+1}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{y}, \frac{g}{g+1}\sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right)$$
$$\sigma^2 \mid \text{resto} \sim \mathsf{GI}\left(\frac{\nu_0 + n}{2}, \frac{\nu_0\sigma_0^2 + \mathsf{SSR}_g(\boldsymbol{\beta})}{2}\right)$$

Para este modelo usamos un muestreador de Gibbs con un periodo de calentamiento de 11000 y donde 1000 hacían parte del periodo de calentamiento.

Modelo 3 - Regresión Rígida:

$$\boldsymbol{\beta} \mid \text{resto} \sim \mathsf{N}_p \left(\left(\frac{1}{\sigma^2}(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda I_p) \right)^{-1} \left(\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\boldsymbol{y} \right), \left(\frac{1}{\sigma^2}(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda I_p) \right)^{-1} \right)$$

$$\sigma^2 \mid \text{resto} \sim \mathsf{GI} \left(\frac{n + p + \nu_0}{2}, \frac{\nu_0\sigma_0^2 + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta} + \mathsf{SSR}(\boldsymbol{\beta})}{2} \right)$$

$$\lambda \mid \text{resto} \sim \mathsf{G} \left(\frac{p}{2} + a_\lambda, \frac{\boldsymbol{\beta}^T\boldsymbol{\beta}}{2\sigma^2} + b_\lambda \right)$$

En este modelo usamos 101000 iteraciones donde 1000 fueron periodo de calentamiento y se hizo un muestreo sistemático cada 10 para finalmente obtener 10000 muestras.

Modelo 4 - Regresión con errores correlacionados:

$$\boldsymbol{\beta} \mid \text{resto} \sim \mathsf{N}_p \left(\left(\frac{1}{\sigma^2}\mathbf{X}^T - \mathbf{C}_\rho^{-1}\mathbf{X} + \Sigma_0^{-1} \right)^{-1} \left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{C}_\rho^{-1}\boldsymbol{y} \right), \left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{C}_\rho^{-1}\mathbf{X} + \Sigma_0^{-1} \right)^{-1} \right)$$

$$\sigma^2 \mid \text{resto} \sim \mathsf{GI} \left(\frac{\nu_0 + n}{2}, \frac{\nu_0 + n}{2}, \frac{\nu_0\sigma_0^2 + SSR_\rho}{2} \right)$$

$$P(\rho \mid \text{resto}) \propto |\boldsymbol{C}_\rho|^{\frac{1}{2}} \exp \left\{ -\frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{C}_\rho^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \right\}$$

En este modelo usamos 41000 iteraciones donde 1000 fueron periodo de calentamiento y se hizo un muestreo sistemático cada 4 para finalmente obtener 10000 muestras.

Dado que el párametro ρ no cuenta con una distribución condicional completa cerrada se uso el algoritmo de metropolis para poder continuar con el modelamiento. Usamos un valor de $\delta = 0{,}325$ y obtuvimos una tasa de aceptación del $40{,}19\,\%$

Log. verosimilitud de los modelos parte 2:

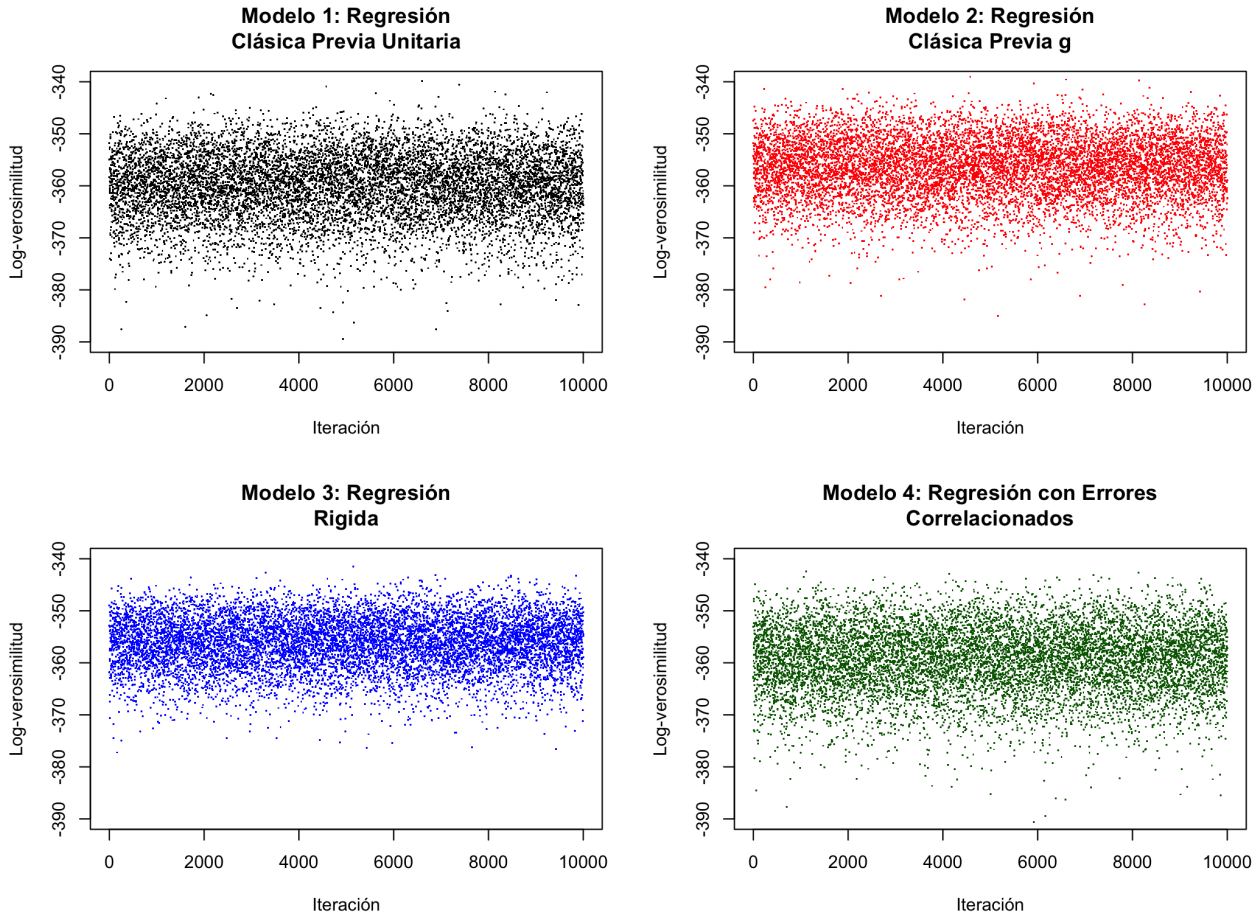

Figura 5: Panel con cadenas de log-verosimilitud de los modelos

Referencias

- https://rpubs.com/jstats1702/1048419
- https://rpubs.com/jstats1702/968742
- https://www.registraduria.gov.co/