Universidad Nacional de Colombia

Facultad de Ciencias

Departamento de Estadística

Estadística Bayesiana

Caso de estudio 3

Autores

Juan David Duitama Correa

jduitama@unal.edu.co

Daniel Hoyos Mateus

dhoyosm@unal.edu.co

Docente

Juan Camilo Sosa Martínez

Bogotá, junio de 2023

Caso de estudio 3

Selección de modelos

Puede ocurrir que en un análisis de regresión haya un gran número de variables independientes x, aunque puede que la mayoría de estas variables no tengan una relación sustancial con la variable dependiente y. En estas situaciones, incluir todas las variables regresoras en el modelo de regresión conduce a modelos saturados poco parsimoniosos difíciles de interpretar con un rendimiento deficiente. Por lo tanto, se recomienda considerar en el modelo final solo aquellas variables x para las que exista evidencia sustancial de una asociación con y. Esto no solo produce análisis de datos más simples, sino que también proporciona modelos con mejores propiedades estadísticas en términos de predicción y estimación.

Datos de diabetes

Considere la base de datos de diabetes dada en la Sección 9.3 de Hoff (2009, p. 161), que contiene datos asociados con 10 medidas basales $x_1, \ldots, x_{10}$ en un grupo de 442 pacientes diabéticos, así como una medida de progresión de la enfermedad y tomada un año después de las medidas basales. A partir de estos datos, el objetivo es hacer un modelo predictivo para y basado en $x_1, \ldots, x_{10}$ (tanto y como las x_j se encuentran estandarizadas). Esto da un total de $p = 10$ variables regresoras (no es necesario considerar el intercepto porque todas las variables se encuentran estandarizadas).

Modelamiento

Se considera un modelo de regresión de la forma $\boldsymbol{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2 \sim \mathsf{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$, donde $\boldsymbol{y}$ es un vector de $n \times 1$ que contiene los valores de la variable respuesta, $\mathbf{X}$ es una matriz de $n \times p$ que contiene los valores de las variables regresoras, $\boldsymbol{\beta}$ es un vector de $p \times 1$ que contiene los parámetros desconocidos, y finalmente, $\mathbf{I}_n$ es la matriz identidad de $n \times n$.

Para evaluar los modelos de regresión, se dividieron aleatoriamente a los 442 individuos con diabetes en 342 individuos de entrenamiento y 100 individuos de prueba, lo que provee un conjunto de datos de entrenamiento $(\boldsymbol{y}_{\text{train}}, \mathbf{X}_{\text{train}})$ y un conjunto de datos de prueba $(\boldsymbol{y}_{\text{test}}, \mathbf{X}_{\text{test}})$. Así, se ajustan los modelos usando $(\boldsymbol{y}_{\text{train}}, \mathbf{X}_{\text{train}})$, y luego, usando los coeficientes de regresión estimados $\hat{\boldsymbol{\beta}} = \mathsf{E}(\boldsymbol{\beta} \mid \boldsymbol{y}_{\text{train}})$, se genera $\hat{\boldsymbol{y}}_{\text{test}} = \mathbf{X}_{\text{test}} \hat{\boldsymbol{\beta}}$. Luego, se evalúa el rendimiento predictivo del modelo comparando $\hat{\boldsymbol{y}}_{\text{test}}$ con $\boldsymbol{y}_{\text{test}}$ por medio de una métrica apropiada.

Modelo 1: 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 2: Regresión Lasso

Distribución previa:

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

con $\tau_0 = 5$, $\nu_0 = 1$ y $\sigma_0^2 = \hat{\sigma}^2_{\text{OLS}}$, donde $\mathsf{DE}(\mu, \sigma)$ es la distribución Doble Exponencial (distribución de Laplace) con media μ y varianza $2\sigma^2$.

Nota: Si $\theta \mid \phi^2 \sim \mathsf{N}(\mu, \phi^2)$ y $\phi^2 \sim \mathsf{G}(1, \frac{1}{2\tau^2})$, entonces $\theta \sim \mathsf{DE}(\mu, \tau)$. Incluir la variable auxiliar ϕ^2 permite desarrollar un muestreador de Gibbs más sencillo de implementar dado que todas las distribuciones condicionales completas tienen forma cerrada (la distribución condicional completa de $1/\phi^2$ es Gaussiana Inversa).

Modelo 3: 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 $b_\rho = 1$.

Modelo 4: Regresión con selección automática de covariables

Distribución muestral:

$$\boldsymbol{y} \mid \mathbf{X}, \boldsymbol{b}, \boldsymbol{z}, \sigma^2 \sim \mathsf{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)\,,$$

donde $\boldsymbol{b} = (b_1, \ldots, b_p) \in \mathbb{R}^p$, $\boldsymbol{z} = (z_1, \ldots, z_p) \in \{0, 1\}^p$ y $\boldsymbol{\beta} = (b_1 \cdot z_1, \ldots, b_p \cdot z_p)$.

Distribución previa:

$$p(\boldsymbol{b}, \boldsymbol{z}, \sigma^2) = \mathsf{N}(\boldsymbol{b}_z \mid \mathbf{0}_{p_z}, n\, \sigma^2 (\mathbf{X}_z^\mathsf{T} \mathbf{X}_z)^{-1}) \cdot \mathsf{GI}(\sigma^2 \mid \nu_0/2, \nu_0\, \sigma_0^2/2) \cdot \prod_{j=1}^{p} \mathsf{Ber}(z_j \mid p_0)\,,$$

con $g = n$, $\nu_0 = 1$, $\sigma_0^2 = \hat{\sigma}^2_{\text{OLS}}$ y $p_0 = 0{,}5$, donde $p_z = \sum_{j=1}^{p} z_j$, $\boldsymbol{b}_z$ es un vector de $p_z \times 1$ conformado por las entradas de β tales que $z_j = 1$ y $\mathbf{X}_z$ es una matriz de $n \times p_z$ correspondiente a la matriz de diseño asociada con las entradas de $\boldsymbol{\beta}$ tales que $z_j = 1$.

Preguntas

Ajustar cada modelo utilizando los datos de entrenamiento $(\boldsymbol{y}_{\text{train}}, \mathbf{X}_{\text{train}})$.
Nota: Incluir un anexo describiendo los detalles del ajuste de cada modelo.

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.

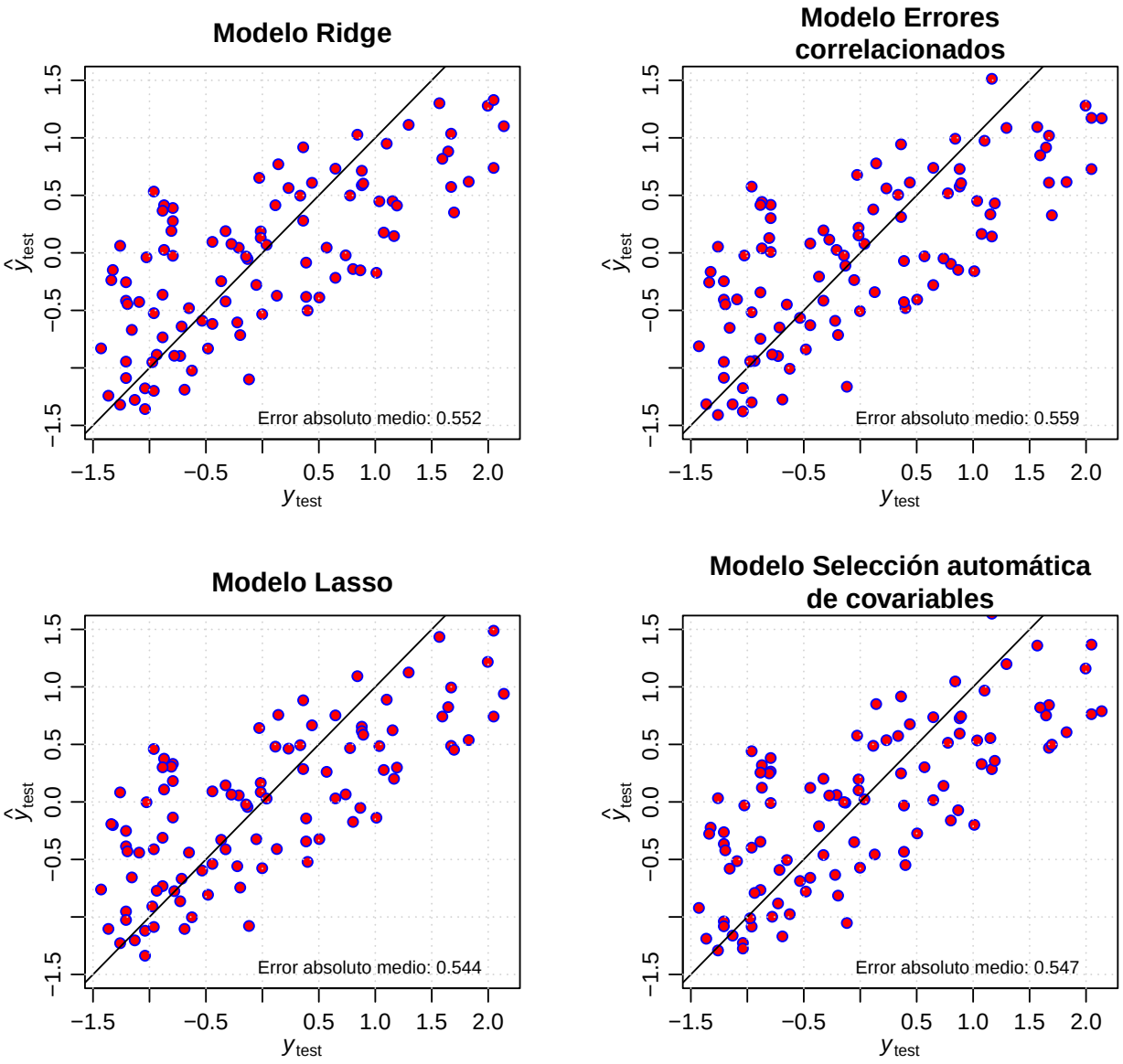


Figura 1: Desempeño de los modelos bayesianos mediante la graficación de $\hat{y}_{test}$ vs y_{test}

Los 4 modelos tuvieron resultados similares en cuanto al error absoluto medio, puesto que sus respectivos valores son muy similares, se evidencia además, la similaridad de las estimaciones del progreso de la enfermedad, ya que la distribución de estas son parecidas en todos los modelos, además de seguir la tendencia de la recta $\hat{y}_{test} = y_{test}$ en los valores de y_{test} entre -1.5 y 0.5, después de estos valores, todos los modelos tuvieron dificultades para que sus estimaciones replicaran el valor verdadero del progreso de la enfermedad; estas fueron menores al verdadero

valor. Estos resultados fueron bastante similares debido a que las estimaciones del vector de parámetros β fueron las siguientes.

Regresión	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$	$\hat{\beta 10}$
Frecuentista	-0.008	-0.134	0.343	0.182	-0.736	0.475	0.176	0.162	0.502	0.070
Ridge	-0.006	-0.132	0.345	0.183	-0.430	0.228	0.044	0.130	0.388	0.070
Lasso	-0.008	-0.133	0.345	0.183	-0.532	0.310	0.089	0.143	0.426	0.071
Errores Cor.	-0.006	-0.134	0.346	0.179	-0.736	0.475	0.176	0.163	0.497	0.075
Selección Aut.	0.000	-0.089	0.358	0.182	-0.232	0.120	-0.058	0.033	0.360	0.007

Donde, sin importar el modelo, incluso el frecuentista, las estimaciones $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_8$ y $\hat{\beta}_{10}$ fueron semejantes.

En la siguiente tabla se registran los intervalos de credibilidad del 95 % de los modelos Ridge, Lasso, Errores Correlacionados y Selección Automática

Modelo	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
Ridge	(-0.088 , 0.076)	(-0.216 , 0.047)	(0.257 , 0.437)	(0.093 , 0.272)	(-0.921 , 0.024)	(-0.158 , 0.641)	(-0.196 , 0.291)	(-0.085 , 0.350)	(0.187 , 0.594)	(-0.022 , 0.162)
Lasso	(-0.090 , 0.074)	(-0.220 , -0.049)	(0.252 , 0.438)	(0.092 , 0.272)	(-1.086 , -0.021)	(-0.118 , 0.769)	(-0.177 , 0.367)	(-0.083 , 0.369)	(0.208 , 0.658)	(-0.022 , 0.163)
Errores Cor.	(-0.088 , 0.076)	(-0.217 , -0.048)	(0.255 , 0.436)	(0.086 , 0.270)	(-1.341 , -0.136)	(-0.023 , 0.977)	(-0.118 , 0.474)	(-0.068 , 0.390)	(0.253 , 0.744)	(-0.021 , 0.169)
Selección Aut.	(-0.001 , 0.000)	(-0.200 , 0.000)	(0.263 , 0.453)	(0.087 , 0.271)	(-0.681 , 0.000)	(-0.107 , 0.518)	(-0.239 , 0.000)	(0.000 , 0.249)	(0.190 , 0.563)	(0.000 , 0.100)

En todos los modelos, β_3, β_4 y β_9 fueron estadísticamente significativos(5 %) mayores a 0 , puesto que sus intervalos de credibilidad no incluyen al 0. Por otro lado, los intervalos de credibilidad de $\beta_1, \beta_6, \beta_7$ y β_{10} sí incluyeron al 0 en todos los modelos, esto puede comprobarse con las distribuciones posteriores de cada β, ya que algunas de estas se encuentran centradas lejos del cero, mientras que otras cercanas.

β_3, β_4 y β_9 son las variables más importantes para modelar el progreso de la diabetes, la selección automática de covariables remarca este hecho.

2. Para cada modelo, chequear la bondad de ajuste usando como estadísticos de prueba de la media y la desviación estándar. Graficar la distribución predictiva posterior de ambos estadísticos simultáneamente por medio de un dispersograma con las muestras correspondientes.

   Las distribuciones predictivas posteriores en los 4 modelos son similares, en cada modelo se utilizaron 10000 muestras provenientes de los muestreadores de Gibbs correspondientes, con las que se evaluó la bondad de ajuste, teniendo así 10000 realizaciones de medias y desviaciones estándar en cada gráfico, la media y desviación estándar observada en los datos reales del progreso de la enfermedad es representada en cada gráfico con un cuadro negro, en todos estos gráficos la media y desviación estándar observadas no se encuentran del todo centrada en la nube de puntos, demostrado con los p-valores predictivos posteriores. La media especialmente contó con valores alrededor de 0.6 en los 4 modelos, por lo que son modelos que no representan la media de la mejor manera, no obstante, modelos como el Ridge, Lasso y selección automática de covariables contaron con buenos ppp para representar la desviación estándar, esto debido a que sus distribuciones muestrales contaban con parámetros de variabilidad más sencillos de modelar.

   Esta bondad de ajuste fue realizada utilizando los datos de entrenamiento, se observó una media de -0.003 y desviación estándar de 1 aproximadamente en los 342 individuos de entrenamiento, por otro lado, los 100 individuos de testeo del modelo reportaron una media de 0.011 y desviación estándar de 0.987 aproximadamente en el avance de la enfermedad. Inesperadamente, estos modelos realizados con los datos de entrenamiento no lograron replicar adecuadamente la media de los datos de entrenamiento, sin embargo, sí lo lograron con la media de los datos de testeo, puesto que sus ppp fueron 0.461, 0.479, 0.494 y 0.45 para el modelo 1, 2, 3 y 4 respectivamente; y, al contrario que la desviación de los datos de entrenamiento, este modelo no replica apropiadamente la desviación estándar de los datos de testeo.

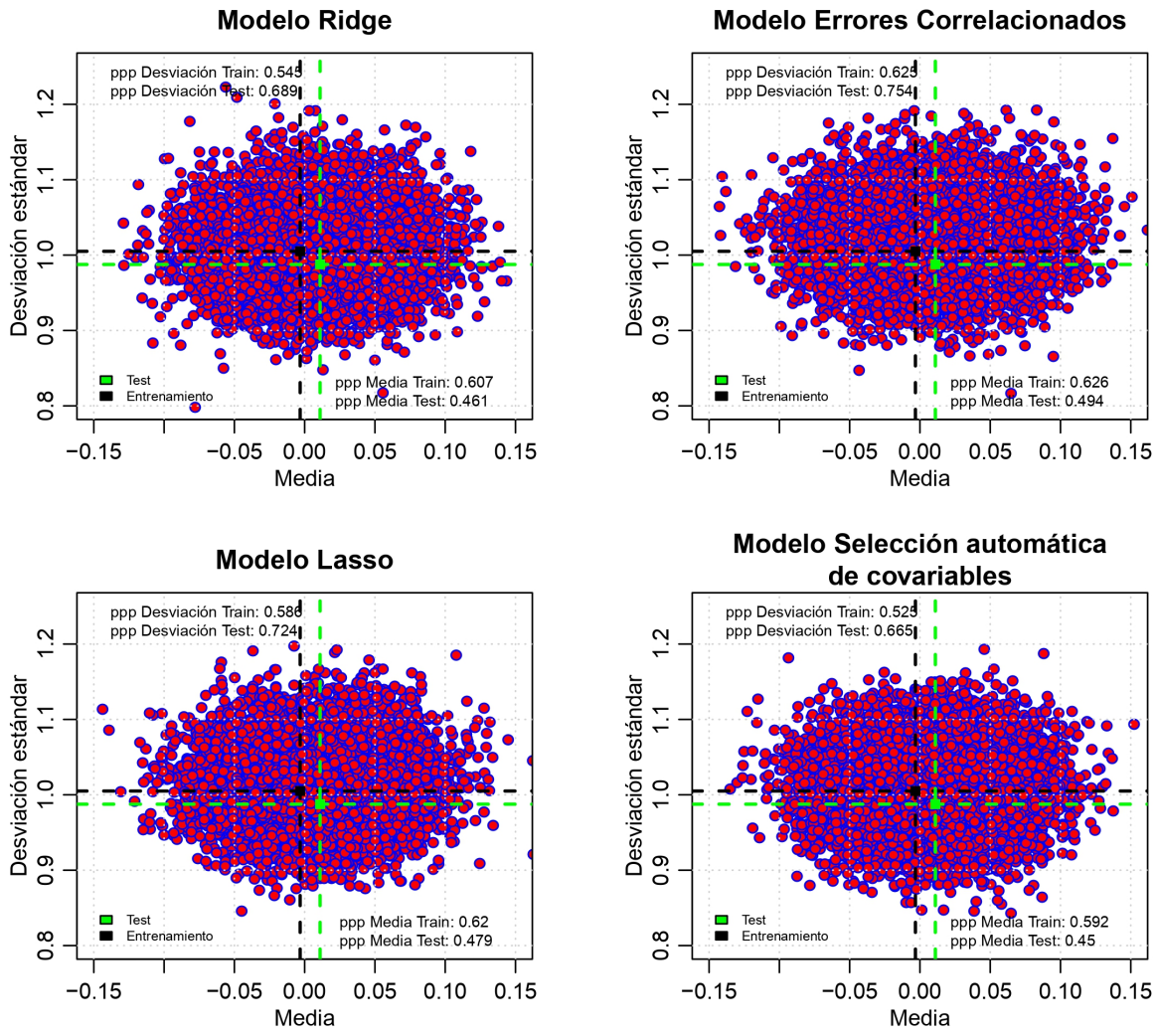


Figura 2: Distribución predictiva posterior de los estadísticos media y desviación estándar.

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

Regresión	Ridge	Lasso	Errores Correlacionados	Selección Automática
DIC	748.575	748.535	749.932	753.330

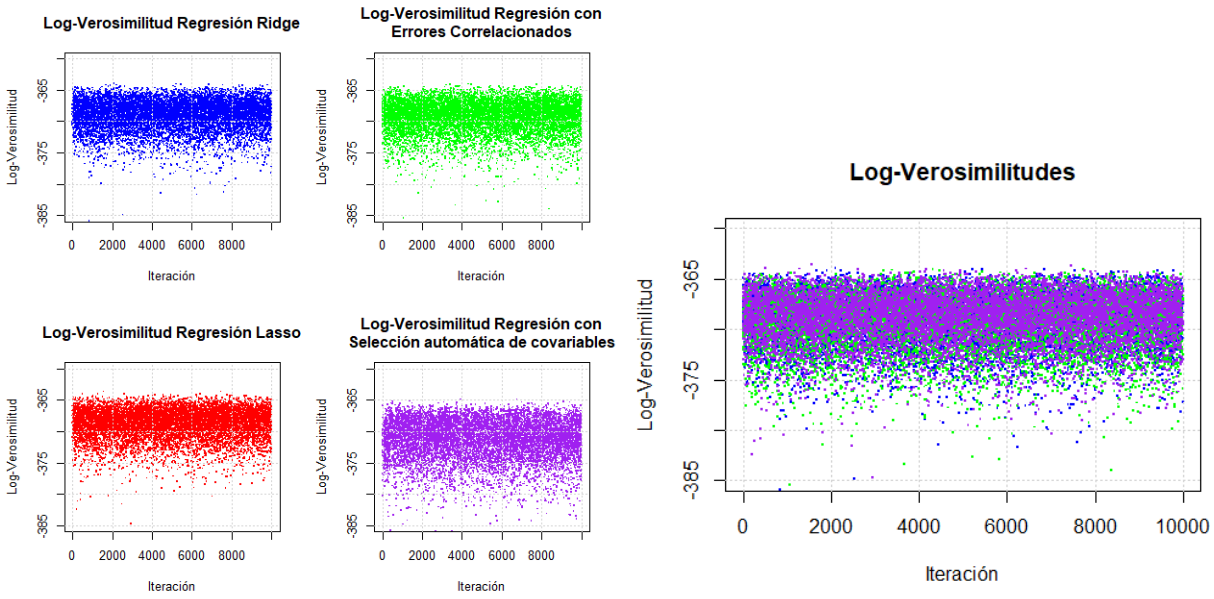
Al igual que en los anteriores apartados, los 4 modelos tuvieron criterios de información DIC similares, el modelo Lasso fue aquel que contó con el más bajo, por lo que podría ser la elección adecuada para modelar el progreso de la enfermedad diabetes con las 10 medidas basales.

En conclusión, los 4 modelos presentaron resultados similares en casi todos los apartados, por lo que los modelos adecuados para representar esta situación son aquellos con menor carga computacional, recordemos que los modelos 3 y 4 requieren de la inversión de matrices y muestreo de más variables auxiliares respectivamente; por lo tanto las alternativas más efectivas para este caso de la enfermedad diabetes, son los modelos Lasso y Ridge.

ANEXO

1. Ajuste de los modelos de regresión:

- **Convergencia:** Los 4 modelos contaron con Logverosimilitudes alrededor de -370, la convergencia fue inmediata, por lo que no se requirió un periodo de calentamiento extenso, en los siguientes gráficos se puede observar las logverosimilitudes marginales y todas superpuestas.

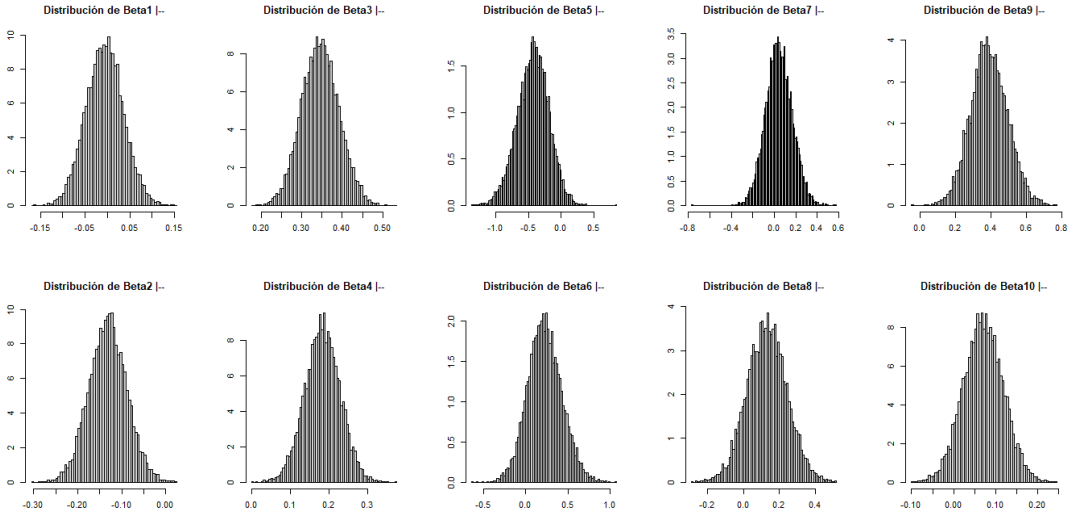


- **Tamaños efectivos de muestra:**

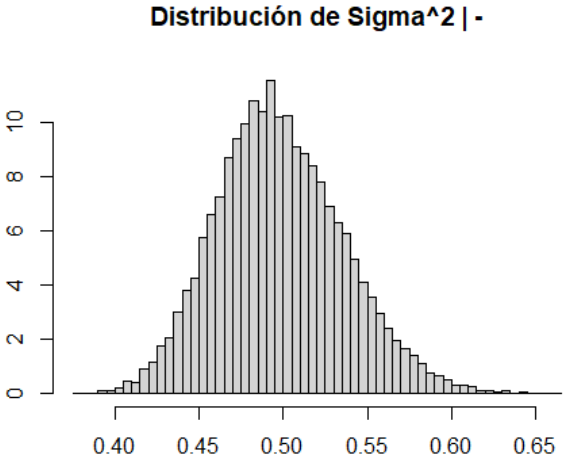
Modelo	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	σ
Ridge	10000	10000	10000	10000	8188	8436	8599	10000	8552	10000	10035
Lasso	10000	10000	10000	10304	17379	16769	14397	11322	15772	10000	9400
Errores Cor.	10337	10000	10000	10000	10000	10000	10328	10000	9587	10000	10000
Selección Aut.	10000	7254	9465	9483	1758	2112	1644	2895	2240	10488	10028

Todos los modelos fueron realizados con el fin de obtener 10000 muestras, la mayoría de parámetros obtuvieron buenos tamaños efectivos de muestras, incluso, por errores de la función *effectivesize* de la librería coda, reportaron mayores a esos 10000, algunos parámetros del modelo 4 reportaron bajos tamaños efectivos de muestra, esto se debe a la cantidad de ceros resultantes por la misma definición de las distribuciones previas del modelo.

- **Distribuciones condicionales completas de los parámetros de interés:**
  - **Modelo Ridge:**
    - β: $\beta \sim \mathsf{N}_{10}\left([\frac{1}{\sigma^2}(\mathbb{X}^T\mathbb{X}+\lambda I_{10})]^{-1}[\frac{1}{\sigma^2}\mathbb{X}^T\mathbb{Y}], [\frac{1}{\sigma^2}(\mathbb{X}^T\mathbb{X}+\lambda I_{10})]^{-1}\right)$



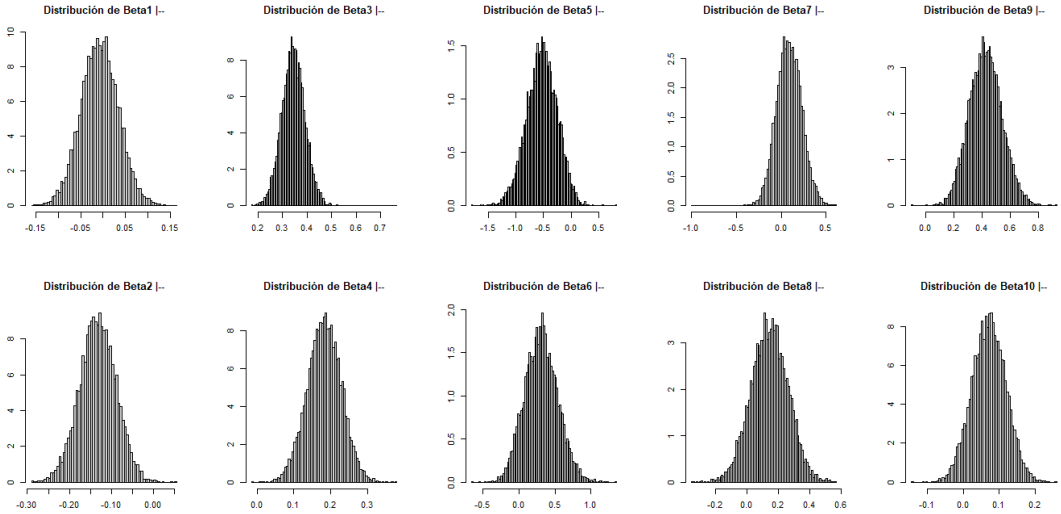
- σ^2: $\sigma^2 \sim \mathsf{GI}\left(\frac{n+10+\nu_0}{2}, \frac{\nu_0\sigma_0^2+\lambda\beta^T\beta+\|\mathbb{Y}-\mathbb{X}\beta\|^2}{2}\right)$



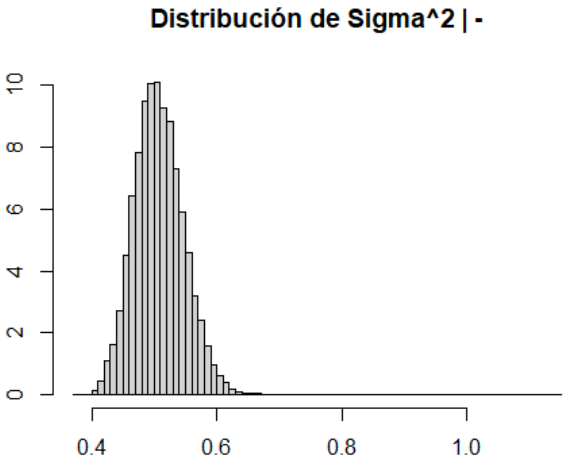
- **Modelo Lasso:**
  - β: $\beta \sim \mathsf{N}_{10}\left([\frac{1}{\sigma^2}\mathbb{X}^T\mathbb{X}+\Phi_{-1}]^{-1}[\frac{1}{\sigma^2}\mathbb{X}^T\mathbb{Y}], [\frac{1}{\sigma^2}\mathbb{X}^T\mathbb{X}+\Phi_{-1}]^{-1}\right)$

    Donde :

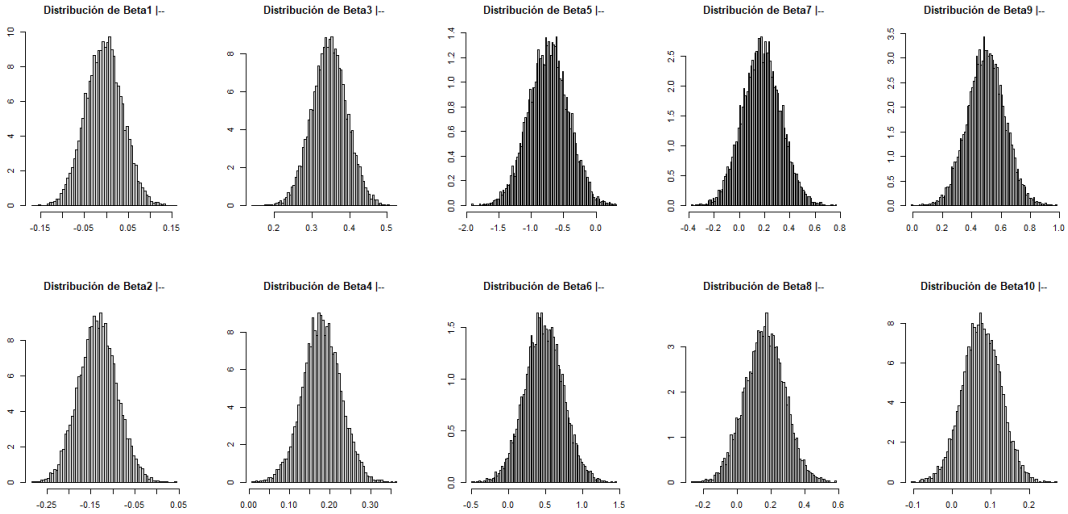
$$\Phi_{-1} = \begin{bmatrix} \frac{1}{\phi_1^2} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{\phi_2^2} & 0 & \cdots & 0 \\ 0 & 0 & \frac{1}{\phi_3^2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{\phi_{10}^2} \end{bmatrix}$$



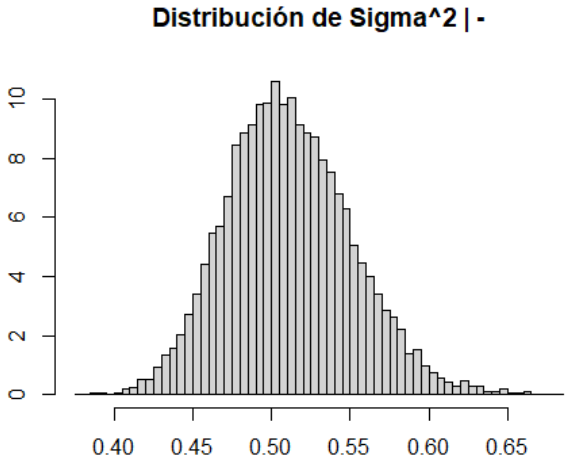
- σ^2: $\sigma^2 \sim \mathsf{GI}\left(\frac{\eta+\nu_0}{2}, \frac{\nu_0\sigma_0^2+\|\mathbb{Y}-\mathbb{X}\beta\|^2}{2}\right)$



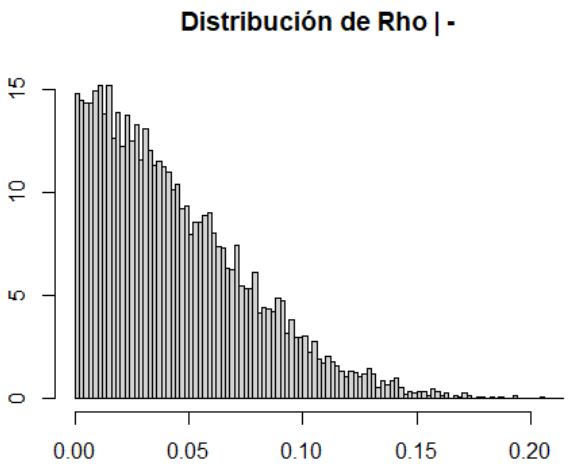
- **Modelo Errores Correlacionados:**
  - β: $\beta \sim \mathsf{N}_{10}\left([\frac{1}{\sigma^2}\mathbb{X}^T C_\rho^{-1}\mathbb{X} + \frac{1}{\tau_0^2}I_{10}]^{-1}[\frac{1}{\sigma^2}\mathbb{X}^T C_\rho^{-1}\mathbb{Y}], [\frac{1}{\sigma^2}\mathbb{X}^T C_\rho^{-1}\mathbb{X} + \frac{1}{\tau_0^2}I_{10}]^{-1}\right)$



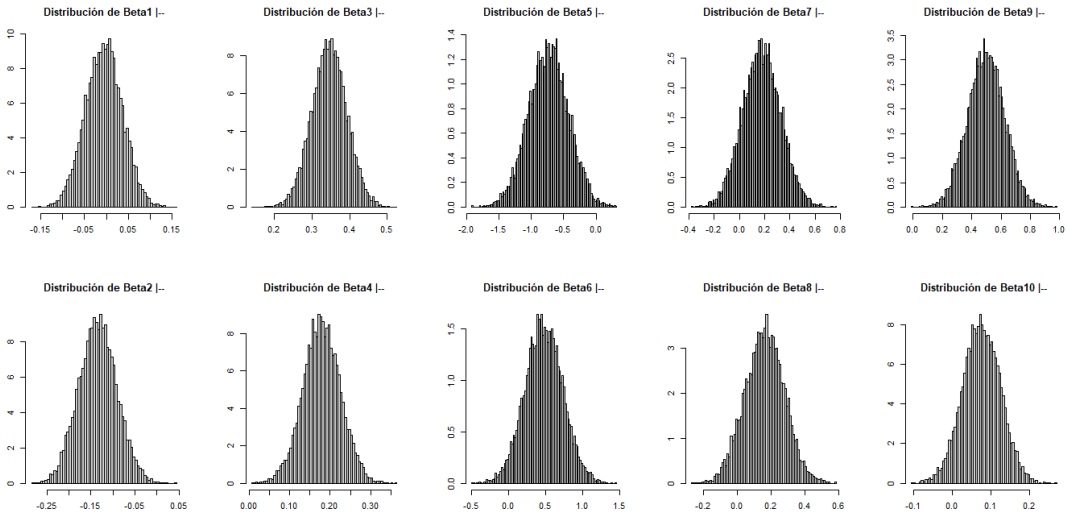
- σ^2:
  $\sigma^2 \sim \mathsf{GI}\left(\frac{n+\nu_0}{2}, \frac{\nu_0\sigma_0^2+(\mathbb{Y}-\mathbb{X}\beta)^T C_\rho^2(\mathbb{Y}-X\beta)}{2}\right)$



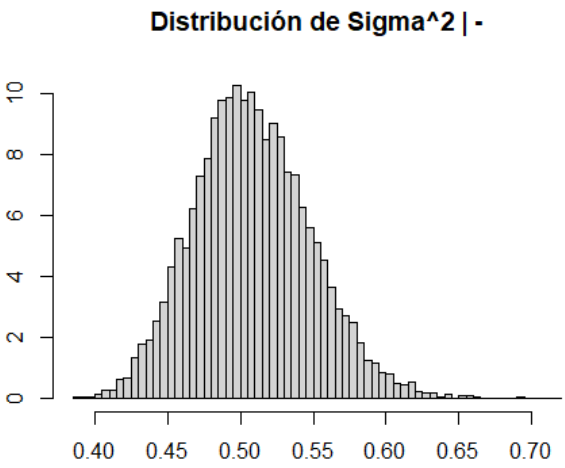
- ρ: $P(\rho|-) \propto |C_\rho|^{-\frac{1}{2}} exp\{-\frac{1}{2\sigma^2}(\mathbb{Y}-\mathbb{X}\beta)^T C_\rho^{-1}(\mathbb{Y}-\mathbb{X}\beta)\}$



- **Modelo Selección Automática de Covariables:**
  - β:



  - σ^2:



- **Autocorrelación:**

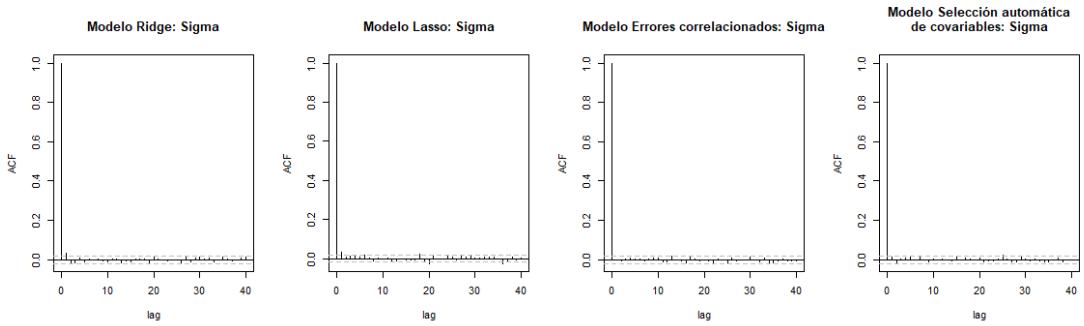


Figura 3: Gráficos de autocorrelación deσ^2

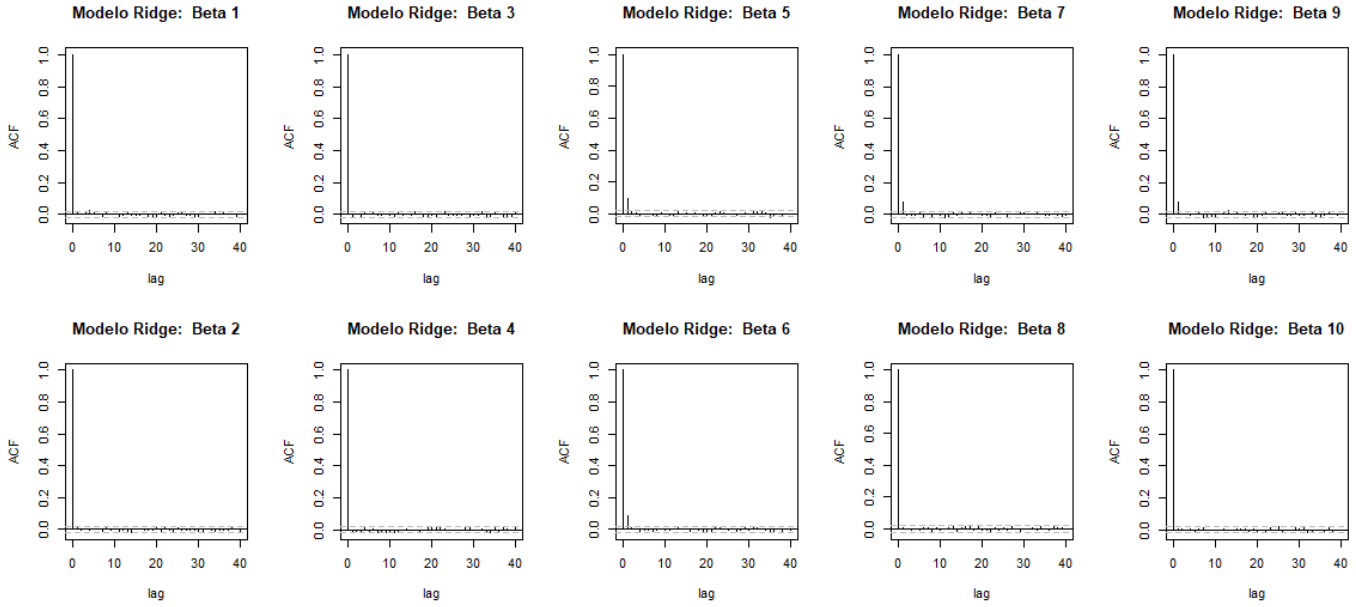


Figura 4: Gráficos de autocorrelación de Betas Modelo Ridge

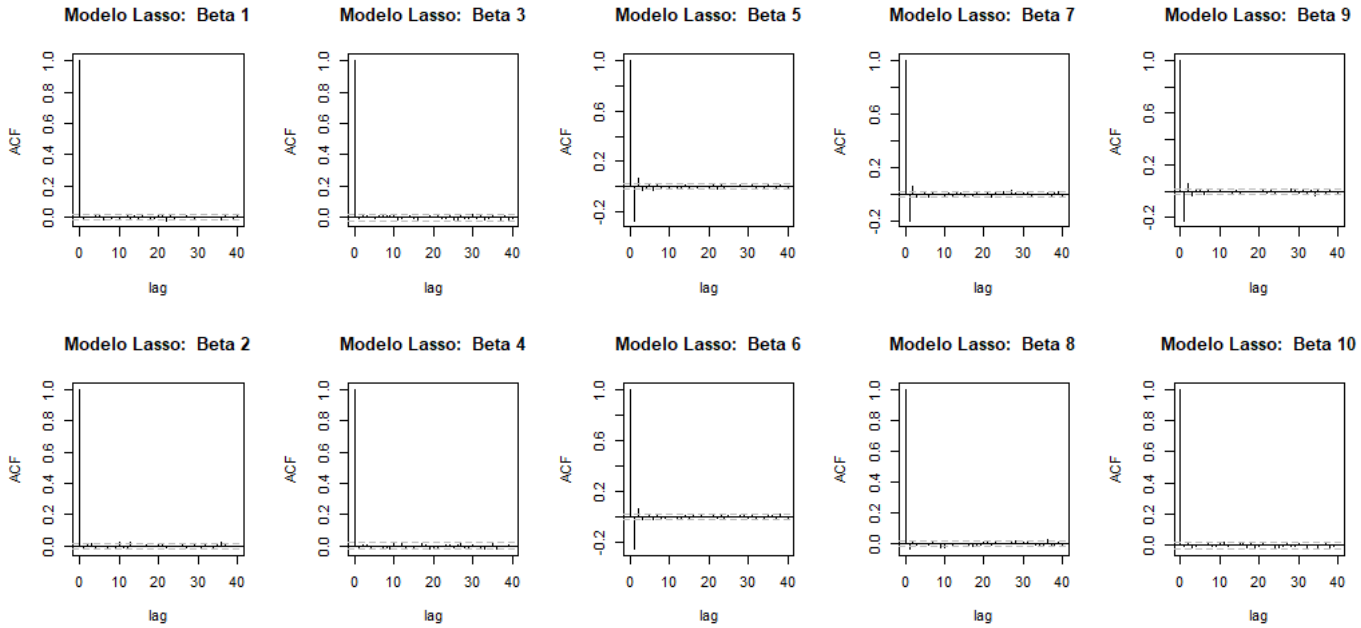


Figura 5: Gráficos de autocorrelación de Betas Modelo Lasso

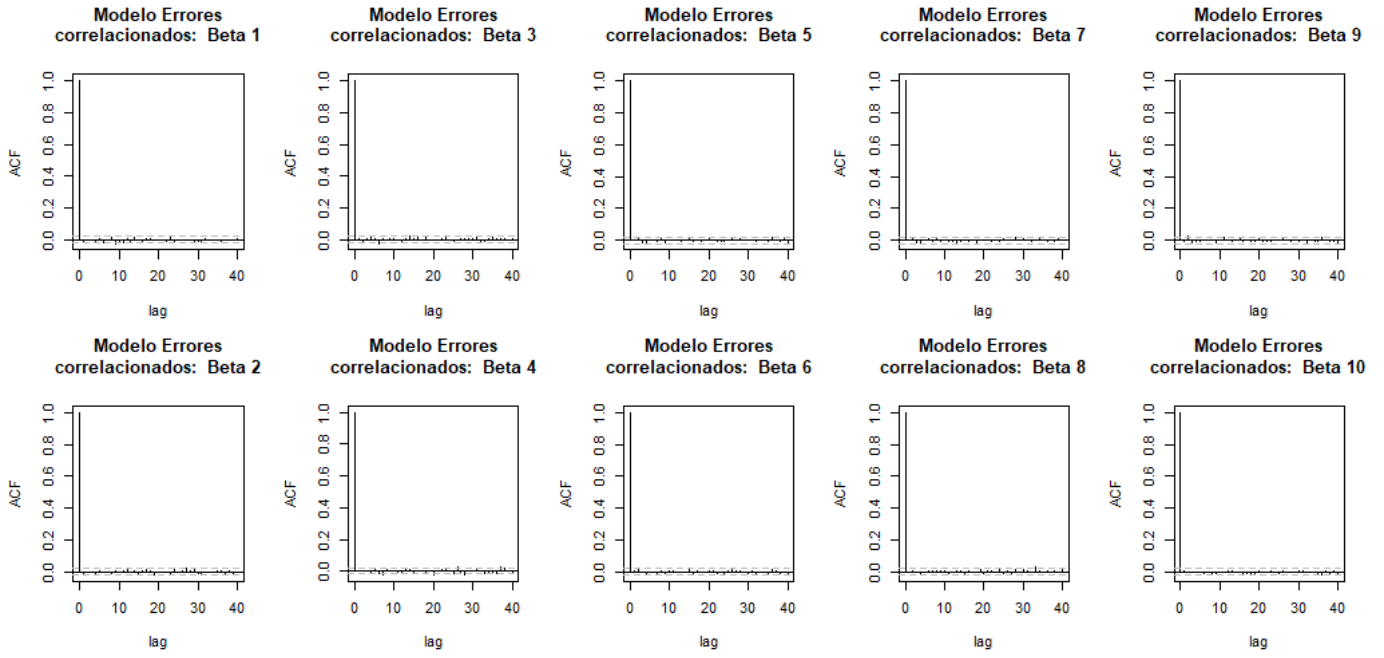


Figura 6: Gráficos de autocorrelación de Betas Modelo Errores Correlacionados

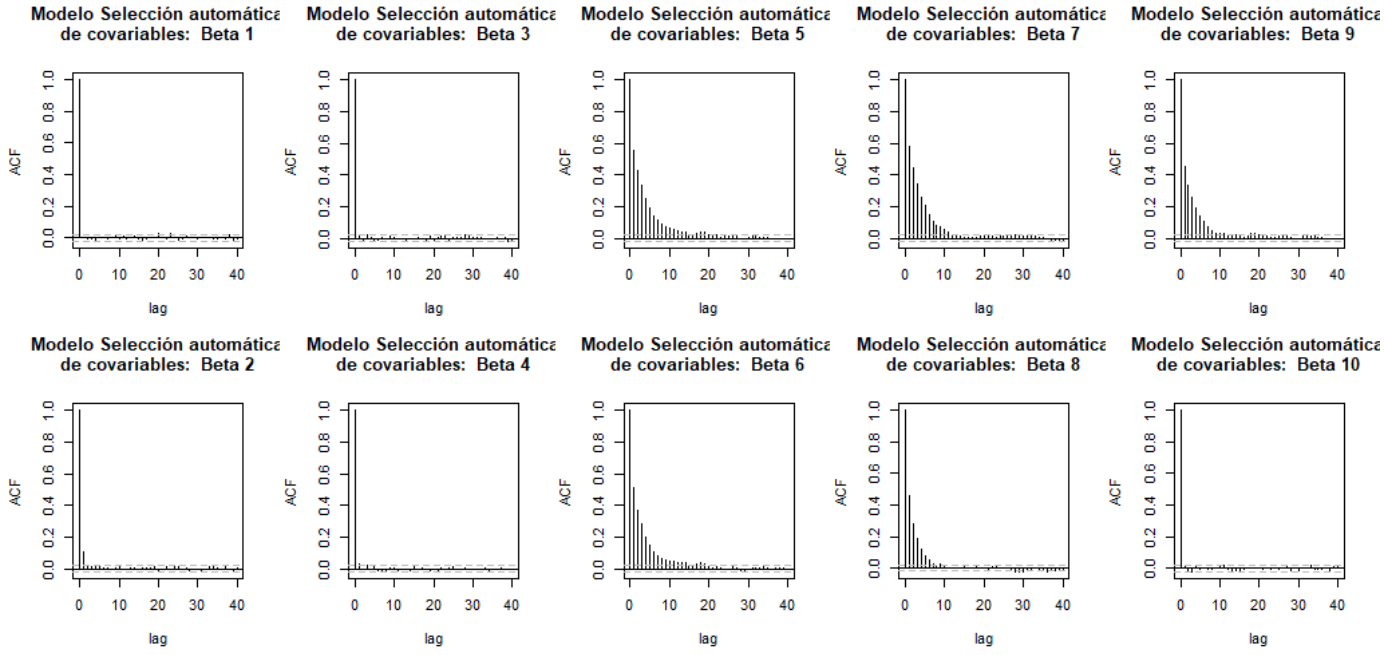


Figura 7: Gráficos de autocorrelación de Betas Modelo Selección automática de covariables

Las únicas variables que mostraron autocorrelación positiva, fueron los $\beta_5, \beta_6, \beta_7, \beta_8$ y β_9 del modelo Selección automática de covariables.