

Pronóstico de contrataciones de préstamos bancarios

Abraham Nieto 51556, Alejandro Hernández 87806 y Omar Reyes 127131

Diciembre 2018

Índice

Introducción	3
Descripción del problema	3
Objetivos	3
Descripción de la información	3
Variables	3
Segmentos	4
Análisis exploratorio de datos	4
Totales	5
Por segmento	7
Variables Numéricas	11
Modelado e implementación	13
Modelos Estáticos	13
Modelo lineal normal	14
Modelo lineal generalizado poisson con liga log	15
Modelo dinámico	16
Interpretación de modelos	19
Modelos estáticos	19
Modelo lineal normal	19
Modelo lineal generalizado poisson con liga log	20
Modelo dinámico	21
Prediccion	26
Conclusiones	27
Referencias	28
Apéndice	29
Gráficas comportamiento cadenas y parámetros	29
Código JAGS	49

Introducción

Es bien sabido que el otorgamiento de créditos es uno de los negocios más rentables para los bancos, es por ello que estos asignan mucho presupuesto a la generación de campañas para su colocación, dirigidas en primera instancia a clientes no sólo para acrecentar el negocio, también para incrementar su fidelidad con el banco. De igual manera, habrá campañas para personas que aún no son clientes con la finalidad de atraerlos hacia el banco.

Para poder analizar el comportamiento de las contrataciones en un banco, se cuenta con una muestra de 64 mil clientes con información de dos años, de tal suerte que hallaremos algunas contrataciones del producto ya sea por campaña o por interés auténtico por parte de la persona.

Nuestro primer objetivo es realizar un pronóstico de la cantidad de créditos que colocaremos en períodos posteriores, de tal modo que podamos utilizarlo como referencia para establecer una meta al área de colocación de créditos. El segundo objetivo será tener la capacidad de crear campañas personalizadas en cuanto a la comunicación de la oferta, esto será posible por toda la información con la que contamos: saldos, gastos, créditos, productos bancarios, movimientos, etc. Tenemos la hipótesis de que esta comunicación se puede diferenciar por una variable adicional: Etapa de vida del cliente. Por ejemplo una persona que es el sustento de una familia no tiene las mismas necesidades que una persona joven que está comenzando su vida laboral, así como alguien que ya está jubilado. Queremos analizar si realmente esta variable es un diferenciador en cuanto a la tasa de contratación.

Descripción del problema

Se quiere establecer un proceso mensual, mediante un modelo, que pronostique el número de contrataciones de préstamos personales con el fin de establecer la meta mensual de colocaciones de crédito a través de campañas. Para ello, contamos con 24 meses de historia (Agosto 2016-Julio 2018), por cada mes tenemos el sumarizado de nuestros datos agrupado por ciclo de vida. La variable target representa el número de préstamos contratados en cada uno de los ciclos de vida por mes, además se cuenta con 5 variables adicionales que será necesario validar que tengan o no influencia en el pronóstico.

Objetivos

- Construir un modelo que pronostique ,puntualmente y por intervalo, el número de préstamos que serán contratados en los siguientes 3 meses y evaluar su desempeño.
- Determinar si existe diferencia significativa en las contrataciones de préstamos por ciclo de vida, es decir, si existe algún o algunos segmentos que tengan mayor influencia en el número de contrataciones totales con el fin de identificar si las estrategias debieran ir dirigidas a ciertos segmentos, además de diferenciar la comunicación.

Descripción de la información

Variables

Las variables presentes en la base de datos son las siguientes:

- **TARGET:** Número de creditos contratados. Es un entero entre 13 y 150. Esta es nuestra variable de respuesta y la renombraremos como y
- **ID:** Número de clientes.
- **FH_REF:** Fecha de referencia, inicia el 01-08-2016 y finaliza el 01-07-2018.

- **TP_SEGMENTO_FINAL2:** Ciclo de vida de los clientes (ADULTO EN PLENITUD, ADULTOS INDEPENDIENTES, DIVORCIADO, HOGARES CON HIJOS, JOVEN PROFESIONAL, JOVEN TRABAJADOR, PAREJA JOVEN, PAREJAS ADULTAS y PAREJAS SENIOR). La renombraremos como x_1 .
- **TO_PROM_TO_CARGOS_3M:** Número promedio de cargos en los últimos 3 meses. La renombraremos como x_2 .
- **NU_VINC_COGNODATA:** Número promedio de productos bancarios. La renombraremos como x_3 .
- **TO_NECESIDAD_FINAN_CAP_3M:** Cociente del gasto promedio y el saldo. La renombraremos como x_4 .
- **IM_PROM_GASTOS_3M:** Importe promedio de gastos (compras+cargos recurrentes) en los últimos 3 meses. La renombraremos como x_5 .
- **IM_SUM_SDO_CORTE_1M:** Importe promedio del saldo de cuenta de cheques. La renombraremos como x_6 .

Segmentos

Las etapas en el ciclo de vida de nuestros clientes y su descripción son las siguientes:

JOVEN TRABAJADOR Clientes entre 18 a 34 años. Obreros calificados. Con ingresos de \$3,300 a \$11,000 mensuales.

JOVEN PROFESIONISTA Clientes entre 18 a 34 años. Joven cuya profesión requiere una calificación superior para ser desempeñada. Con ingresos superiores a los \$3,200 mensuales.

PAREJA JOVEN Clientes entre 18 a 34 años. Sin hijos y conviven en pareja. Con ingresos superiores a los \$5,000 mensuales.

HOGARES CON HIJOS Clientes entre 18 a 65 años. Conviven con sus hijos en el hogar. Con ingresos promedio de \$25,000 mensuales.

PAREJA ADULTA Clientes entre 35 a 44 años. Sin hijos en el hogar, que conviven en pareja. Con ingresos superiores a los \$7,000 mensuales.

PAREJA SENIOR Clientes entre 45 a 65 años. Sin hijos en el hogar, que conviven en pareja. Con ingresos superiores a los \$7,000 mensuales.

DIVORCIADO Clientes entre 35 a 65 años. Divorciados.

ADULTO INDEPENDIENTE Clientes entre 35 a 65 años. Sin hijos en el hogar, no conviven en pareja. Con ingresos superiores a \$6,000 mensuales.

ADULTO EN PLENITUD Clientes mayores a 65 años.

Análisis exploratorio de datos

Veamos un breve resumen sobre algunas estadísticas relevantes de cada variable

```
summary(camp)
```

```
##          FH_REF           TP_SEGMENTO_FINAL2      target
## 2016-08-01: 9    ADULTO EN PLENITUD :24      Min.   : 13.00
## 2016-09-01: 9    ADULTOS INDEPENDIENTES:24    1st Qu.: 37.00
## 2016-10-01: 9    DIVORCIADO              :24    Median : 49.00
## 2016-11-01: 9    HOGARES CON HIJOS       :24    Mean   : 53.48
```

```

## 2016-12-01: 9 JOVEN PROFESIONAL :24      3rd Qu.: 64.00
## 2017-01-01: 9 JOVEN TRABAJADOR :24      Max.    :150.00
## (Other)   :162 (Other)       :72
##           ID TO_PROM_TO_CARGOS_3M NU_VINC_COGNODATA
## Min.    :1939 Min.    : 2.756     Min.    :1.571
## 1st Qu.:3666 1st Qu.: 6.292     1st Qu.:2.327
## Median :4514 Median : 7.647     Median :2.713
## Mean   :4766 Mean   : 7.483     Mean   :2.744
## 3rd Qu.:5559 3rd Qu.: 9.010     3rd Qu.:3.063
## Max.   :8436  Max.   :12.964     Max.   :4.203
##
## TO_NECESIDAD_FINAN_CAP_3M IM_PROM_GASTOS_3M IM_SUM_SDO_CORTE_1M
## Min.    :-1367.4      Min.    : 24058     Min.    : 828293
## 1st Qu.: 952.7      1st Qu.:207558    1st Qu.:1730313
## Median : 2523.3     Median :282743    Median :1923435
## Mean   : 17539.3     Mean   :327599    Mean   :1981975
## 3rd Qu.:12378.9     3rd Qu.:409213    3rd Qu.:2305378
## Max.   :843804.7     Max.   :834860    Max.   :3364973
##
## cat
## Min.  :1
## 1st Qu.:3
## Median :5
## Mean   :5
## 3rd Qu.:7
## Max.   :9
##

```

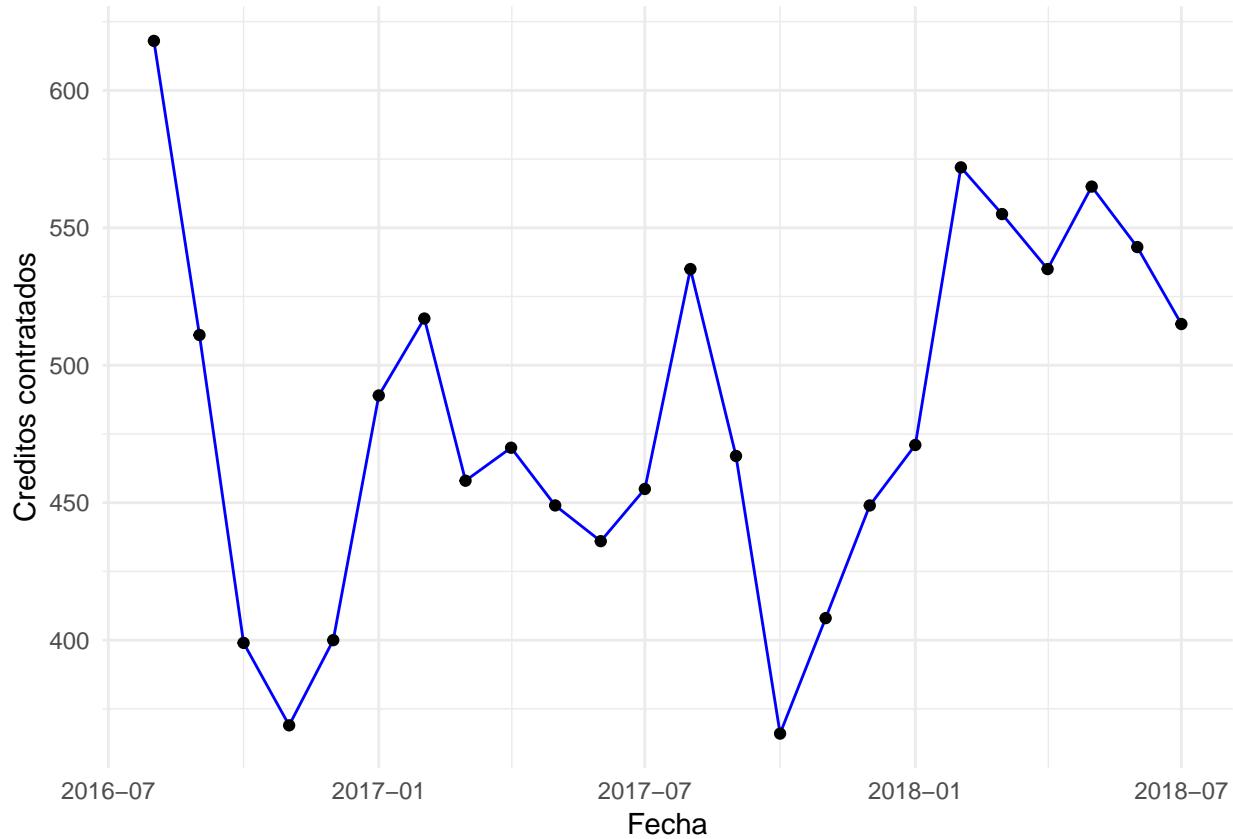
Como podemos observar, las variables x_2, \dots, x_6 están en escalas con distintos órdenes de magnitud, por lo que las reescalamos.

Es importante destacar que como la variable x_1 es una variable categórica tenemos que incorporar la restricción para sus coeficientes asociados, lo cual se detallará en cada uno de los modelos.

A continuación, se muestra un análisis exploratorio. Estudiaremos el comportamiento de las ventas contra el tiempo, para identificar tendencia o estacionalidad en los datos para los 9 segmentos diferentes que son objeto de análisis. De igual manera, analizaremos la relación que existe entre las diferentes variables. Además, con el fin de mejorar la tasa de contratación queremos estudiar si la etapa del ciclo de vida puede ser un diferenciador potente en cuanto al comportamiento de las colocaciones, desde la perspectiva de negocio esto ayudaría a hacer campañas cada vez más personalizadas en cuanto a la comunicación en el momento adecuado de la forma adecuada.

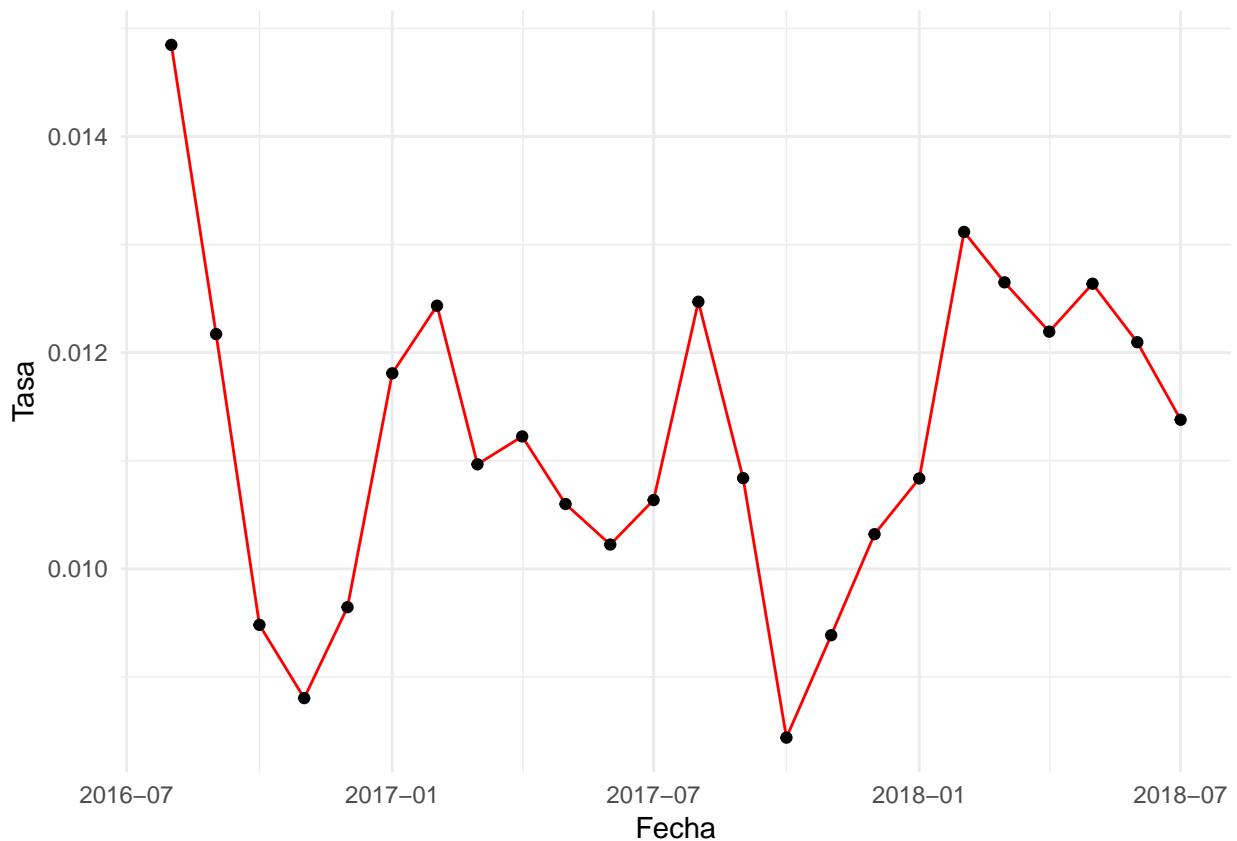
Totales

En la siguiente gráfica se observa el número de contrataciones totales por mes. Vemos que en agosto de 2016 tenemos el mayor número histórico de contrataciones (más de 600), debido a algunas estrategias comerciales en ese año. Por otro lado, aunque la estacionalidad no es precisamente mensual, sí se puede observar por períodos, por ejemplo, tenemos caídas muy claras en las contrataciones de los últimos meses del año, de octubre a diciembre, lo usual es que en este periodo de tiempo la gente tenga mayor liquidez. Notemos que, al comienzo de año, hay un repunte importante ya que normalmente los clientes están gastados y su necesidad de financiarse con un crédito aumenta. Del mismo modo, en el verano de cada año aproximadamente, siempre hay un número importante de contrataciones, esto puede ser debido al regresos a clases.



Ahora observemos las tasas de contratación a través del tiempo, donde en los últimos meses del año se tiene una tasa menor al 1%. Los meses con mayor número de contrataciones superan el 1.2%.

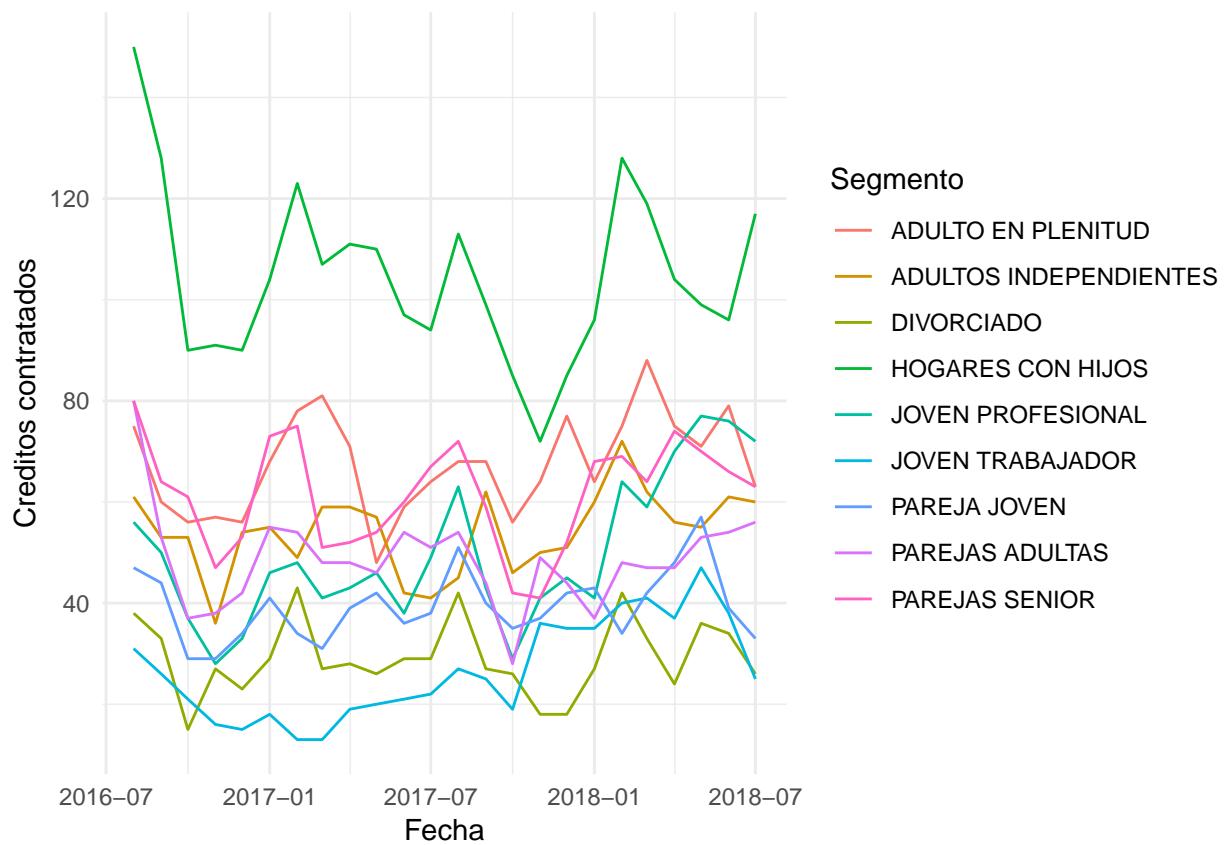
```
ggplot(tot,aes(x=as.Date(FH_REF),y=tasa)) + geom_line(col="red") +geom_point()+
  xlab("Fecha") +
  ylab("Tasa") +theme_minimal()
```



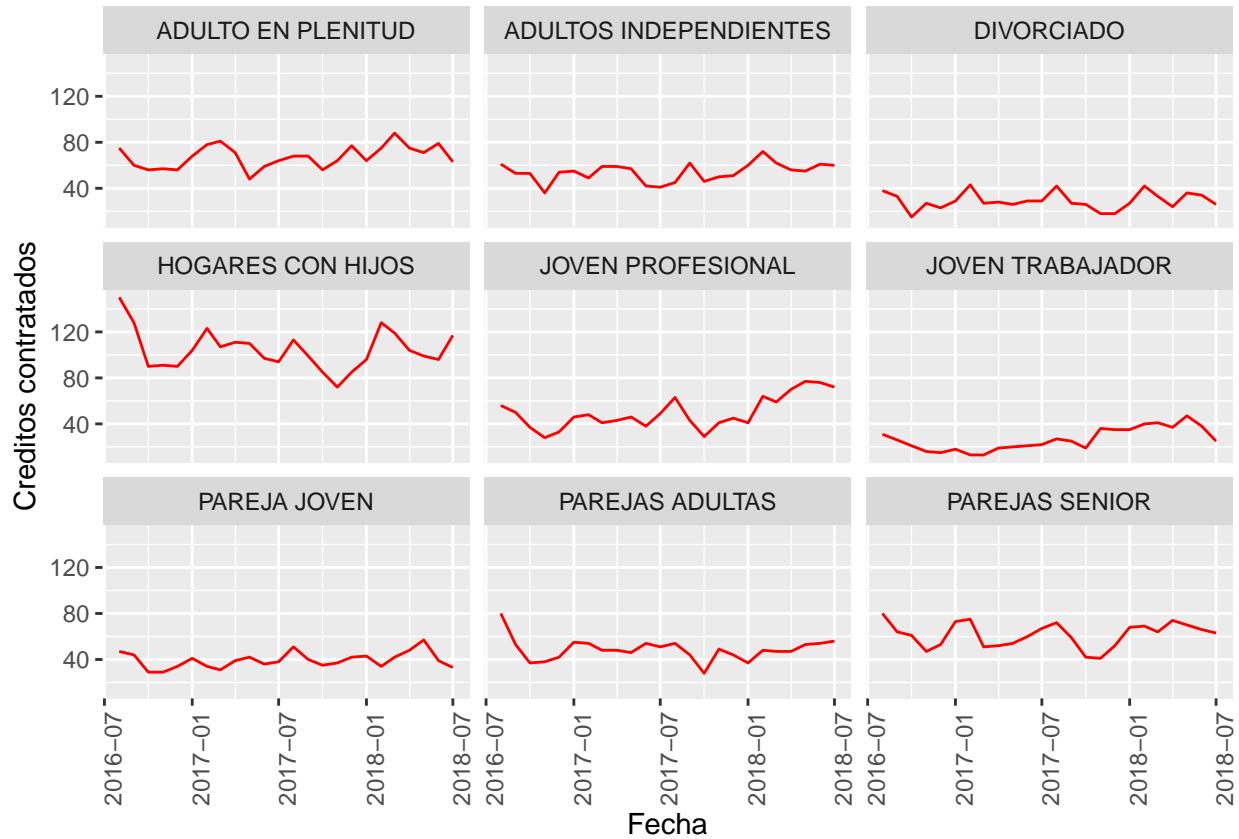
Por segmento

De forma similar, en la siguiente gráfica se puede observar la serie de número de créditos contratados por segmento.

En esta gráfica sobresalen los HOGARES CON HIJOS. En contraste con la tendencia general, vemos que los segmentos de JOVEN PROFESIONAL y JOVEN TRABAJADOR tuvieron una tendencia creciente constante a partir del 2018, la causa pudo ser que más jóvenes cumplieron los requisitos para adquirir préstamos.



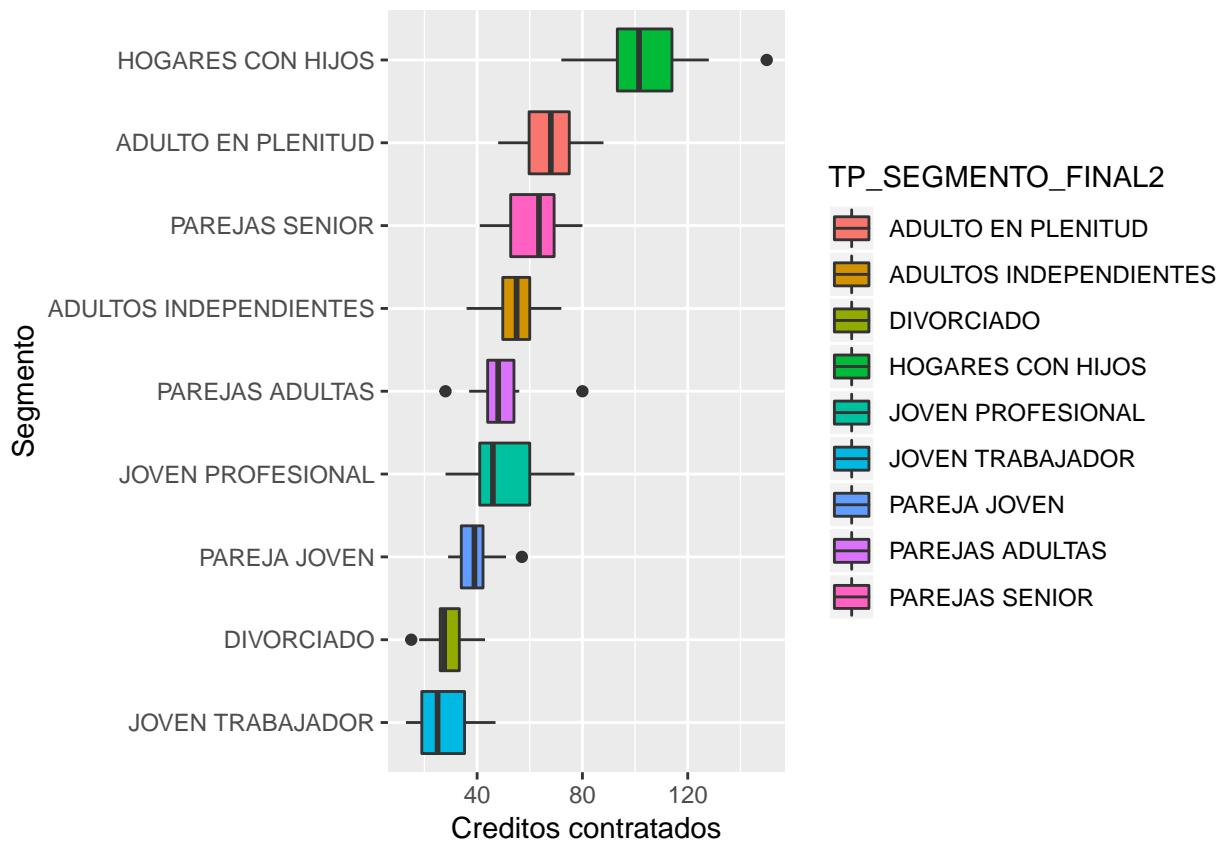
Y si separamos las tendencias:



En la gráfica anterior se muestra que el segmento de HOGARES CON HIJOS es el que tiene una tendencia más parecida a la total, es decir con más variaciones. Podemos pensar que este segmento tiene una influencia importante en las contrataciones con respecto al total por mes. En contraste, PAREJA JOVEN y DIVORCIADO muestran una tendencia más constante.

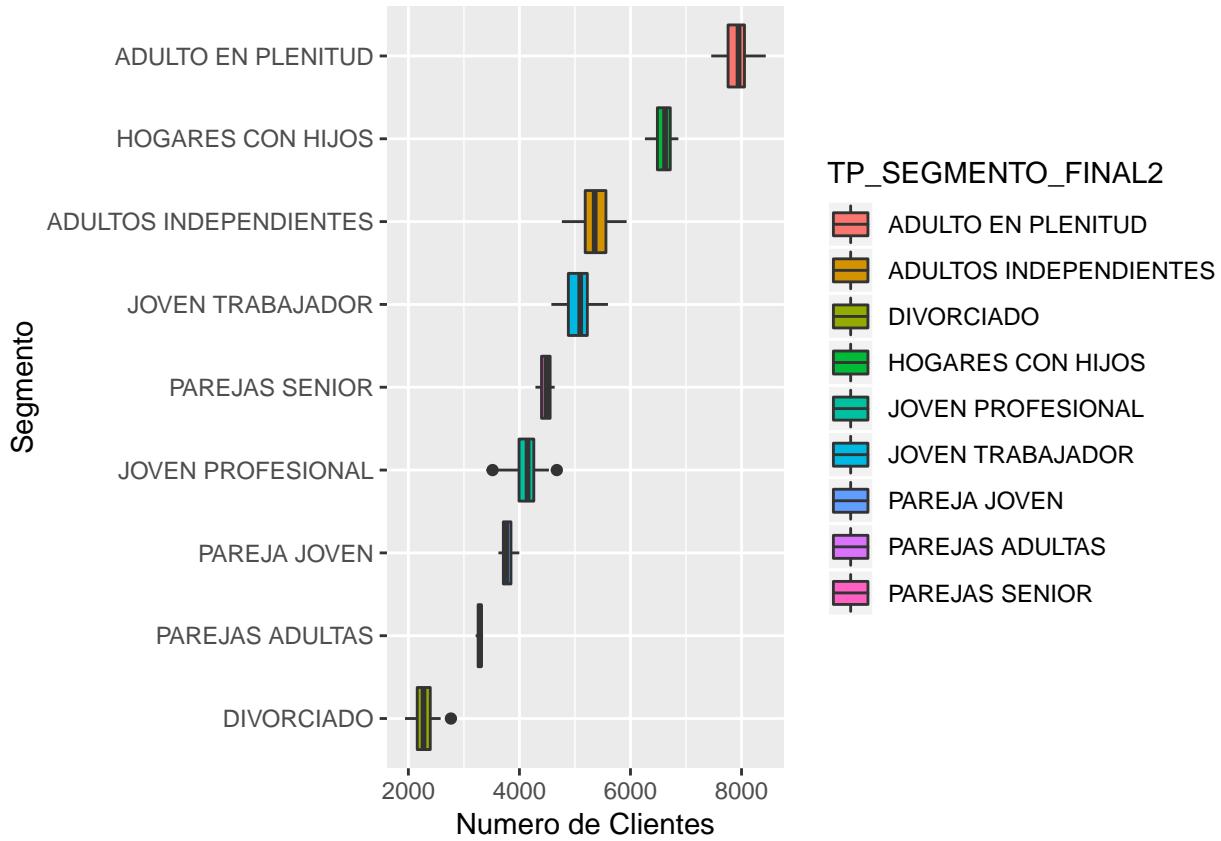
Los segmentos con mayor número de contrataciones de forma mensual son HOGARES CON HIJOS y ADULTO EN PLENITUD. Los que tienen menos colocación son DIVORCIADO y JOVEN TRABAJADOR.

Se presenta una gráfica de caja y brazos donde se notan diferencias entre los distintos segmentos. Como se había mencionado, el segmento de HOGARES CON HIJOS es un grupo totalmente separado al comparar su rango intercuartílico con respecto a los demás, y el que tiene mayor volumen de contrataciones. Por otro lado, se puede observar que desde el segmento de ADULTO EN PLENITUD hasta JOVEN PROFESIONAL, al menos el 75% de los meses han sobrepasado las 40 contrataciones. Existe un tercer grupo donde el volumen y la distribución de contrataciones mensuales es el más bajo: PAREJA JOVEN, DIVORCIADO y JOVEN TRABAJADOR.



Es importante resaltar que estas distribuciones no necesariamente tienen una relación directamente proporcional al número de clientes que tenemos en cada segmento por mes, pero sí nos muestra que hay segmentos a los que de forma más recurrente se les otorgan los préstamos. Veamos la distribución de los clientes por segmento.

```
ggplot(camp, aes(x=fct_reorder(TP_SEGMENTO_FINAL2, ID), y=ID, fill=TP_SEGMENTO_FINAL2)) +
  geom_boxplot() + coord_flip() + xlab('Segmento') + ylab('Número de Clientes')
```



En la tabla siguiente observamos los estadísticos puntuales de cada segmento. Se puede confirmar lo explicado en las figuras anteriores de manera puntual: los HOGARES CON HIJOS es el segmento con mayor número de contrataciones promedio mensual con 104, también en la mediana con 102; tiene la mayor tasa promedio de contrataciones por mes con 1.58%. Despues, los segmentos ADULTO EN PLENITUD y PAREJA SENIOR tienen estadísticos similares entre sí, aunque la tasa de contratación promedio de los primeros es la segunda más baja. Al final, los DIVORCIADO y JOVEN TRABAJADOR son los segmentos con menor número de contrataciones de manera mensual, estos últimos también tienen la tasa de contratación promedio mensual más baja .5%.

```
## # A tibble: 9 x 4
##   TP_SEGMENTO_FINAL2   mediana promedio tasa_promedio
##   <fct>          <dbl>    <dbl>      <dbl>
## 1 HOGARES CON HIJOS 102.     104.       1.58 
## 2 ADULTO EN PLENITUD 68       67.5      0.852 
## 3 PAREJAS SENIOR    63.5     61.5      1.37 
## 4 ADULTOS INDEPENDIENTES 55     54.1      1.01 
## 5 JOVEN PROFESIONAL 46       49.8      1.21 
## 6 PAREJAS ADULTAS   48       48.6      1.48 
## 7 PAREJA JOVEN     39       39.4      1.04 
## 8 DIVORCIADO        27.5     29.2      1.27 
## 9 JOVEN TRABAJADOR  25       26.7      0.523
```

Variables Numéricas

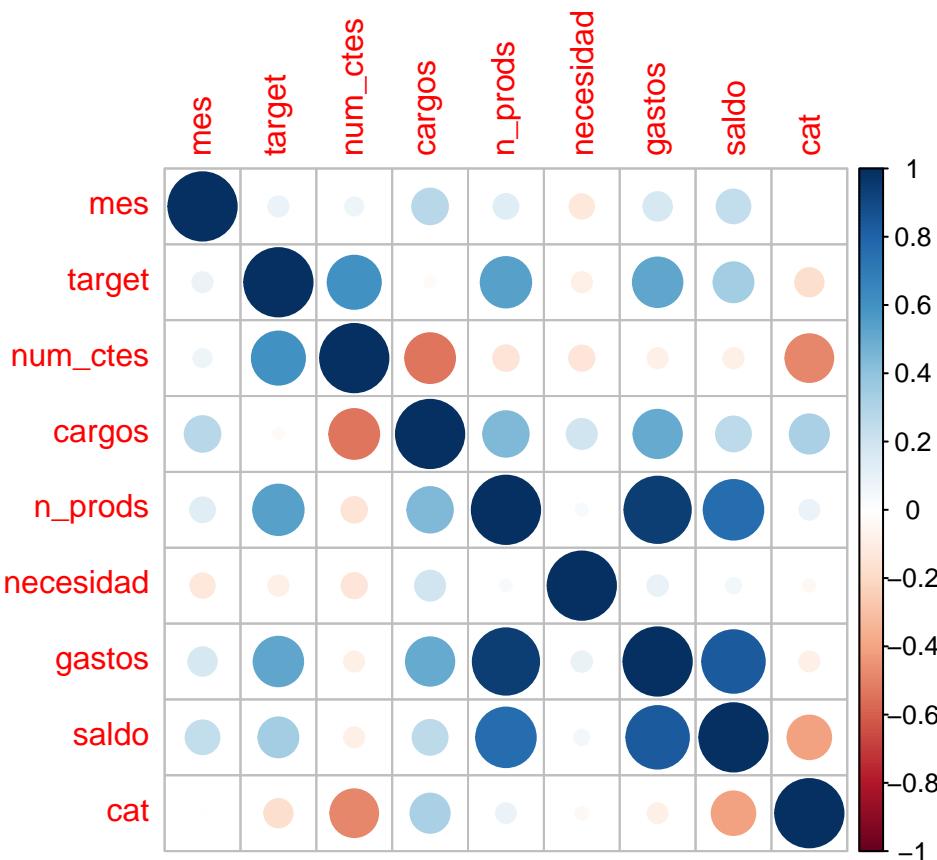
Para realizar una primera exploración a la interacción de las variables, graficamos un correlograma de nuestros datos. Para hacer esto se transformó la variable de fecha a numérica:

```

library(corrplot)

## corrplot 0.84 loaded
campx<-camp
campx$FH_REF<-as.numeric(campx$FH_REF)
names(campx) <- c("mes","seg","target","num_ctes","cargos","n_prods","necesidad","gastos","saldo","cat")
nums <- unlist(lapply(campx, is.numeric))
corrplot(cor(campx[,nums]))

```



Podemos observar que las variables más relacionadas linealmente con nuestro target son: el número de clientes, el número de productos y los gastos. Otras correlaciones que podemos ver son entre los cargos y el número de productos de forma positiva, así como con los gastos; también el número de productos tiene una relación positiva con los saldos de cuenta de cheques. Por otro lado, vemos una relación negativa entre la variable cat, que es el segmento construido con un valor numérico de forma descendente con respecto a la edad, y los saldos, es decir tienes más liquidez mientras más grande eres.

El tiempo tiene mayor relación positiva con los cargos y los saldos lo cual suena coherente en el sentido de que los gastos son distintos en el año y por tanto la liquidez también.

Ahora podemos ver las distintas variables en el tiempo...

```

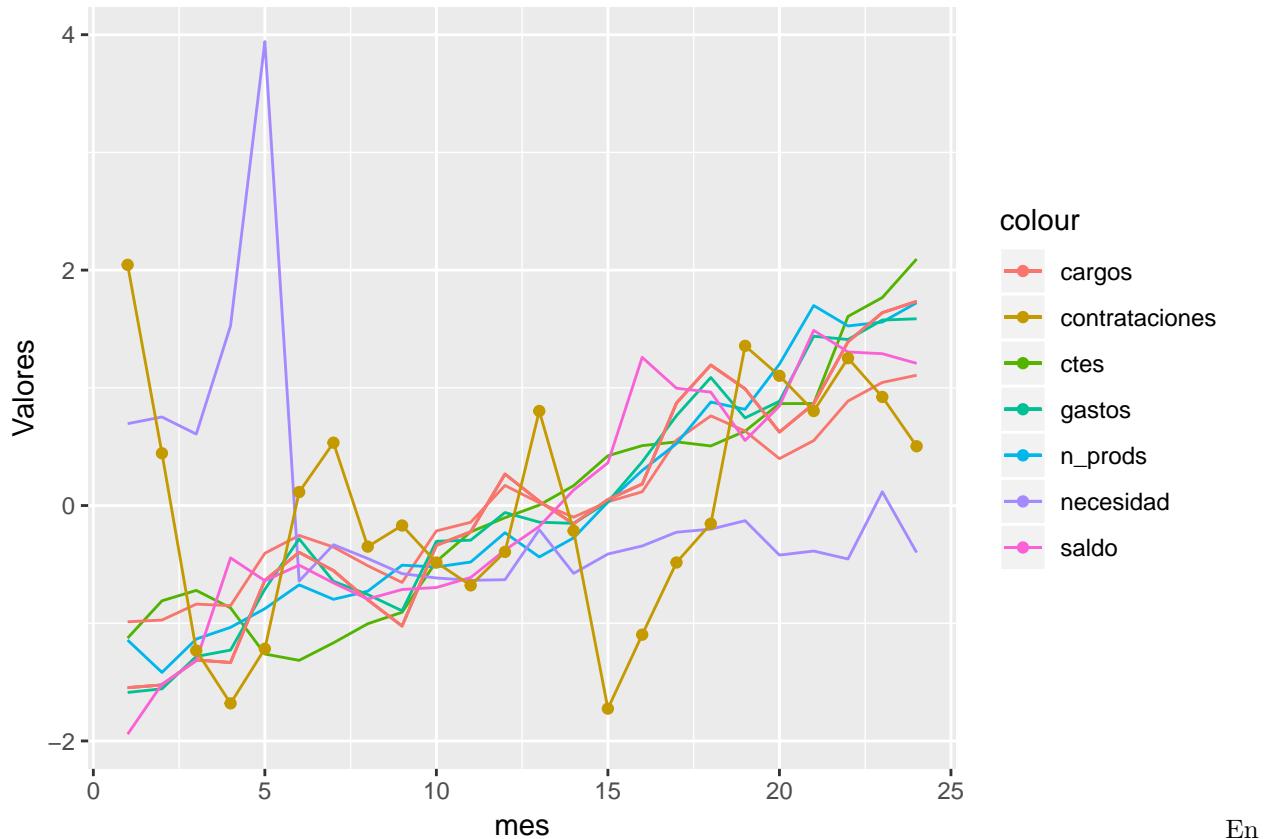
totp<-campx %>% group_by(mes)%>%summarise(total=sum(target),ctes=sum(num_ctes),cargos=mean(cargos),n_prods=mean(n_prods))
#ggplot(tot, aes(x=as.Date(FH_REF))) + geom_line(y=total, col="blue")
ggplot(totp, aes(x=mes)) + geom_line(aes(y = scale(ctes), colour = "ctes")) +
  geom_line(aes(y = scale(cargos), scale = FALSE), colour = "cargos")+
  geom_line(aes(y = scale(n_prods), colour = "n_prods"))+
  geom_line(aes(y = scale(cargos), colour = "cargos"))+
  geom_line(aes(y = scale(necesidad), colour = "necesidad"))+

```

```

geom_line(aes(y = scale(cargos), colour = "cargos"))+
geom_line(aes(y = scale(gastos), colour = "gastos"))+
geom_line(aes(y = scale(cargos), colour = "cargos"))+
geom_line(aes(y = scale(saldo), colour = "saldo"))+
geom_point(aes(y = scale(total), colour = "contrataciones"))+
geom_line(aes(y = scale(total), colour = "contrataciones"))+
ylab('Valores')

```



En la gráfica anterior se presentan las variables reescaladas con el fin de poder comparar mejor las variables entre sí y vs las contrataciones a lo largo del tiempo.

Podemos observar que para estos clientes la tendencia de las variables es creciente, excepto la de necesidad financiera lo cual tiene sentido por la forma en que se construyó. Sabemos que las variables con mayor correlación con las contrataciones son: el número de clientes (ctes), el importe promedio de gastos de los últimos 3 meses (gastos) y el número de productos bancarios promedio que poseen los clientes (n_prods), sin embargo, en ninguno de los casos la relación es tan evidente en la gráfica anterior.

Modelado e implementación

Modelos Estáticos

Primero comenzamos utilizando dos modelos estáticos, una regresión normal y un modelo lineal generalizado poisson con liga log para ver como se ajustaban a los datos pero primordialmente para ver el comportamiento y poder interpretar cada una de las variables. Asimismo, nos interesa saber si el tipo de categoría guarda relación con el número de créditos otorgados.

Modelo lineal normal

Se utilizó el siguiente modelo:

$$Y_i \sim N(\mu_i, \tau)$$

$$\mu_i = \alpha + \gamma_j I_{categoria} + \beta_1 X2_i + \beta_2 X3_i + \beta_3 X4_i + \beta_4 X5_i + \beta_5 X6_i$$

Donde:

$$\sum_{j=1}^9 \gamma_j = 0$$

$$\alpha \sim N(0, 0.001)$$

$$\gamma_j \sim N(0, 0.001)$$

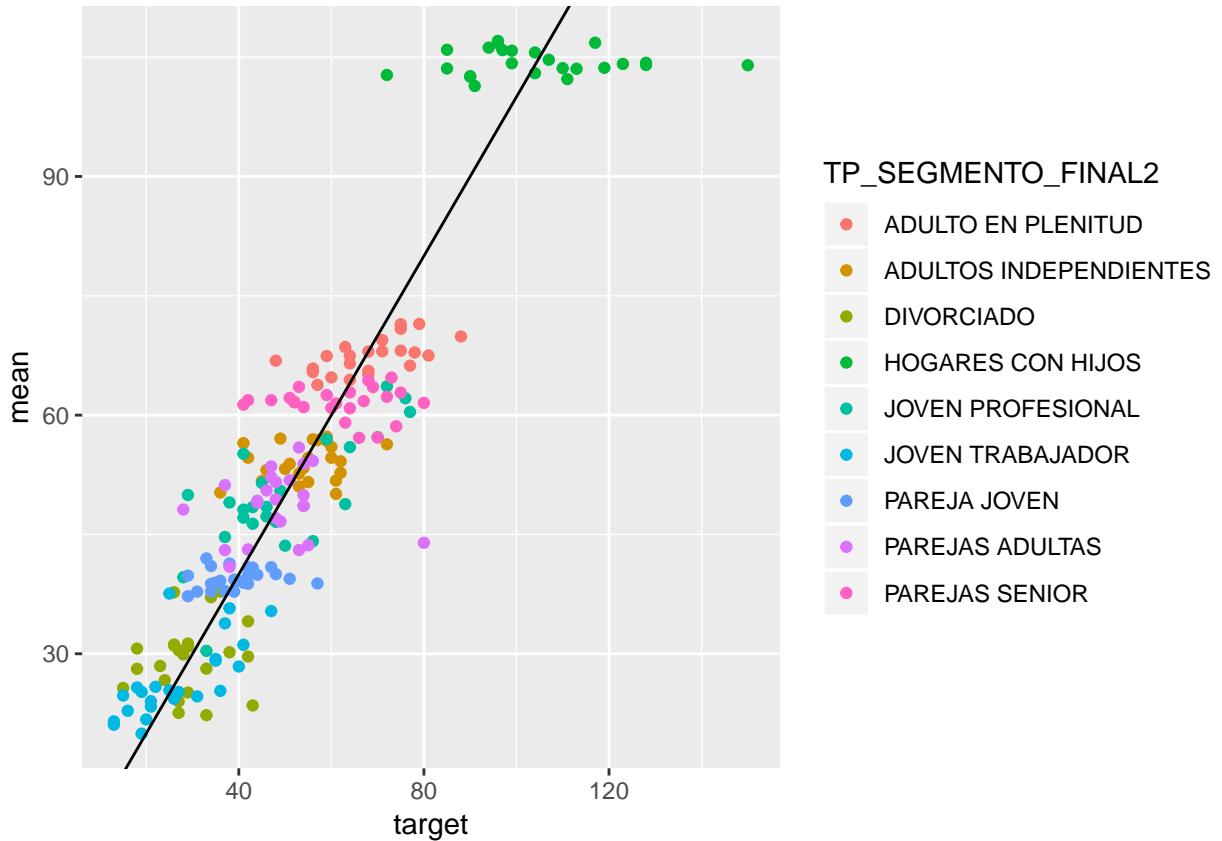
$$\tau \sim Gama(0.001, 0.001)$$

$$\beta_j \sim N(0, 0.001)$$

Se utilizó una regresión lineal con errores que se distribuyen normal con media cero y varianza constante. Como la categoría es una variable categórica se incluyó la restricción de que la suma de los coeficientes (γ_j) fuera cero. Se definió que el intercepto α , γ_j y las β_j provienen de una distribución normal con media 0 y precisión 0.001 para que sean no informativas. Asimismo, se definió que la τ se distribuye gamma con $a = 0.001$ y $b = 0.001$.

Se corrieron dos cadenas y se revisaron tanto las cadenas como la convergencia para cada uno de los coeficientes. En todos los casos se apreció convergencia como se puede constatar en las gráficas que se incluyen en el Apéndice.

Graficamos los valores observados contra las estimaciones y notamos que en general las estimaciones no eran tan malas con excepción del caso de hogares con hijos en el que el modelo no esta captando adecuadamente la información de este segmento particularmente.



Proseguimos a utilizar un modelo lineal generalizado como se describe en el siguiente apartado para ver si se mejoraba el ajuste.

Modelo lineal generalizado poisson con liga log

Este modelo se definió de la siguiente manera:

$$Y_i \sim Po(\mu_i)$$

$$\mu_i = n e_i \cdot \lambda_i$$

$$\log(\lambda_i) = \alpha + \gamma_j I_{categoria} + \beta_1 X2_i + \beta_2 X2_i + \beta_3 X4_i + \beta_4 X5_i + \beta_5 X6_i$$

Donde:

$$\sum_{j=1}^9 \gamma_j = 0$$

$$\alpha \sim N(0, 0.001)$$

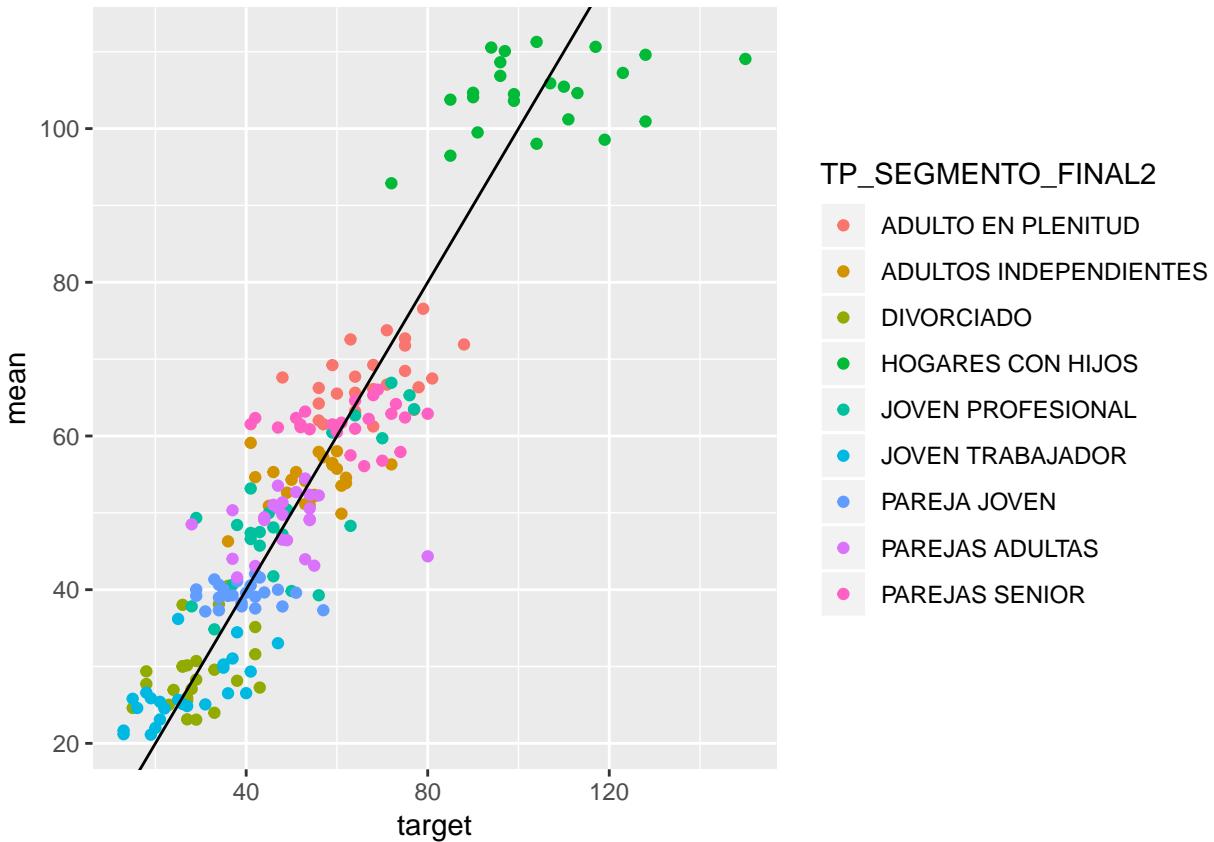
$$\gamma_j \sim N(0, 0.001)$$

$$\beta_j \sim N(0, 0.001)$$

Se utilizó un modelo lineal generalizado Poisson con liga log. De nueva cuenta como los segmentos son una variable categórica se incluyó la restricción de que la suma de los coeficientes (γ_j) fuera cero. Asimismo, se definió que el intercepto α , γ_j y las β_j provienen de una distribución normal con media 0 y precisión 0.001 para que sean no informativas.

Para este modelo también se utilizaron 2 cadenas y se constató que la convergencia para cada uno de los coeficientes fue correcta como se puede visualizar en las gráficas que se incluyen en el Apéndice.

Por otro lado, apreciamos que este modelo mejoró la estimación para la categoría hogares con hijos como se puede notar a continuación:



No obstante lo anterior, decidimos definir un modelo que incorpore el efecto del tiempo buscando “captar” de manera más apropiada la información de los datos.

Modelo dinámico

Decidimos aprovechar la temporalidad de nuestros datos incorporando la variable tiempo a través de un modelo dinámico lineal generalizado. Sabemos de antemano que este modelo debe ajustar mucho mejor, pues recordamos que suelen ajustar perfecto a la serie de tiempo. Con este modelo tenemos como objetivo entender más a detalle la temporalidad de las variables, si su significancia tiene alguna especie de temporalidad o no, y de igual manera poder comparar como se comportan las variables representativas diferenciando por segmento para saber si hace sentido nuestra hipótesis de acotar campañas específicas para los diferentes ciclos de vida de los clientes. Cabe mencionar también que utilizaremos únicamente este modelo para predecir por su naturaleza temporal.

El modelo dinámico lo definimos como sigue:

Observación:

$$Y_t | \mu_t \sim Po(\mu_t)$$

$$\mu_t = n e_t \cdot \lambda_t$$

$$\log(\lambda_t) = \alpha_t + \gamma_{jt} I_{categoria} + \beta_{1t} X_{2t} + \beta_{2t} X_{3t} + \beta_{3t} X_{4t} + \beta_{4t} X_{5t} + \beta_{5t} X_{6t}$$

Evolución:

$$\beta_{ijt} \sim N(\beta_{ijt-1}, \tau_t) \quad i = 1, \dots, 5$$

Donde:

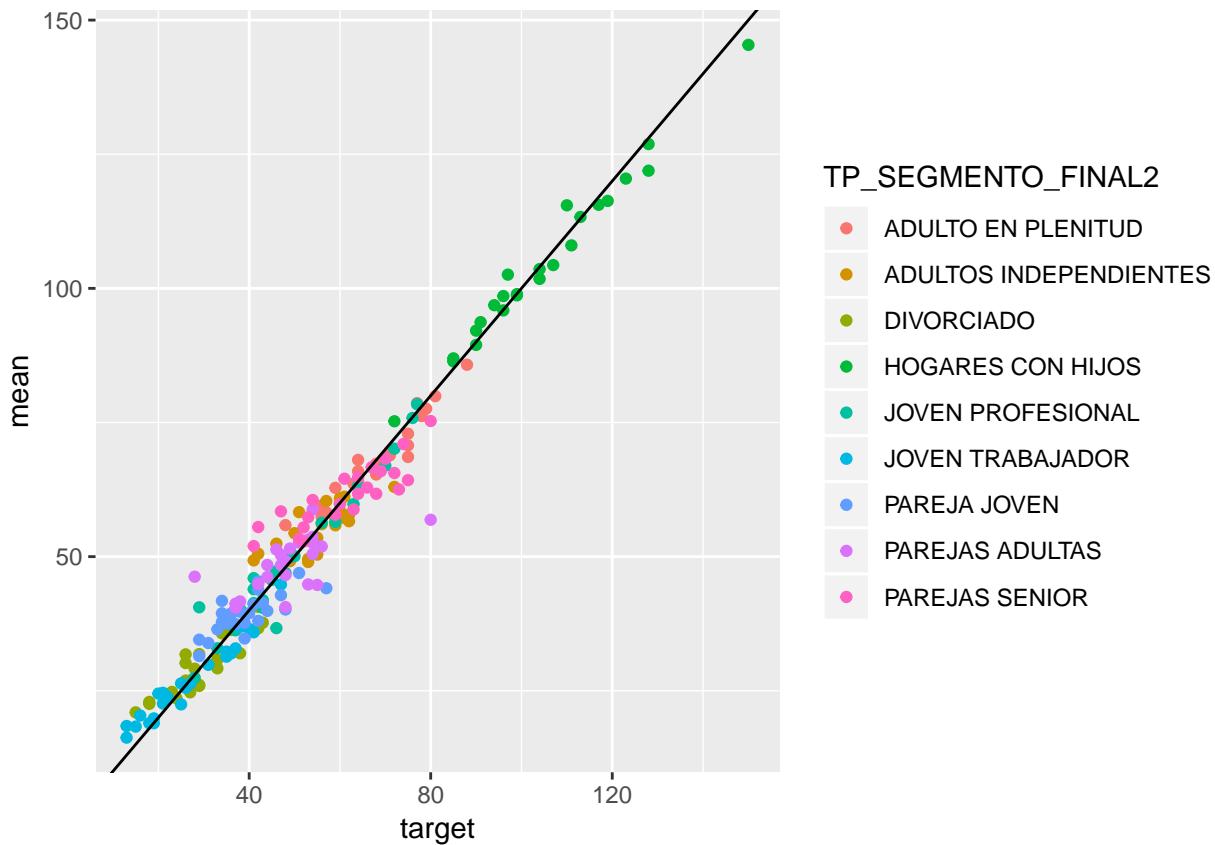
$$\sum_{j=1}^9 \gamma_j = 0$$

$$\beta_{1j1}, \dots, \beta_{5j1} \sim N(0, 0.001)$$

$$\alpha \sim N(0, 0.001)$$

$$\gamma_j \sim N(0, 0.001)$$

$$\tau \sim Gama(100, 1)$$



El DIC de los tres modelos se muestra a continuación:

```
## DIC
## Regresion Lineal Normal 1634.55
## Glm poisson liga log 1625.64
## Dinamico poisson liga log 1638.54
```

Notamos que el mejor DIC lo obtuvimos con el GLM Poisson con liga log, seguido de la regresión lineal normal y por último el modelo dinámico. Esto es contraintuitivo pues por el buen ajuste que tiene el último modelo, pensaríamos que debería tener el DIC más bajo. Aunque como no varía tanto respecto a los otros pensamos que puede ser un problema de JAGS.

```
## Pseudo-R2
## Regresion Lineal Normal 0.8375450
## Glm poisson liga log 0.8476439
## Dinamico poisson liga log 0.9665040
```

La mejor pseudo R^2 se obtuvo con el modelo dinámico por mucha diferencia, lo cual era de esperarse pues, como sabemos, los modelos dinámicos ajustan casi perfectamente a la serie de tiempo.

Interpretación de modelos

Modelos estáticos

Modelo lineal normal

Para el modelo lineal normal se aprecia que el sector 1 (ADULTO EN PLENITUD) y el 4 (HOGARES CON HIJOS) son los que tienen mayor influencia positiva en el número de créditos, mientras que el 3 (DIVORCIADO) y el 7 (PAREJA JOVEN) son los que menor aporte tienen. Esto cuadra perfectamente con lo que esperábamos dado el análisis exploratorio.

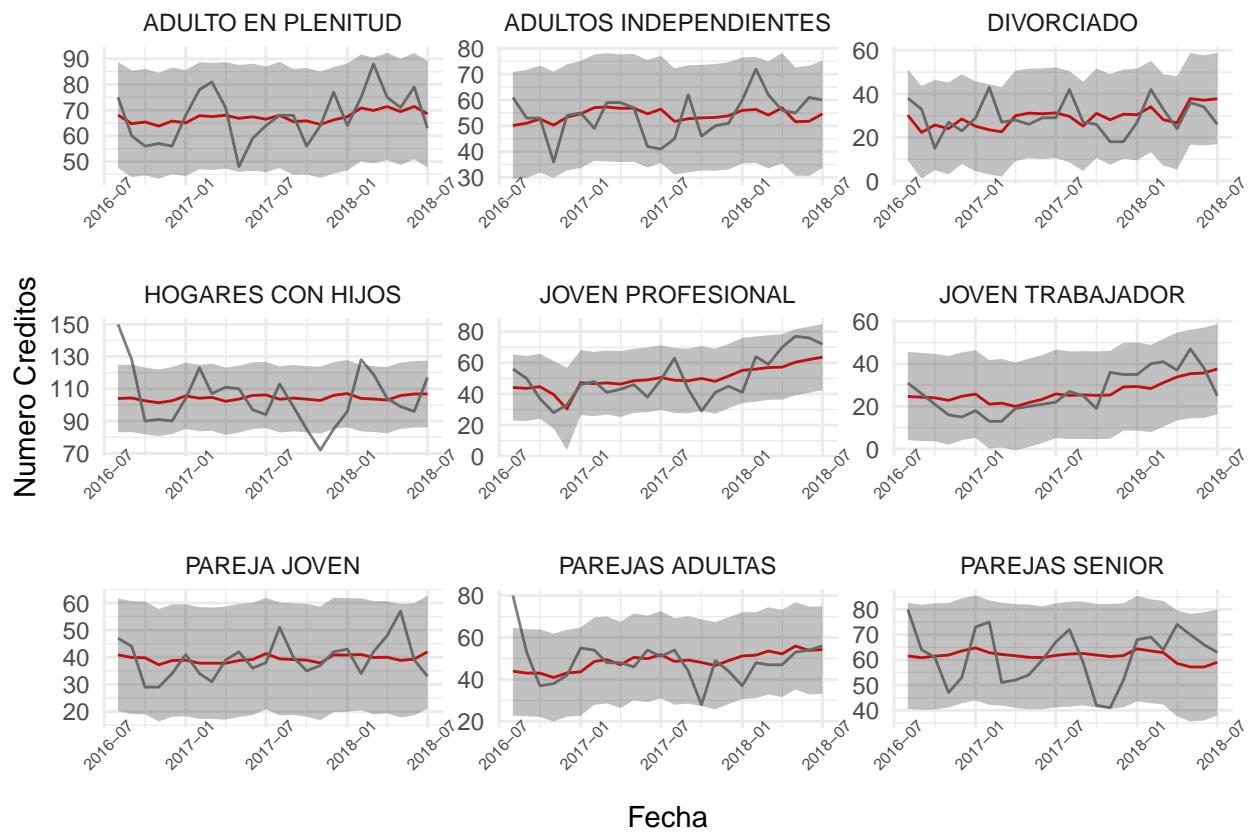
En cuanto a las demás variables, la referente a otros productos bancarios (β_2) es la que tiene mayor efecto sobre el número de créditos que se otorgan, mientras que ni el cociente de gasto vs saldo (β_3) ni el importe promedio de los últimos 3 meses (β_4) son significativos, pues su intervalo de confianza incluye al cero.

```
##               mean      2.5%     97.5%
## alpha.adj    53.461526  52.092326  54.8430916
## beta[1]      8.957241   1.932520  15.9206825
## beta[2]     18.176980   3.062320 33.5507424
## beta[3]     -1.344589  -2.942725  0.2510475
## beta[4]     -6.475591 -15.102006  2.0955961
## beta[5]     -7.364654 -13.406689 -1.4042234
## gama.adj[1]  40.928899  28.973600  53.0750253
## gama.adj[2]  15.905802   8.177202 23.8014755
## gama.adj[3] -24.262371 -34.213550 -14.5204402
## gama.adj[4]  29.671149   7.048531  52.0867014
## gama.adj[5] -10.011851 -24.463812  4.4359098
## gama.adj[6] -18.885073 -40.336583  2.4792849
## gama.adj[7] -21.159869 -26.720041 -15.5827211
## gama.adj[8] -18.693273 -25.764264 -11.7091271
## gama.adj[9]  6.506586  -1.934038 14.9268886
```

En la siguiente gráfica se muestra la serie de datos observados (en color negro) y los valores estimados (en color rojo). La franja gris, representa el intervalo al 95% de confianza.

Como era de esperarse, debido a la interpretación de los coeficientes, los grupos correspondientes a ADULTOS EN PLENITUD y HOGARES CON HIJOS, son los que presentan valores más altos en el número de créditos colocados. Mientras que DIVORCIADO presenta los valores más bajos.

Este modelo nos da una idea general de la relación de cada sector con respecto al número de créditos que se otorgan, pero no se recomienda su uso para hacer pronósticos. Además, podemos observar que no ajusta adecuadamente por sector.

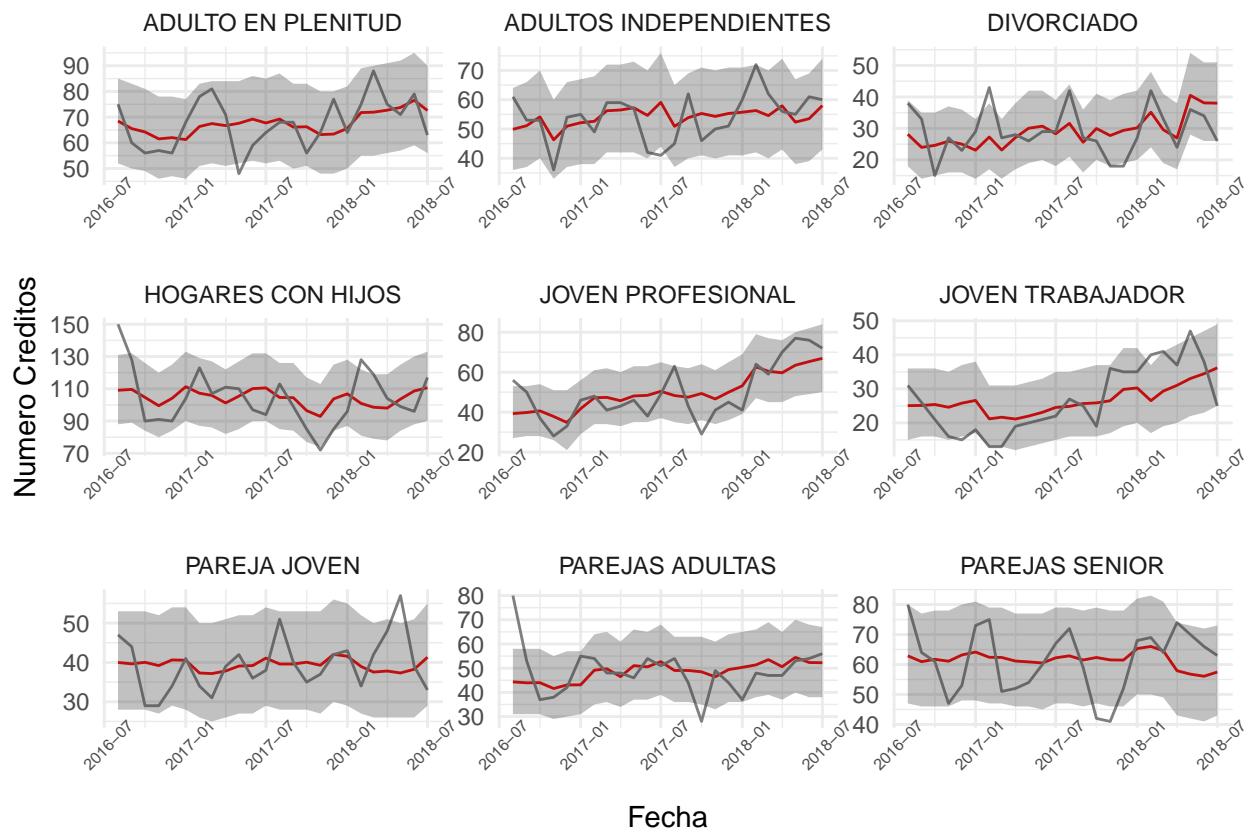


Modelo lineal generalizado poisson con liga log

Al igual que en el modelo anterior la variable con más influencia es β_2 que corresponde al número promedio de productos bancarios, mientras que β_3 cociente de gasto vs saldo y β_4 importe promedio de los últimos 3 meses, siguen sin ser significativas. De manera similar, el sector con más influencia positiva es el 1 (ADULTO EN PLENITUD). Por su parte, el sector 6 (JOVEN TRABAJADOR) tiene la mayor influencia negativa.

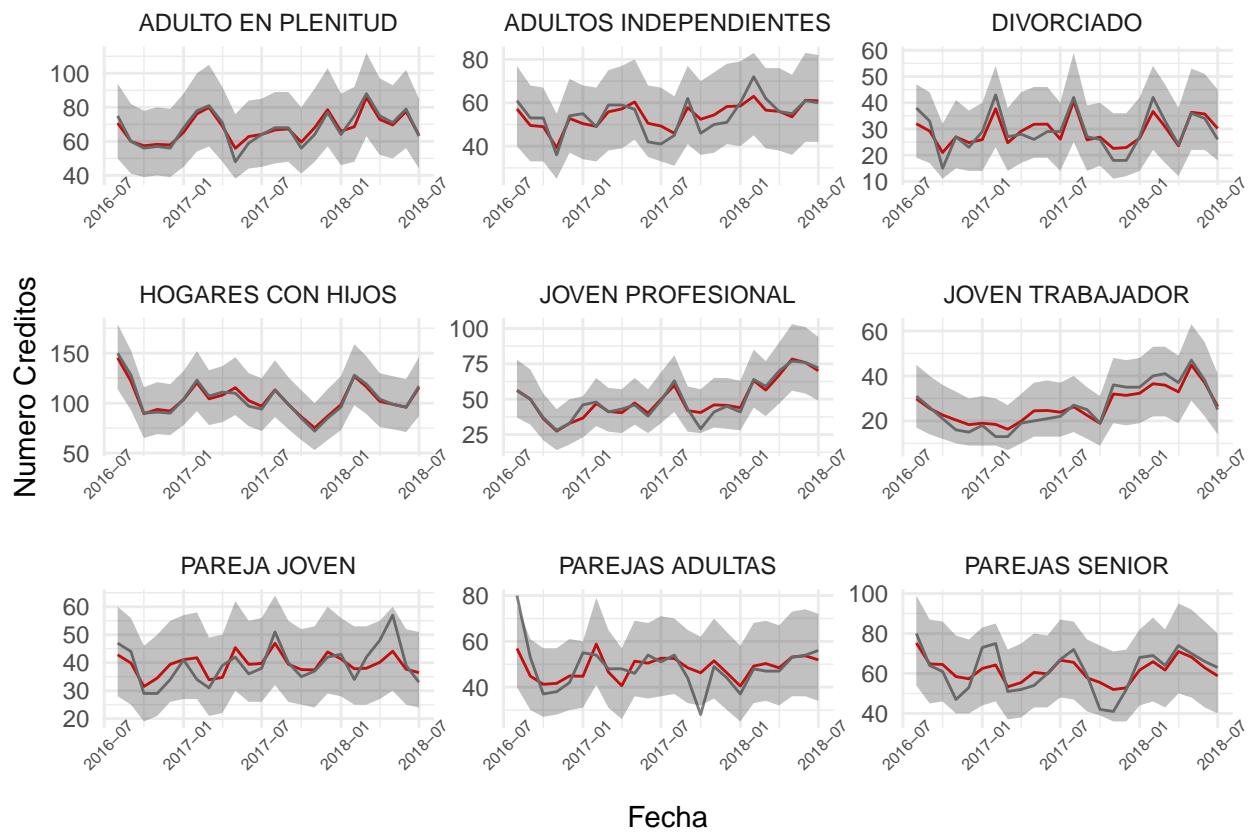
```
##               mean      2.5%     97.5%
## alpha.adj   -4.51485088 -4.53482878 -4.49491303
## beta[1]      0.19966883  0.09448857  0.30553854
## beta[2]      0.27393434  0.03310018  0.52977918
## beta[3]     -0.01331146 -0.03925244  0.01132821
## beta[4]     -0.09404217 -0.21760990  0.02405724
## beta[5]     -0.20336334 -0.29062311 -0.11453282
## gama.adj[1]  0.28030152  0.10978850  0.45321276
## gama.adj[2]  0.21178649  0.09177483  0.33218832
## gama.adj[3]  0.26378635  0.10774962  0.41985172
## gama.adj[4]  0.10249441 -0.27322441  0.47546454
## gama.adj[5] -0.11405538 -0.33976054  0.11169083
## gama.adj[6] -0.77898042 -1.12647873 -0.42521296
## gama.adj[7] -0.21744491 -0.30000131 -0.13799201
## gama.adj[8]  0.02906226 -0.07509319  0.12671733
## gama.adj[9]  0.22304968  0.09168130  0.35302375
```

El modelo lineal generalizado Poisson con liga log, en general da resultados similares a la regresión normal; sin embargo, en algunos casos como HOGARES CON HIJOS y JOVEN PROFESIONAL genera mejores estimaciones, aunque como observamos, la estimación sigue sin ser adecuada.



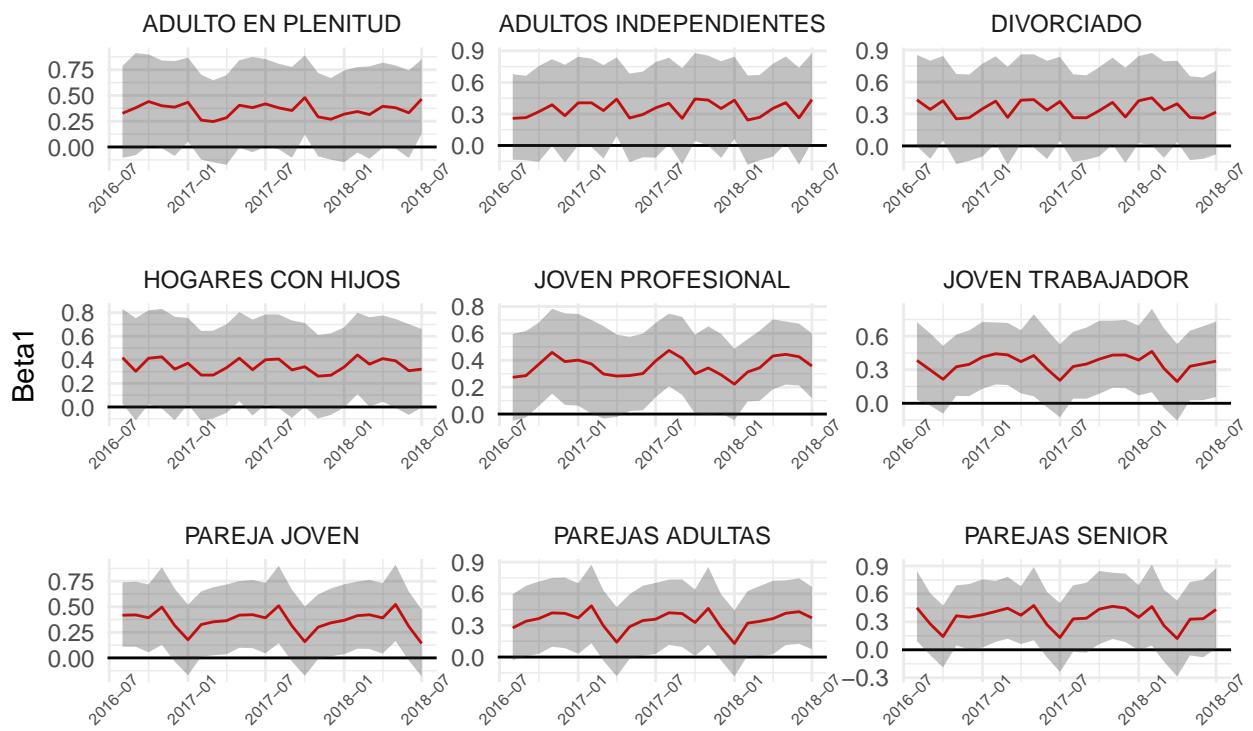
Modelo dinámico

Por su construcción los datos se ajustan casi perfecto a la serie de tiempo, seguimos identificando que hay sectores donde el ajuste es mejor que otros pero aún así, el modelo es muy superior tanto en la inspección visual como con la pseudo R^2 . Por el momento nos limitaremos a hacer un análisis sobre la estimación de los parámetros para más adelante ahondar más en el tema del pronóstico.

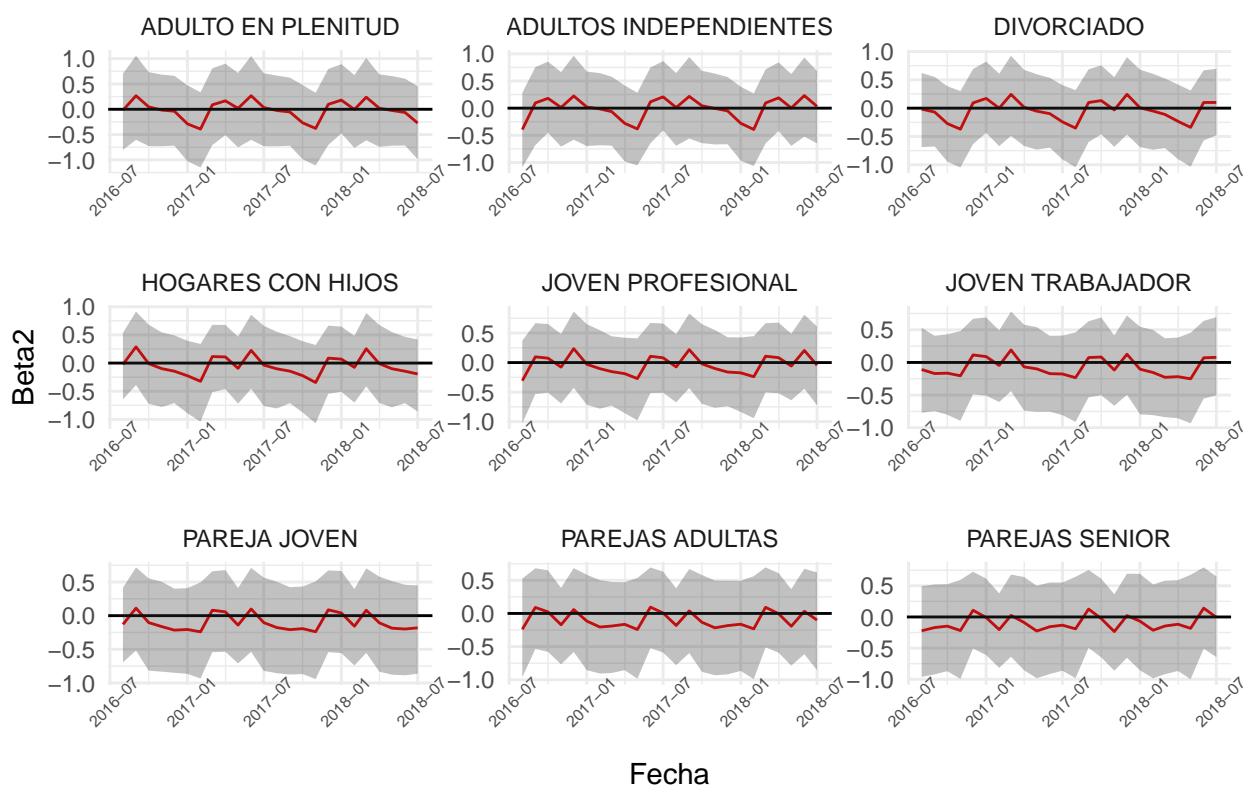


Para el modelo dinámico buscaremos si una variable es representativa o no, lo que es interesante es que podemos encontrar estacionalidad en su importancia para predecir el número de créditos colocados. A continuación graficamos el valor de las betas por segmento para analizar su representatividad a través del tiempo.

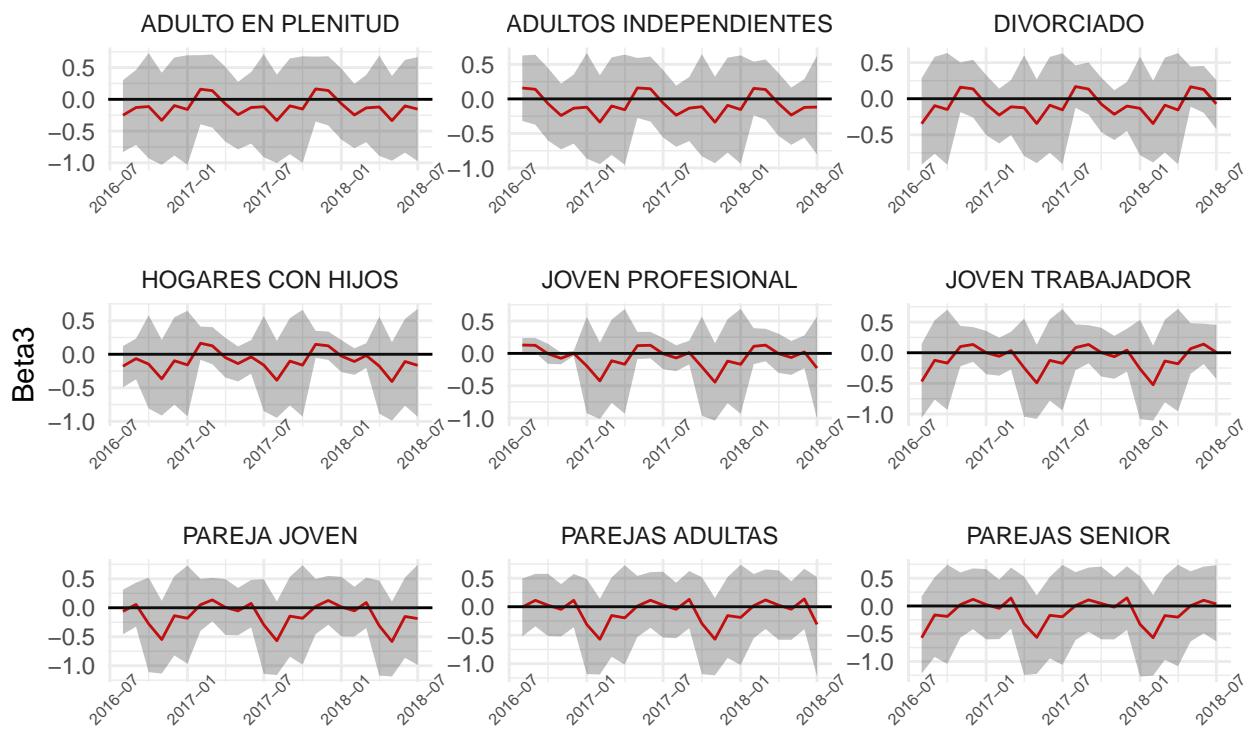
Beta1 por segmento a traves del tiempo



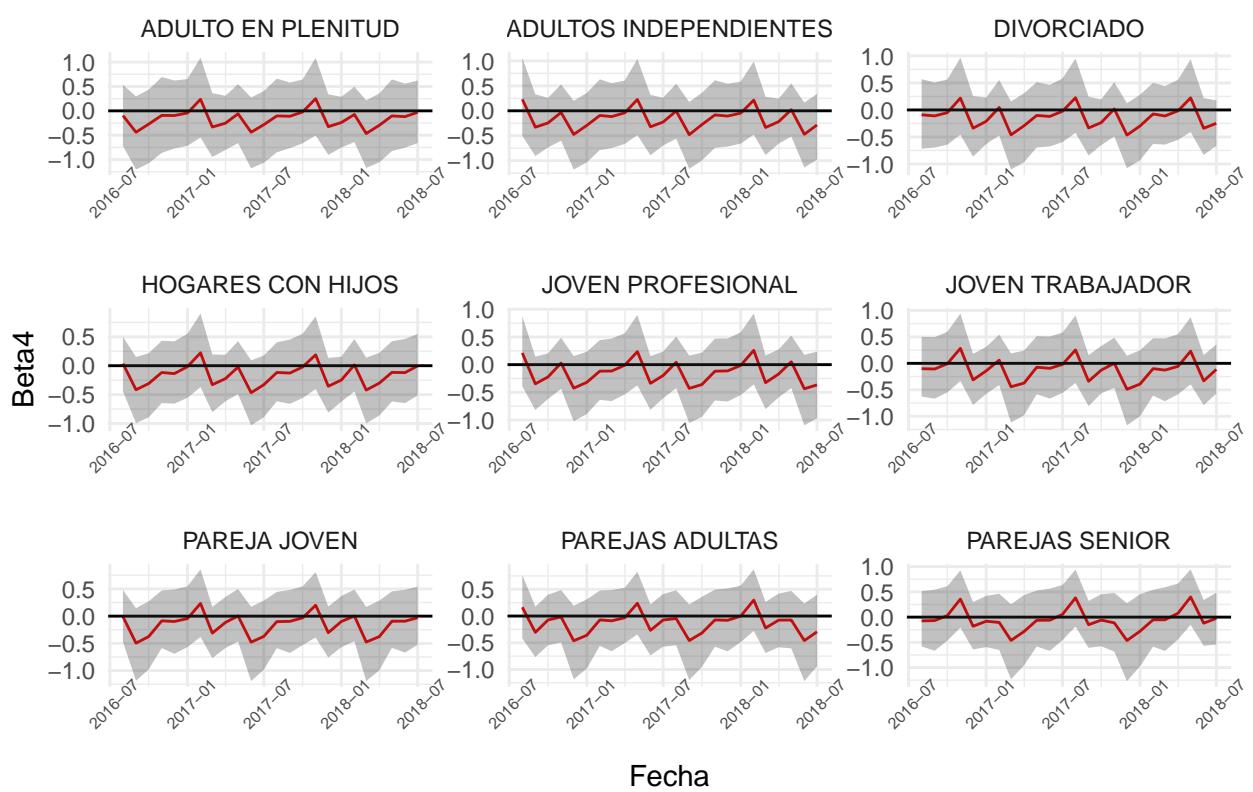
Fecha Beta2 por segmento a traves del tiempo



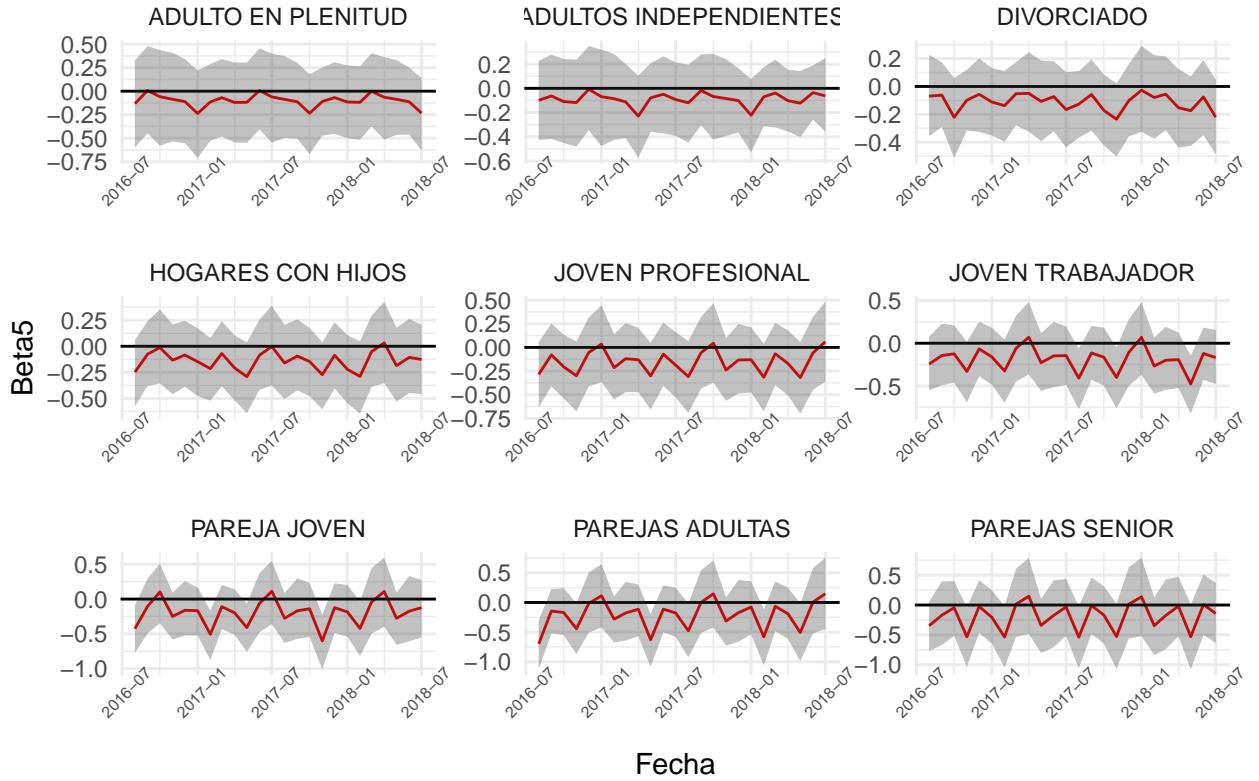
Beta3 por segmento a traves del tiempo



Beta4 por segmento a traves del tiempo



Beta5 por segmento a través del tiempo



Observamos que, todas las betas contienen al cero en su intervalo de confianza durante la mayor parte del periodo observado a excepción de β_{1t} que es la correspondiente al número promedio de cargos en tres meses. Ahora, queremos identificar el cambio porcentual que resulta de aumentar en una unidad esta variable en cada una de las segmentaciones.

Sabiendo que el modelo es Poisson y la liga es logarítmica, podemos llegar al siguiente resultado.

Recordemos el modelo que utilizamos para este apartado

$$\log(\lambda_t) = \alpha_t + \gamma_{jt} I_{\text{categoria}} + \beta_{1t} X_{2t} + \beta_{2t} X_{3t} + \beta_{3t} X_{4t} + \beta_{4t} X_{5t} + \beta_{5t} X_{6t}$$

Sea $x_j = x_i + k$, entonces

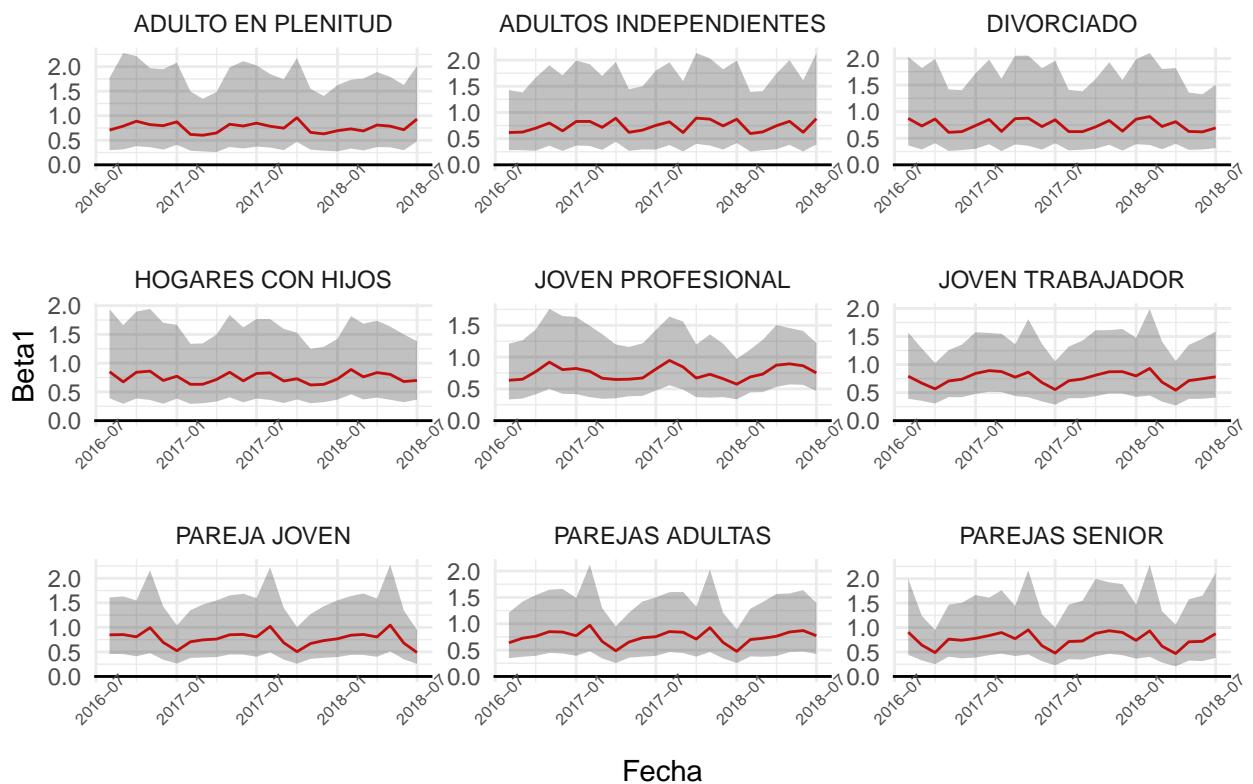
$$\log(\lambda_{ti}) - \log(\lambda_{tj}) = \beta_{1t} + k\beta_{1t}$$

Por lo tanto,

$$e^{k\beta_{1t}} = \lambda_{ti}/\lambda_{tj}$$

Por lo que aplicaremos esta transformación a la β_1 para analizar con mayor detalle en cuántas unidades aumentaría en cada segmento el número de contrataciones por cada k unidades en la variable correspondiente a dicha β . Para que las cantidades se ilustren adecuadamente tomaremos $k = 2$

Cambio porcentual de beta1 al aumentar en dos unidades



Podemos observar claramente la estacionalidad en los segmentos PAREJA JOVEN, PAREJAS ADULTAS, PAREJAS SENIOR y JOVEN TRABAJADOR. En HOGARES CON HIJOS también parece haber una estacionalidad aunque un poco menos marcada. Los demás segmentos muestran estacionalidad muy débil, sin embargo, la conclusión es que podemos hacer estrategias de mercado de manera personalizada para cada segmento del Ciclo de Vida del cliente, lo que resulta súper relevante de este estudio es que podemos no solo saber a quiénes, sino cuándo hacer la comunicación para que la campaña resulte más efectiva y mejor aún, estudiando una sola variable: Número promedio de cargos en 3 meses.

Prediccion

Para no desaprovechar ninguna fecha que tenemos en el dataset, decidimos realizar el estudio descriptivo de las variables por un lado y el análisis predictivo por otro. Recordemos que, como comentamos previamente, únicamente usaremos el modelo dinámico para hacer predicciones por su naturaleza temporal. Sabemos de antemano que las proyecciones se estabilizan después de cierto número de períodos por la construcción del modelo, por lo que decidimos pronosticar solamente 3 meses hacia adelante en la serie de cada segmento. Quitaremos las últimas tres observaciones de cada uno para poder probar qué tan bien está ajustando nuestro modelo.

Comparado con el modelo dinámico con todos los datos, su estimación no cambia mucho. Vemos que en general ajusta bastante bien como era de esperarse por ser un modelo dinámico. Incluso las observaciones nuevas están bastante bien pronosticadas.

Comparando los DIC del modelo dinámico anterior, este resulta ser mucho mejor. No es intuitivo pero ya sabemos como es JAGS.

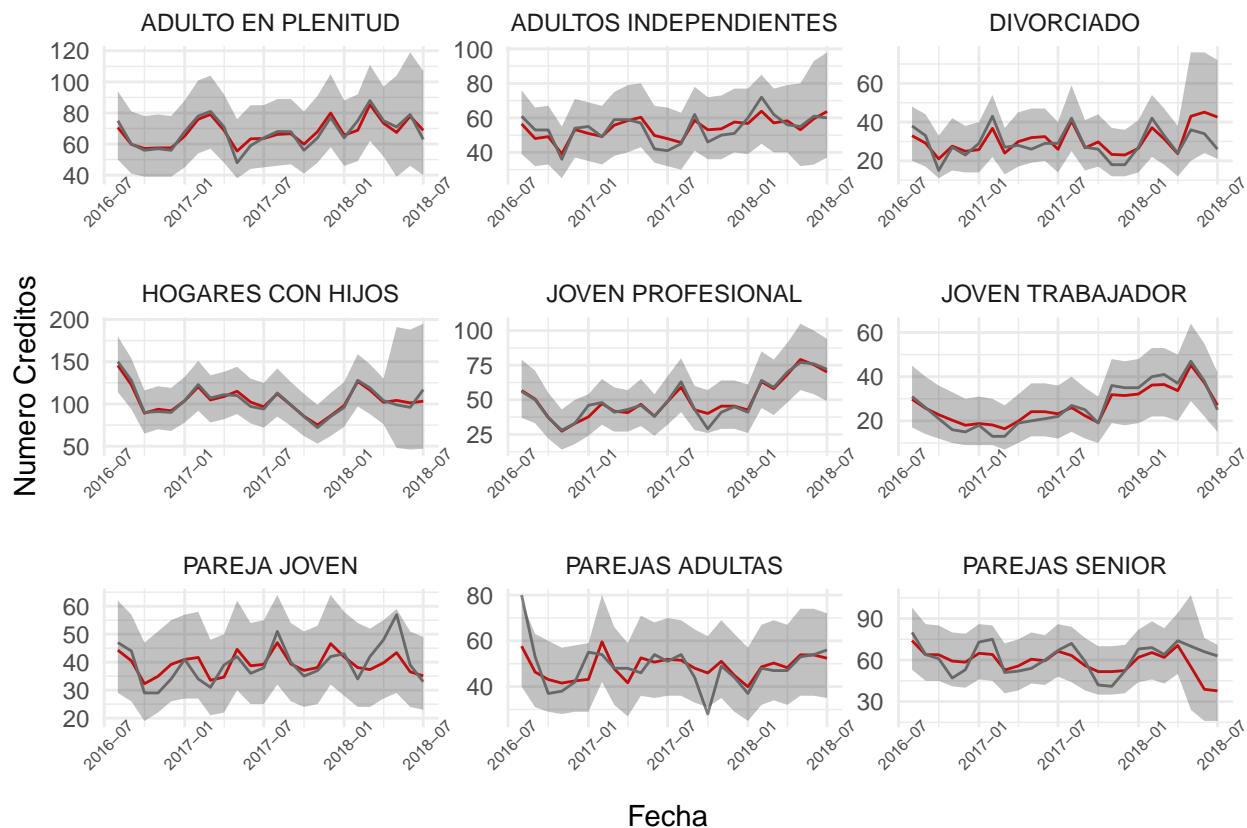
```
## DIC
## Dinamico poisson liga log      1638.54
```

```
## Dinamico poisson liga log pronostico 1524.77
```

Recordemos que el modelo pronosticó las últimas tres observaciones, y siendo que por su naturaleza los modelos dinámicos ajustan muy bien a los datos observados, era lógico pensar que la pseudo R² sería menor en el modelo predictivo.

##	Pseudo-R2
## Dinamico poisson liga log	0.966504
## Dinamico poisson liga log pronostico	0.950216

Como era de esperarse encontramos que las observaciones que se predicen, mantienen la tendencia con la que venía la serie en observaciones pasadas. Para aquellos segmentos que la tendencia se mantiene el ajuste es bastante bueno, sin embargo el modelo no tiene la sensibilidad de corregir o predecir cambios en la tendencia, esto se puede ilustrar con la serie de DIVORCIADO. Vemos que la línea roja (predicciones) mantiene un cierto nivel dado lo que se observa previamente, sin embargo la serie real cambia su tendencia y esto provoca que nuestra estimación puntual quede lejos de lo observado, a pesar de eso, vemos que el valor real esta dentro de nuestros intervalos de confianza, que a medida que se alejan de la última observación no pronosticada, cada vez se hacen más amplios.



Conclusiones

En cuanto al objetivo número uno, pudimos observar que nuestra hipótesis sobre hacer campañas diferenciadas dependiendo del ciclo de vida y la fecha hace sentido. Insisto en que con esas gráficas podemos contestar el a quién y cuándo sería efectivo comunicar al cliente para hacerle llegar una oferta para contratación de crédito. Resulta interesante analizar la estacionalidad de la serie que se ve más claramente en los intervalos de confianza de la β_1 transformada. No olvidemos que será muy fácil además saber el momento en que se puede hacer la comunicación dado que la única variable significativa durante la mayoría del tiempo analizado

fue la del Número de cargos promedio en tres meses. También hace sentido hacia negocio pues entre más retiros tengas de tu dinero disponible, menos liquidez tendrás y, por lo tanto, más necesitarás un crédito para financiarte.

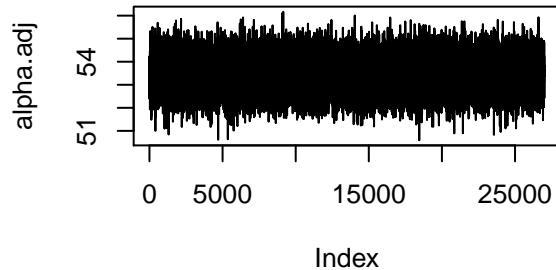
Hablando sobre el pronóstico, vemos que el modelo dinámico ajusta bastante bien, aunque como lo hemos repetido varias veces a lo largo del trabajo, esto es por la propia naturaleza del modelo. La carencia más importante del modelo es su baja sensibilidad a cambios en la tendencia tratando de predecir más de un periodo hacia adelante. Las campañas de colocación de crédito son trimestrales, por lo que la propuesta a negocio sería actualizar la cifra cada tres meses.

Referencias

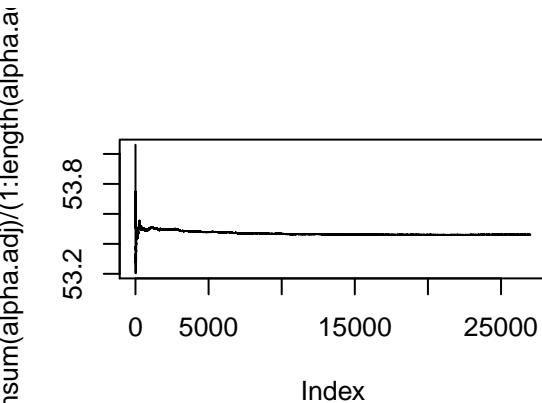
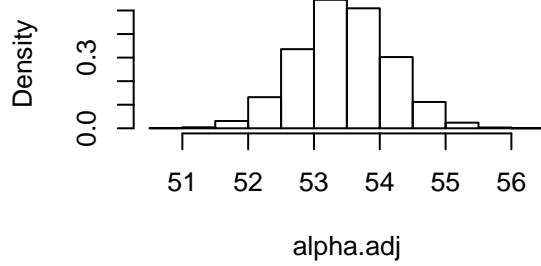
- Los datos utilizados fueron obtenidos de un Banco y constituyen una muestra de 64 mil clientes durante 24 meses (agosto 2016 a agosto 2018), en la cual se resumen las contrataciones de créditos personales de forma mensual a lo largo de este periodo. Es importante mencionar que los datos fueron modificados a petición de la institución para poder compartirlos; no obstante, no modifica la esencia del problema y la solución.
- Notas de clase del profesor Luis Enrique Nieto Barajas, en particular los capítulos 4 y 5 que hablan sobre modelos lineales generalizados y modelos dinámicos, respectivamente.
- Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. Bayesian Data Analysis, 2002, 2a edición. Chapman & Hall: Boca Raton.
- Gelman, A., Hill, J. Data Analysis Using Regression and Multilevel / Hierarchical Models, 2008, 6a edición, Cambridge University Press.

Apéndice

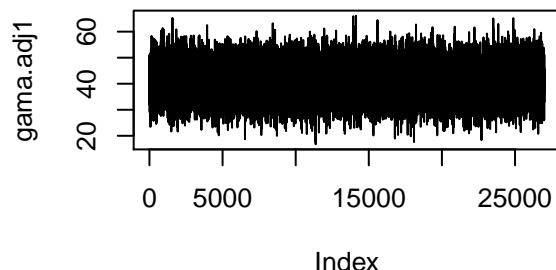
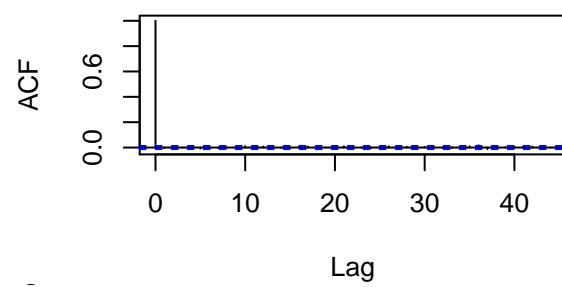
Gráficas comportamiento cadenas y parámetros



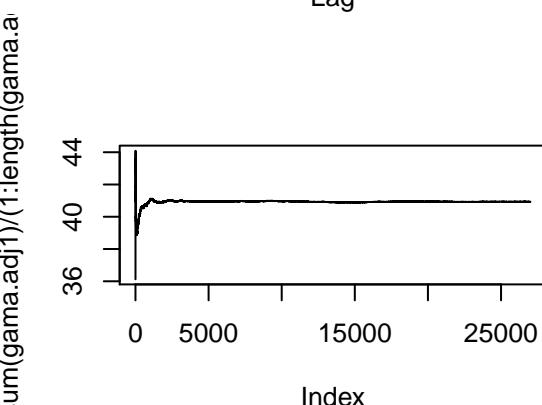
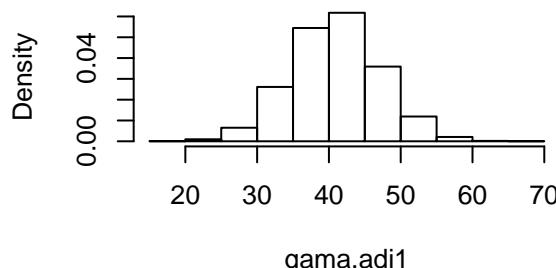
Histogram of alpha.adj



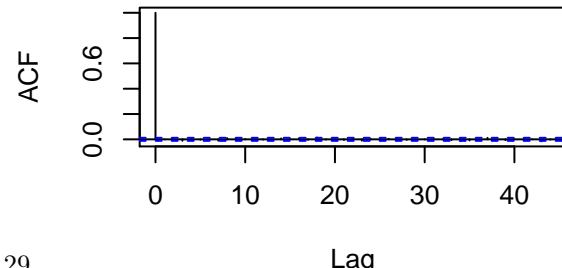
Series alpha.adj

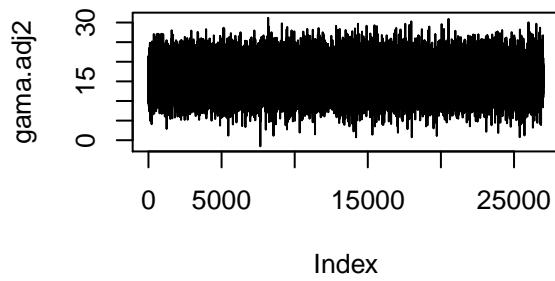


Histogram of gama.adj1

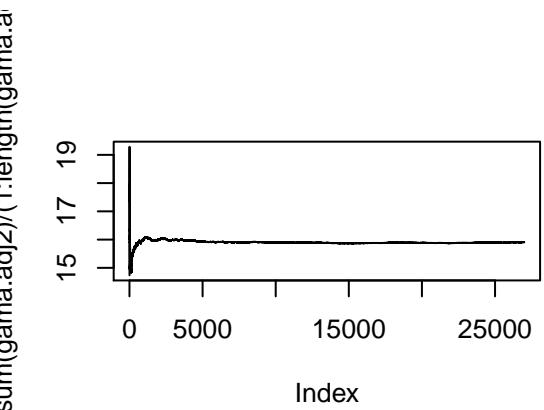
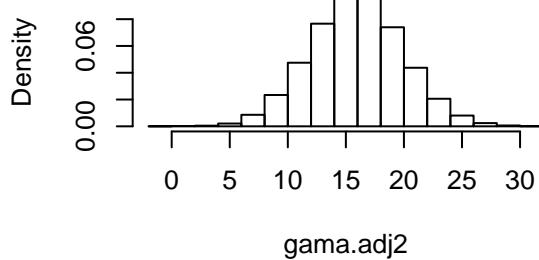


Series gama.adj1

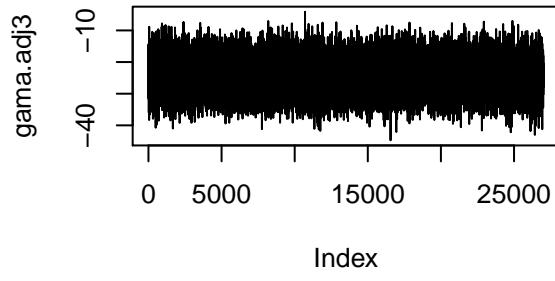
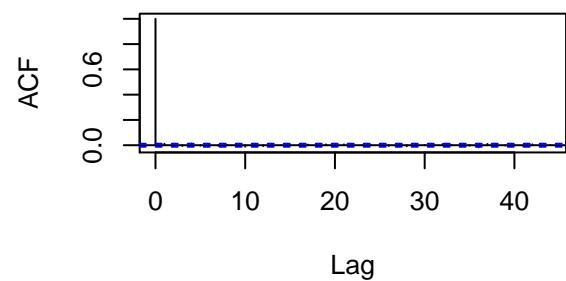




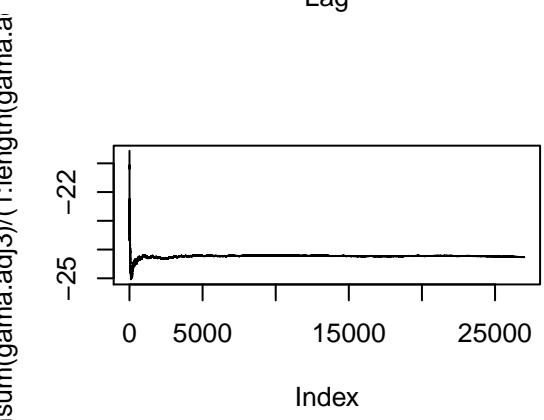
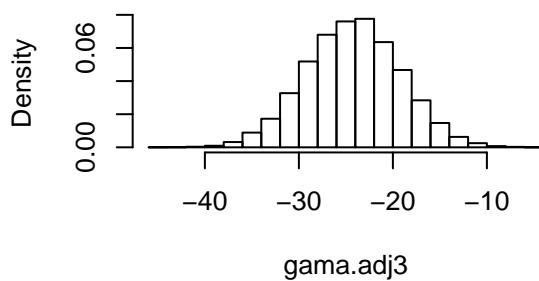
Histogram of gama.adj2



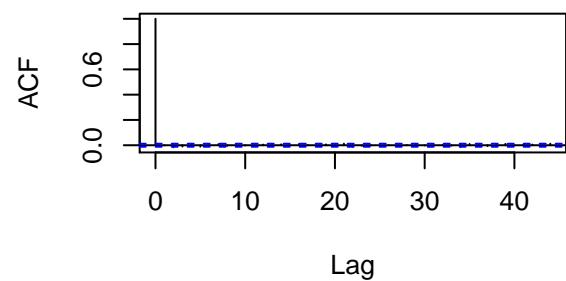
Series gama.adj2

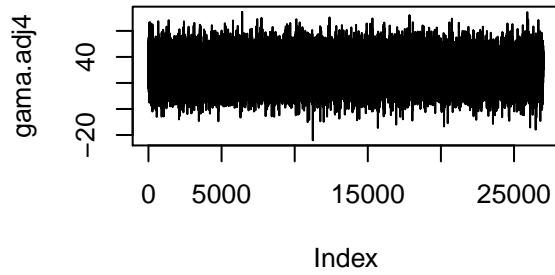


Histogram of gama.adj3

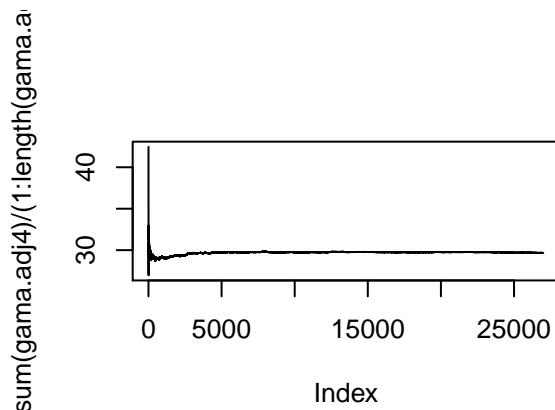
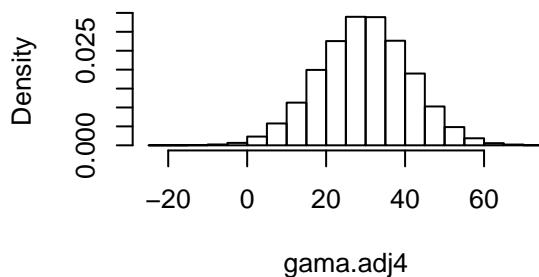


Series gama.adj3

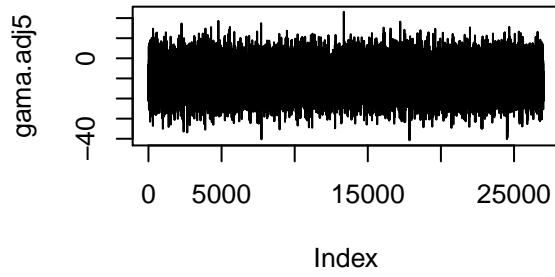
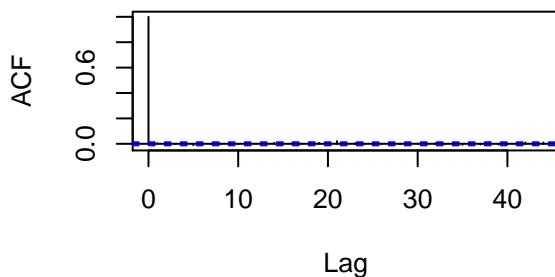




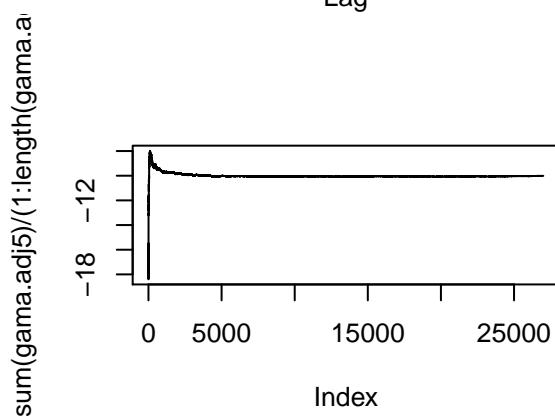
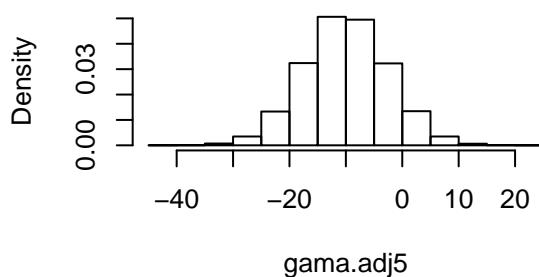
Histogram of gama.adj4



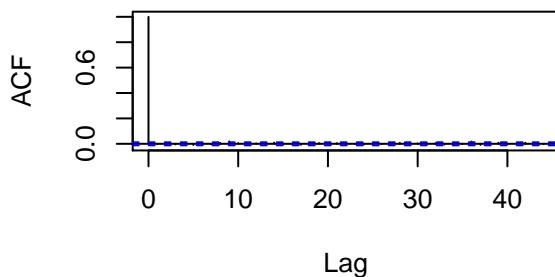
Series gama.adj4

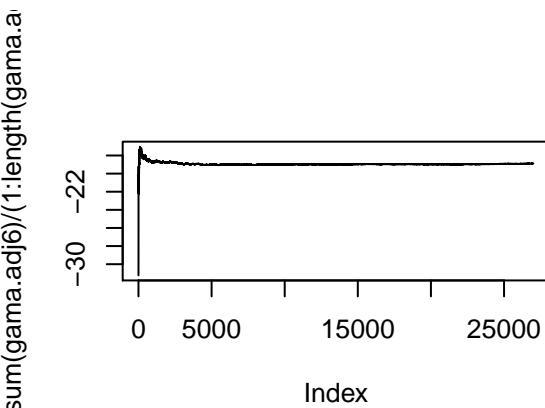
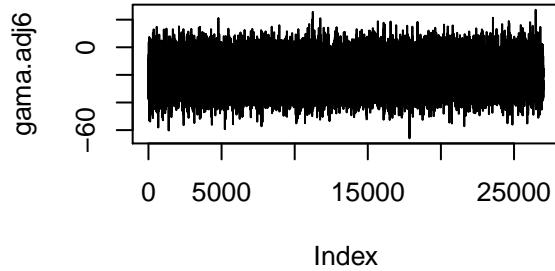


Histogram of gama.adj5

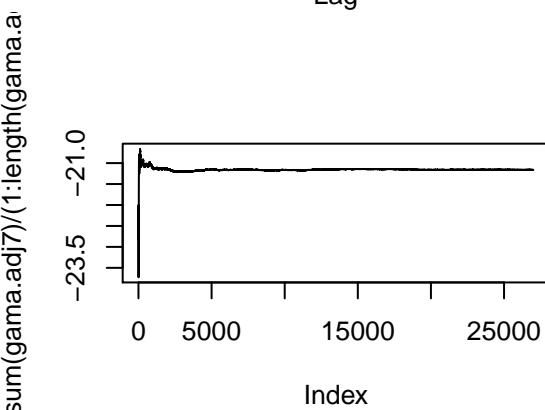
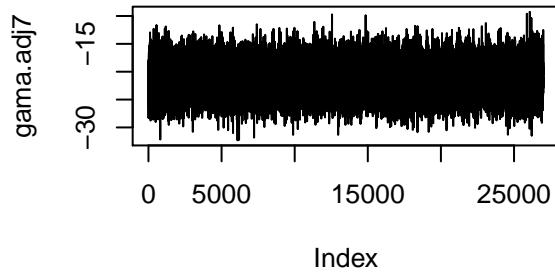
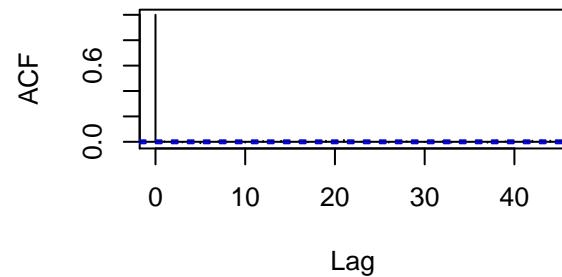
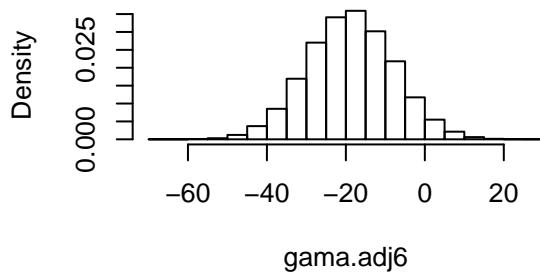


Series gama.adj5

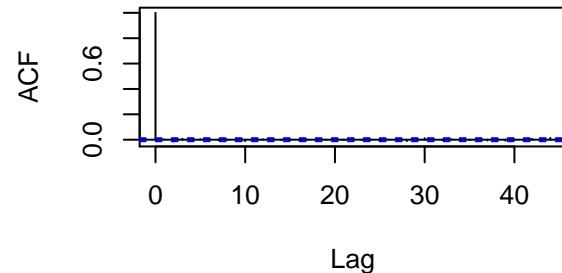
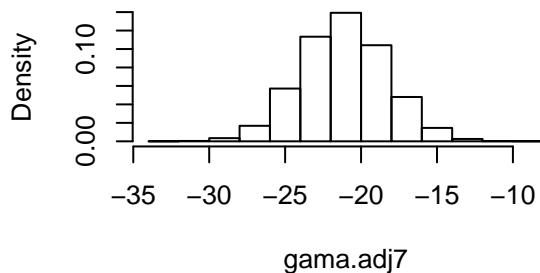


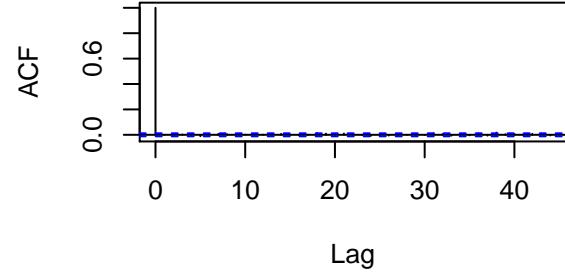
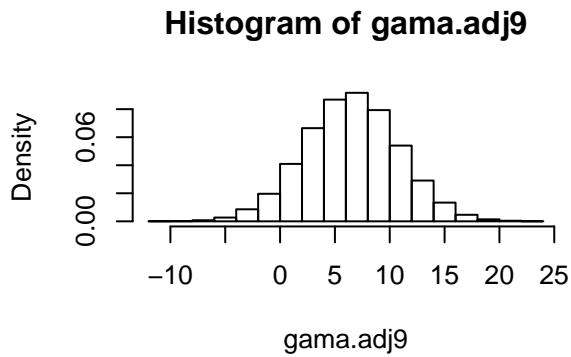
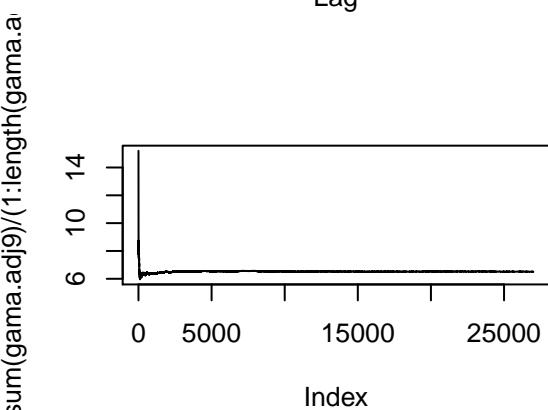
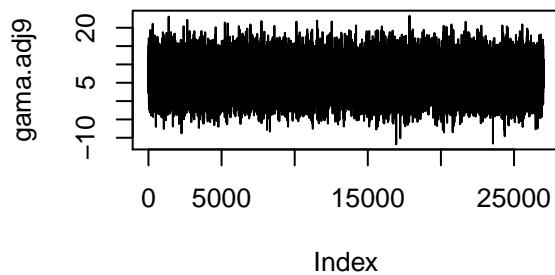
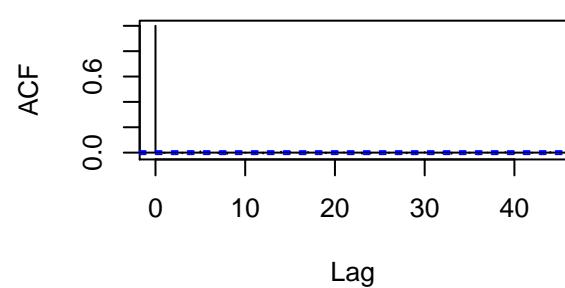
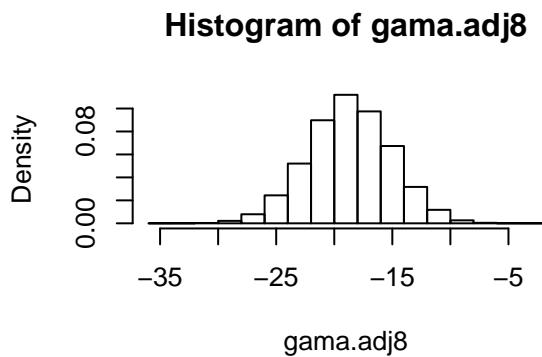
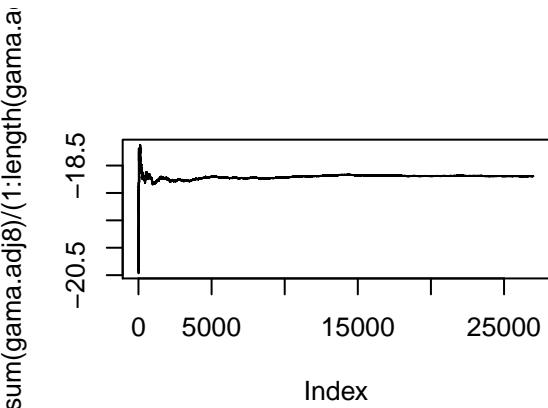
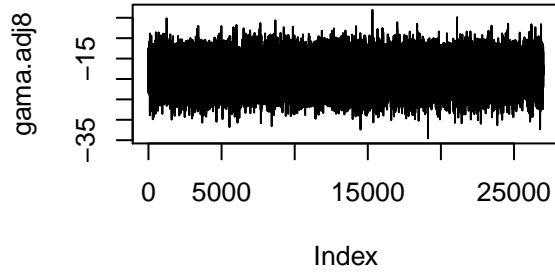


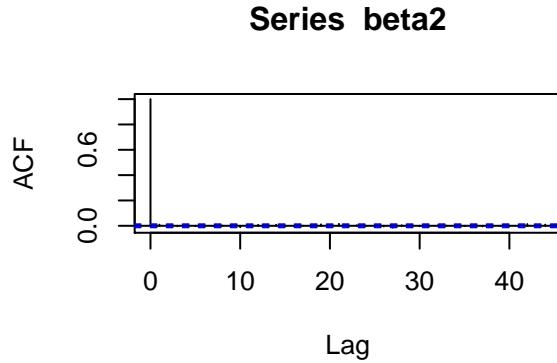
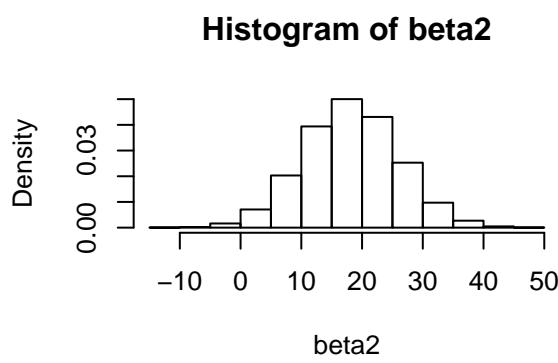
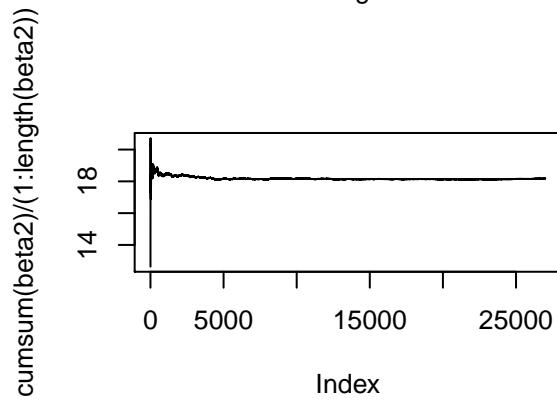
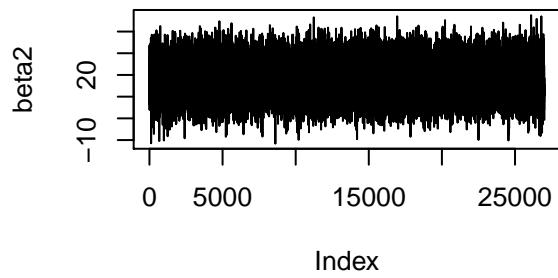
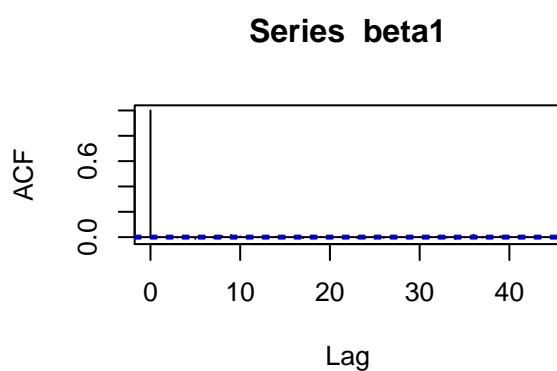
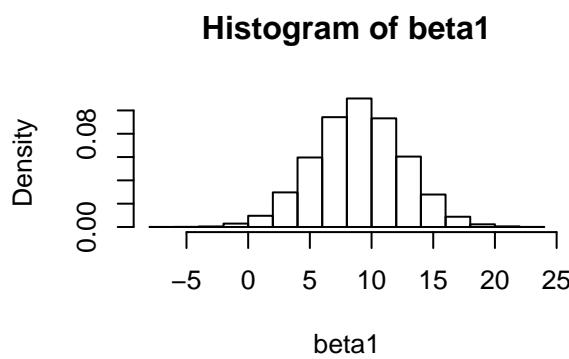
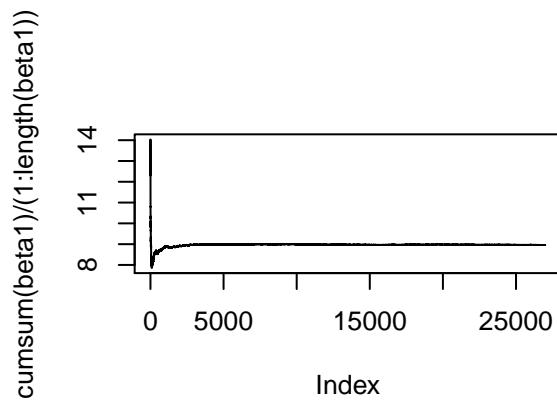
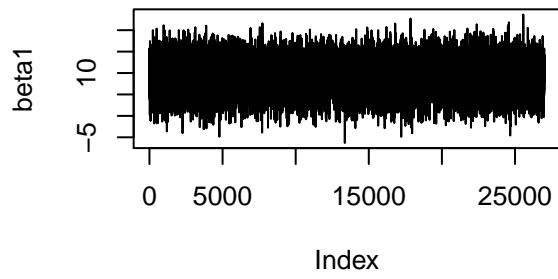
Histogram of gama.adj6

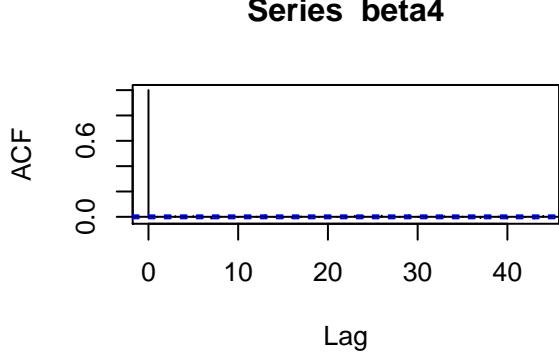
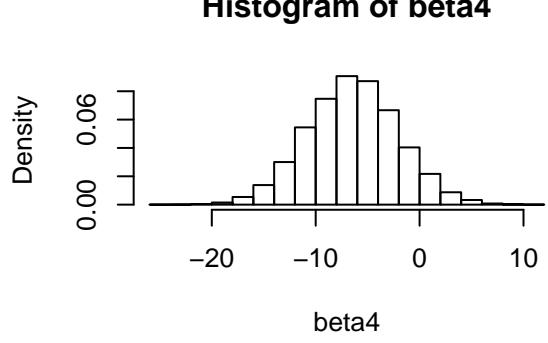
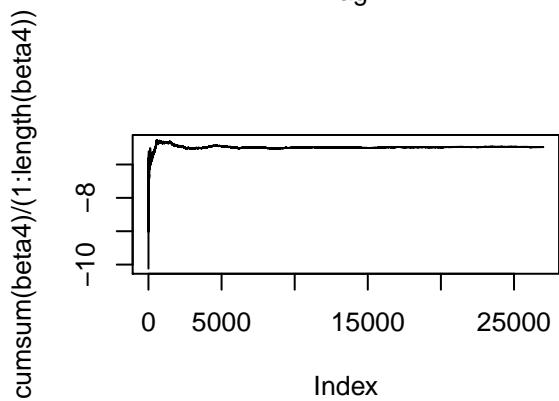
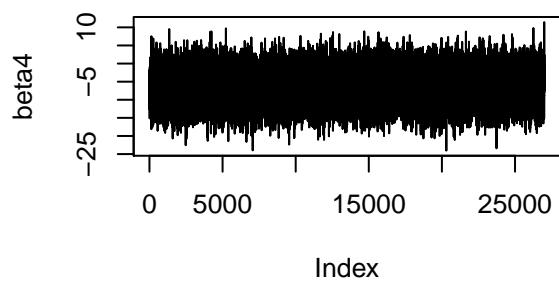
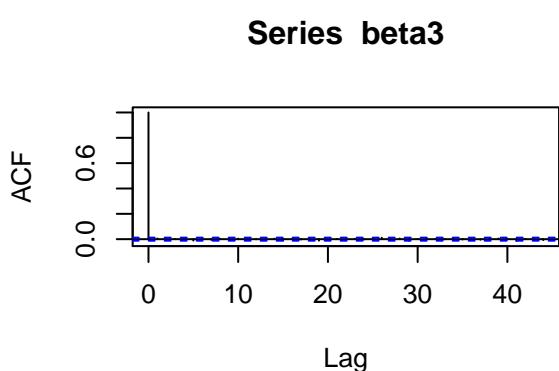
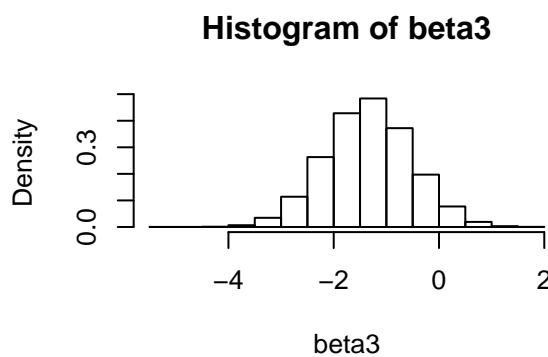
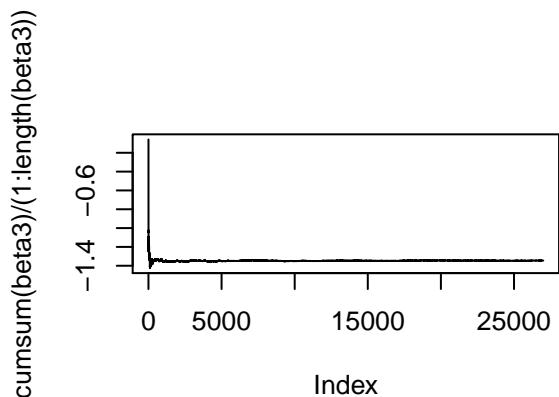
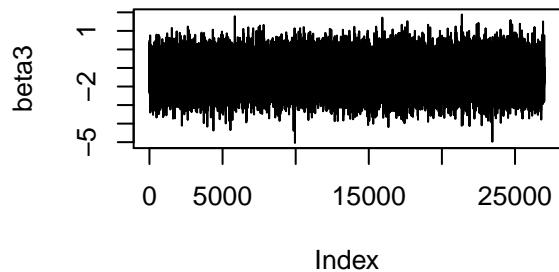


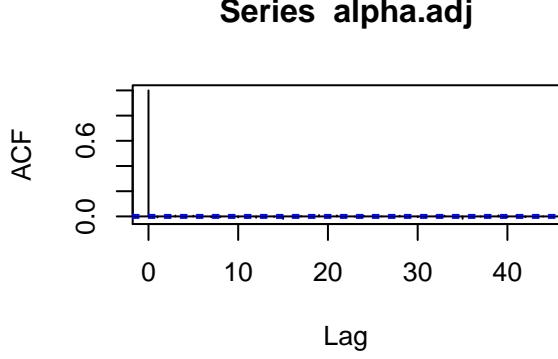
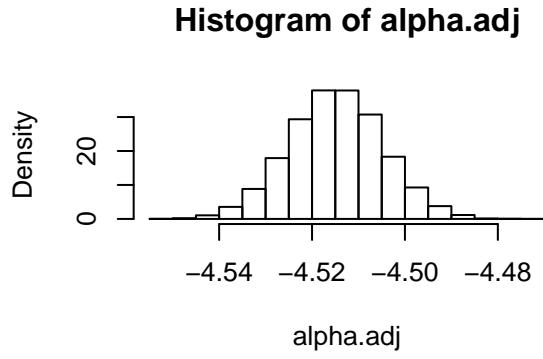
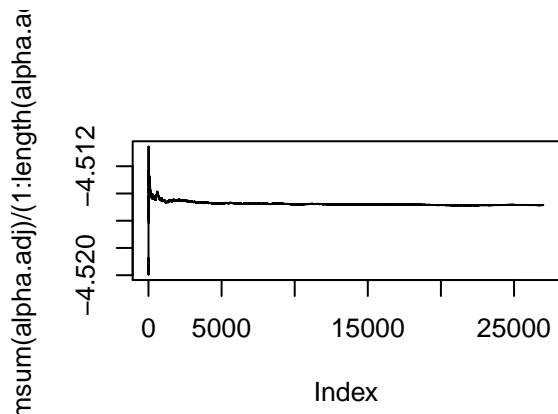
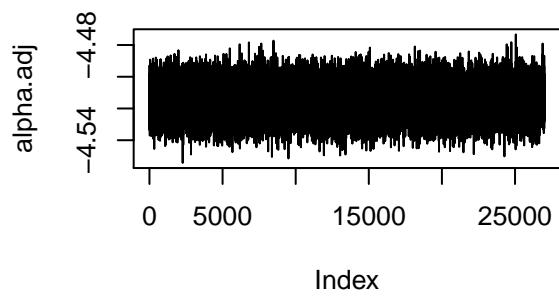
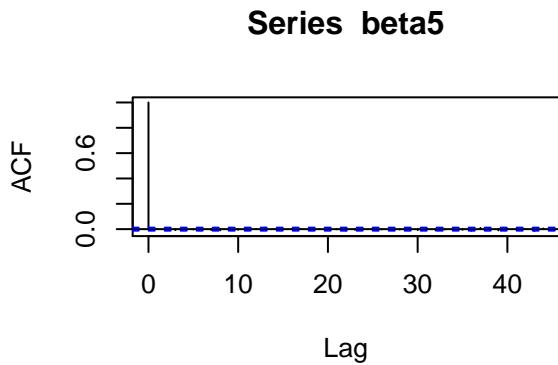
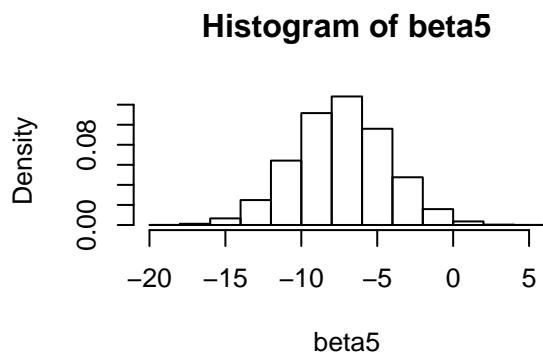
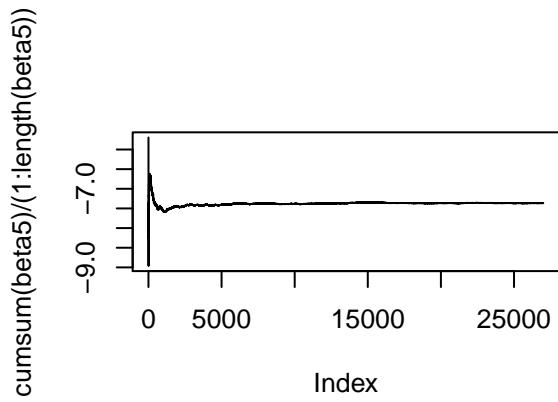
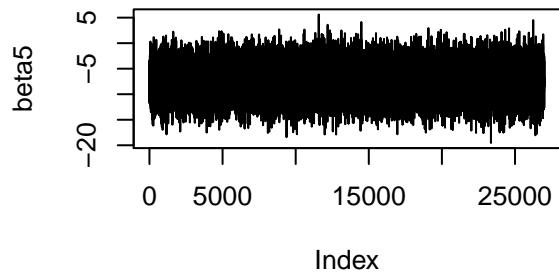
Histogram of gama.adj7

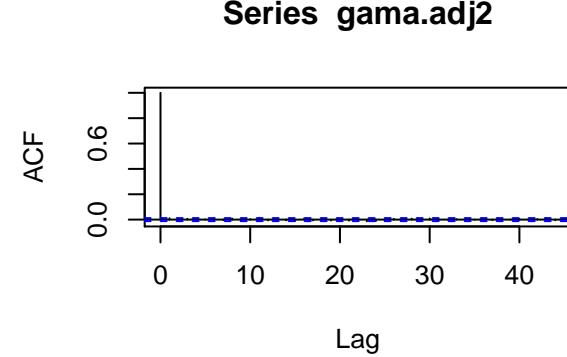
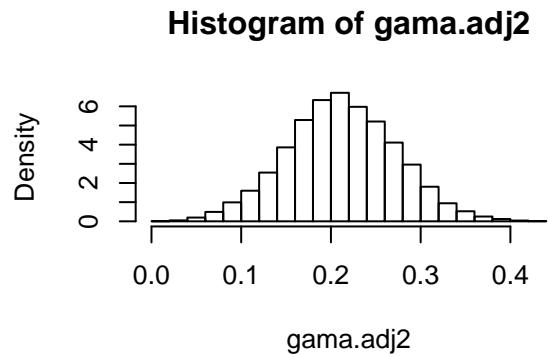
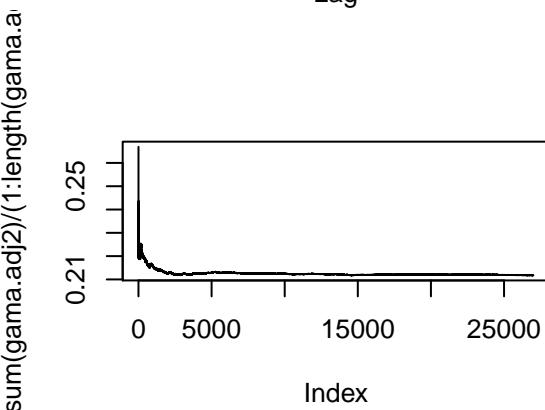
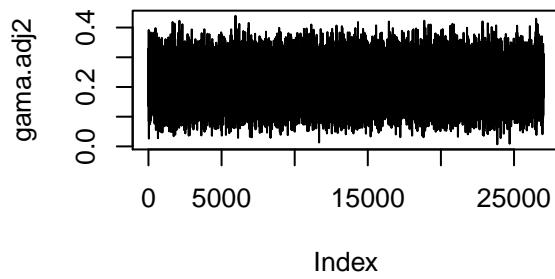
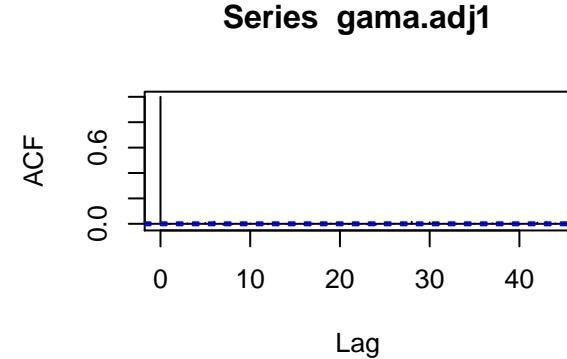
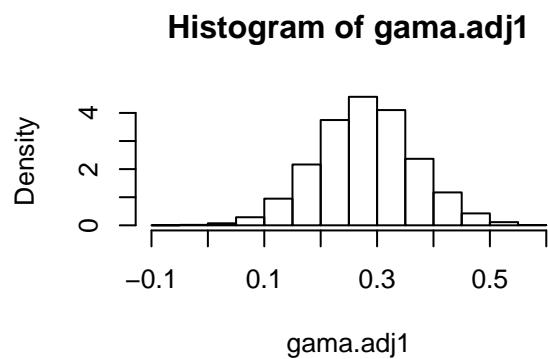
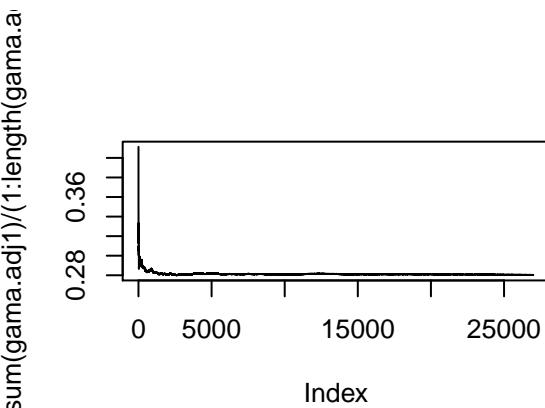
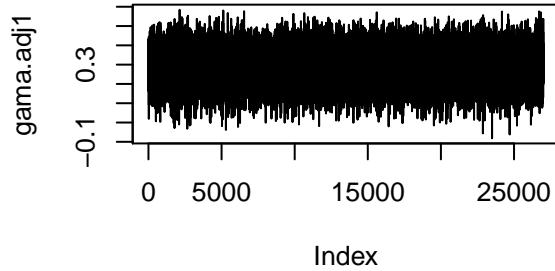


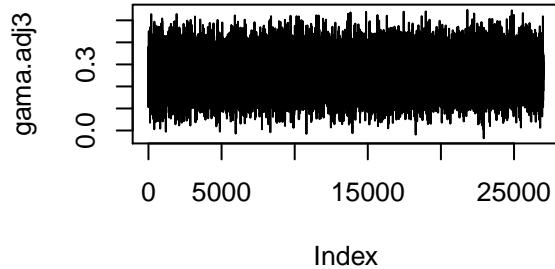




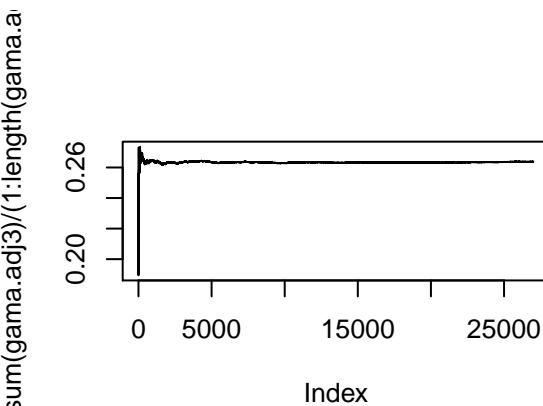
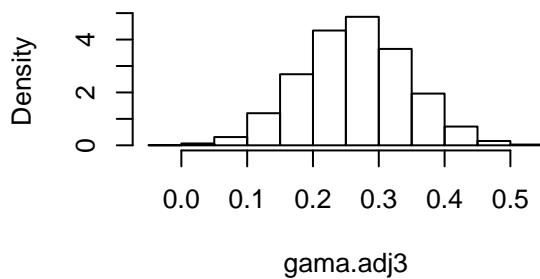




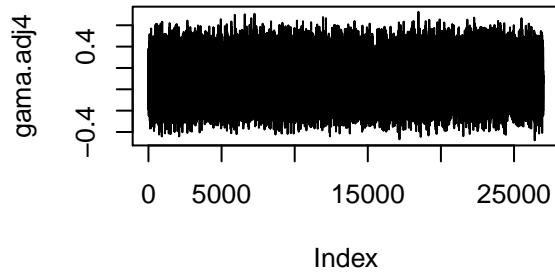
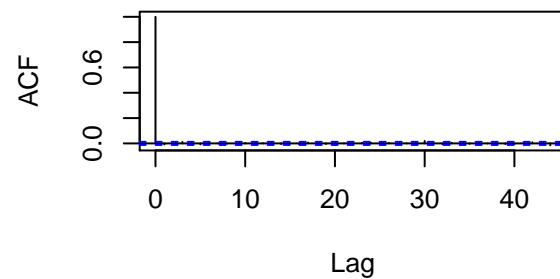




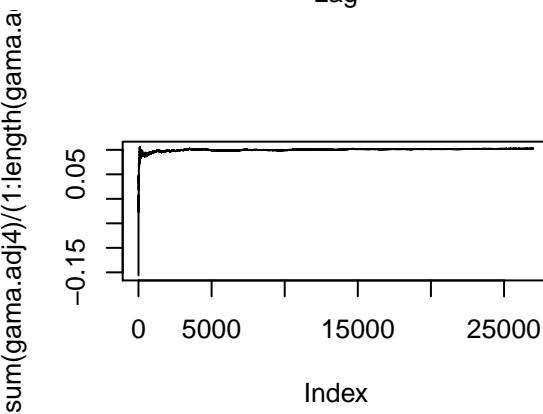
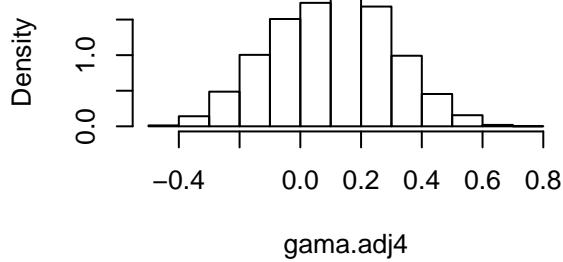
Histogram of gama.adj3



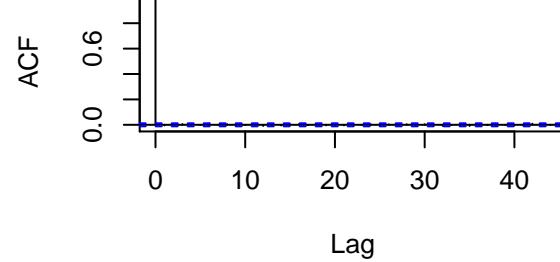
Series gama.adj3

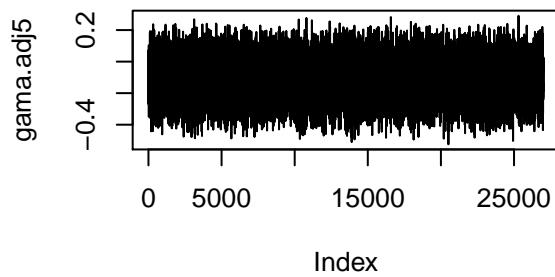


Histogram of gama.adj4

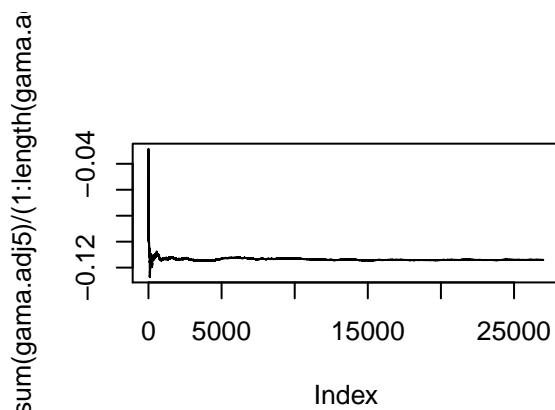
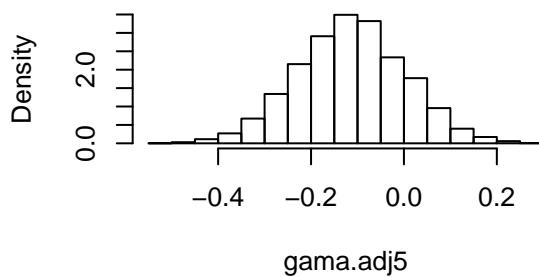


Series gama.adj4

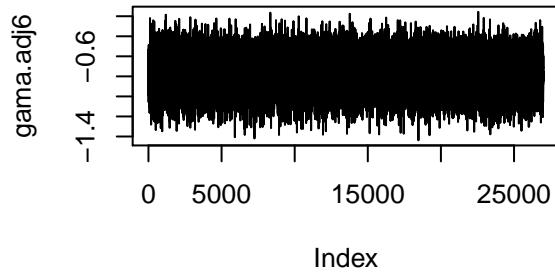
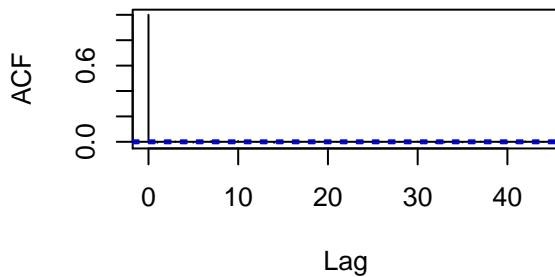




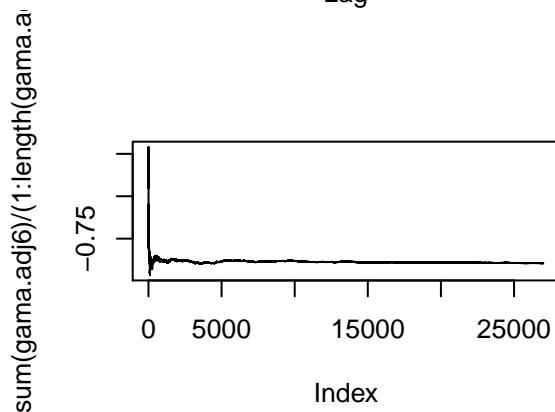
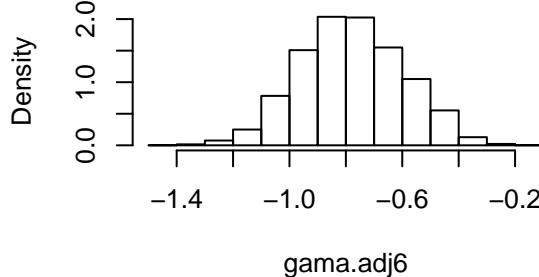
Histogram of gama.adj5



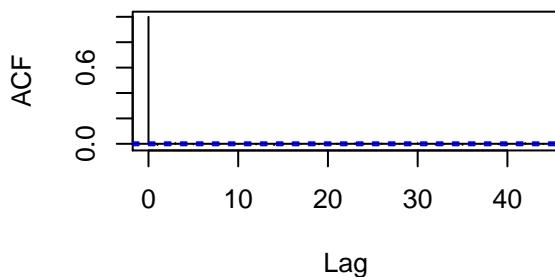
Series gama.adj5

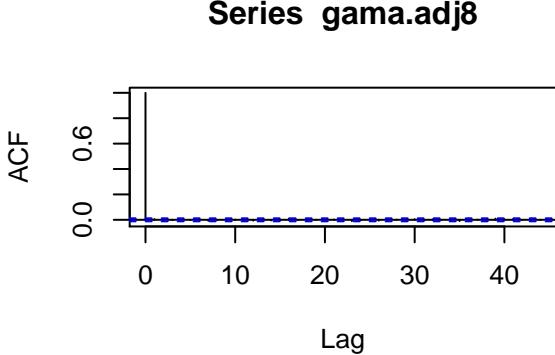
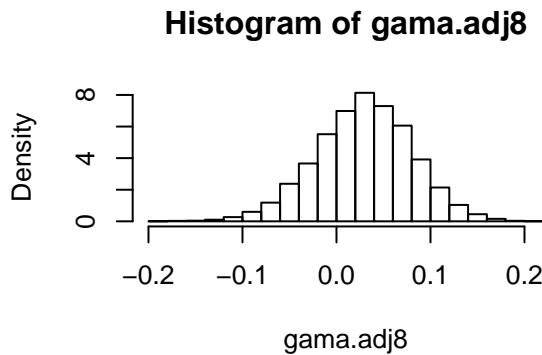
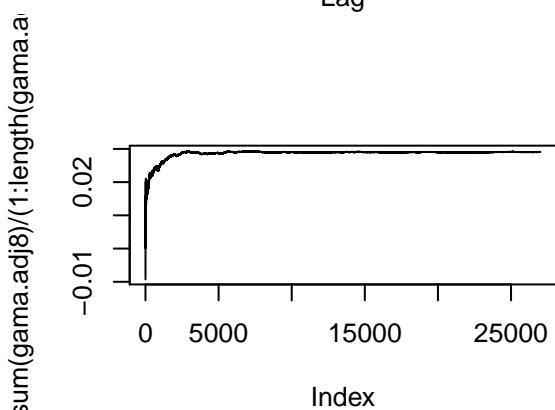
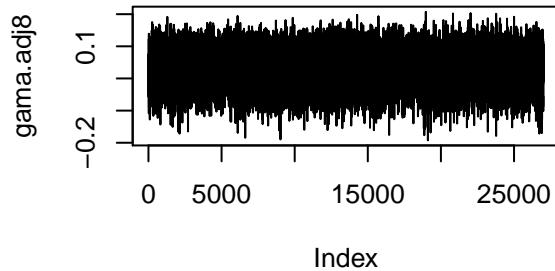
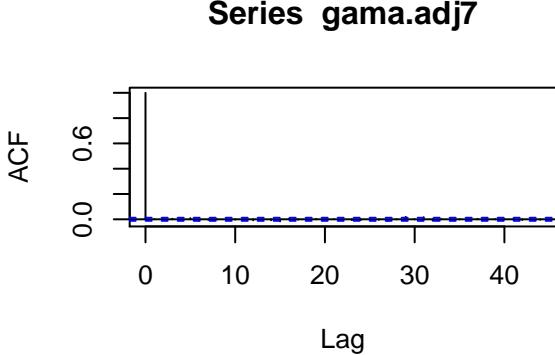
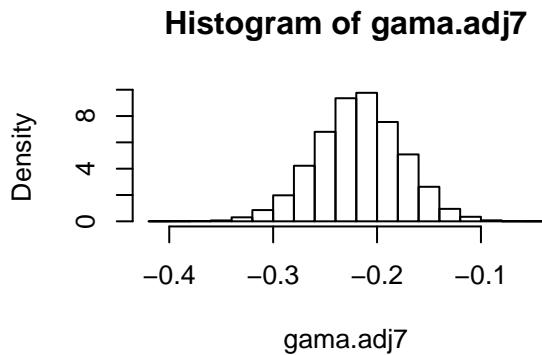
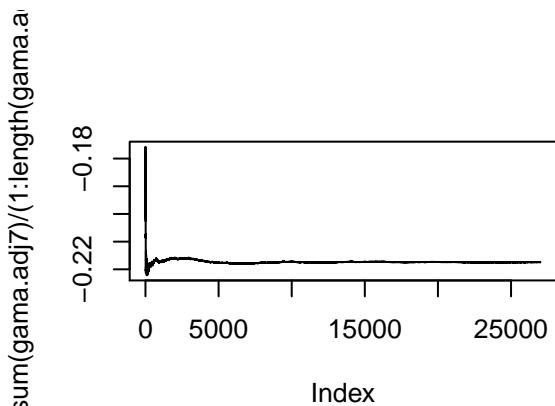
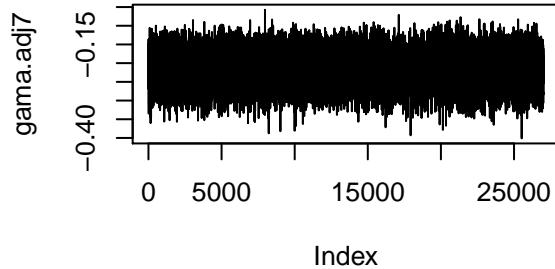


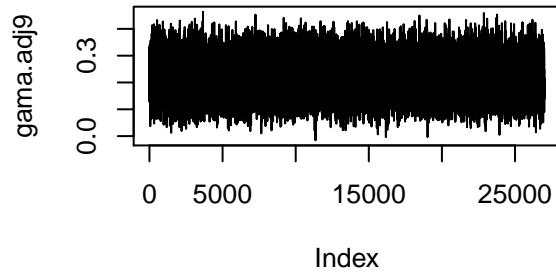
Histogram of gama.adj6



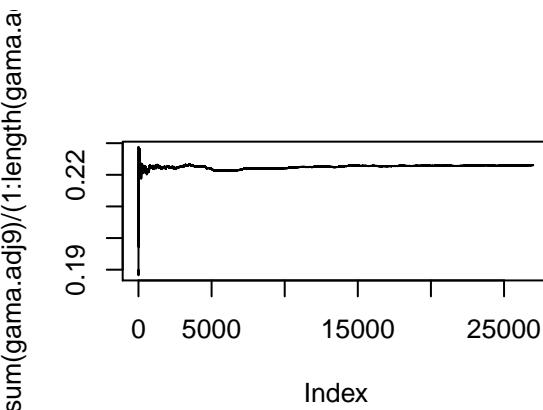
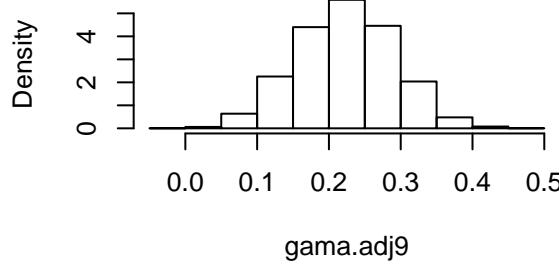
Series gama.adj6



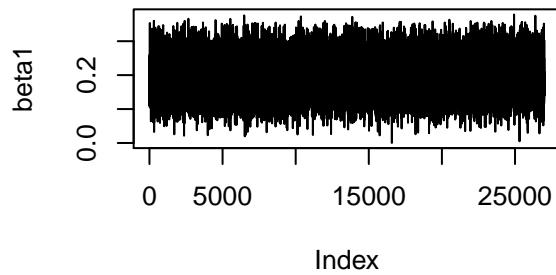
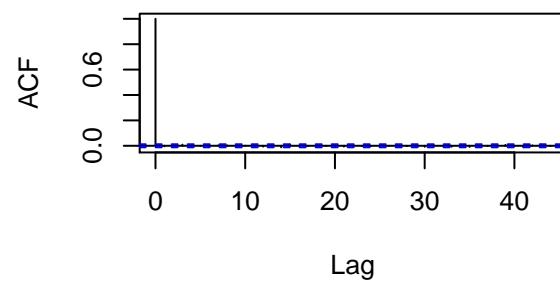




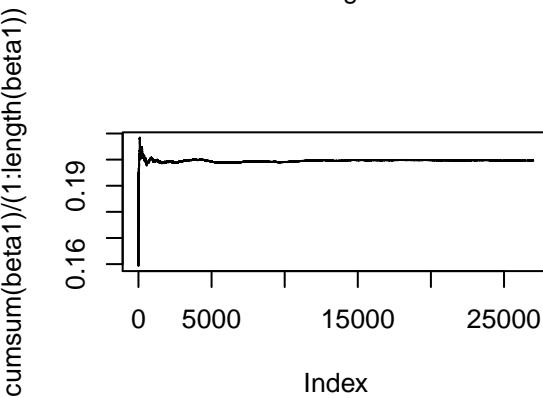
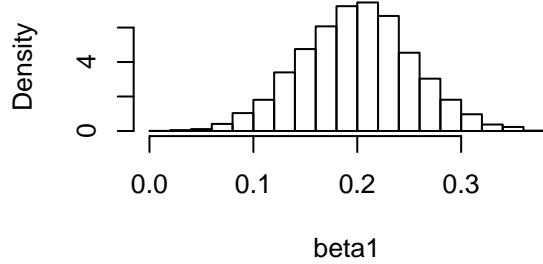
Histogram of gama.adj9



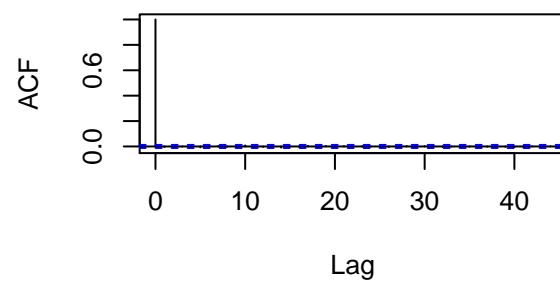
Series gama.adj9

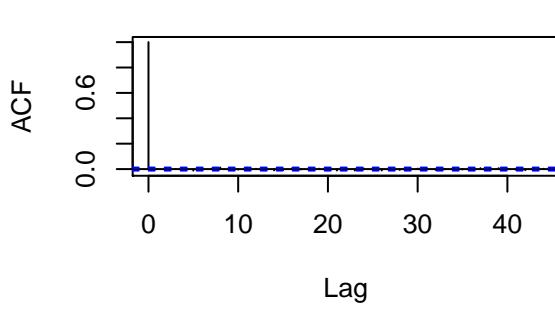
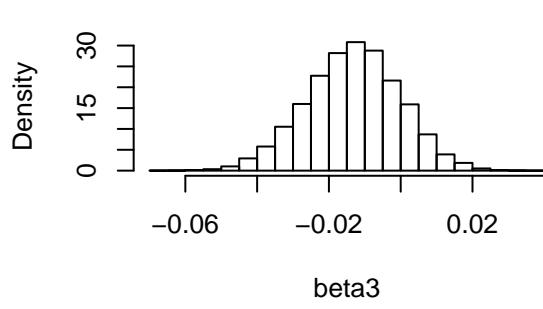
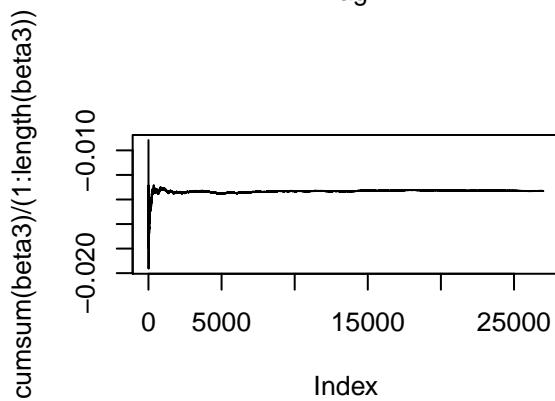
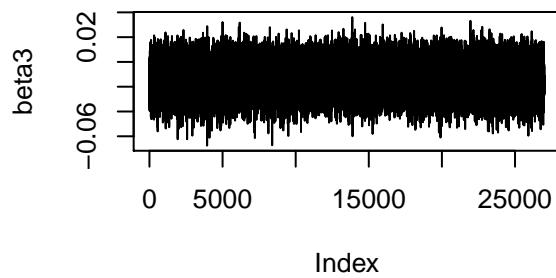
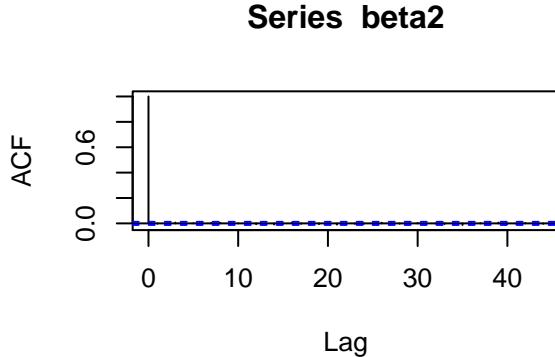
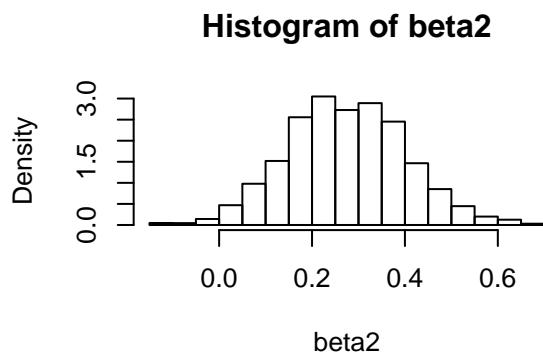
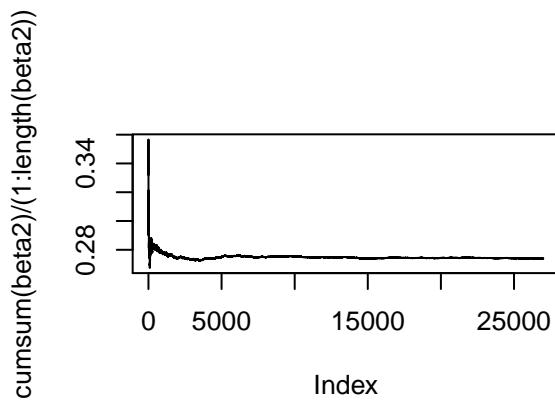
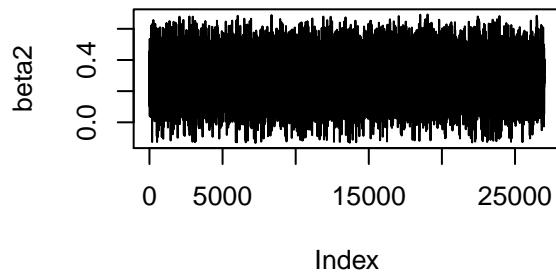


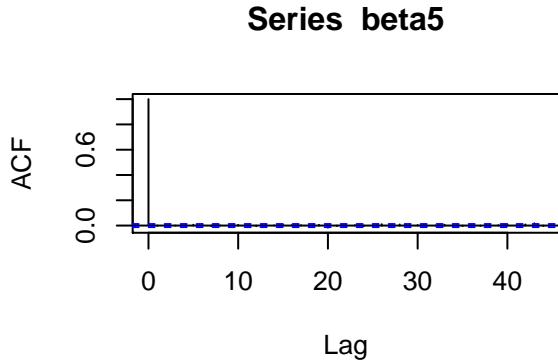
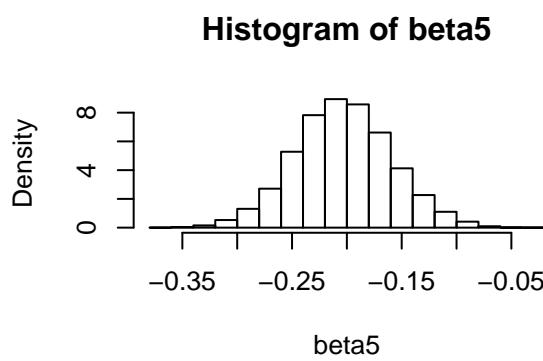
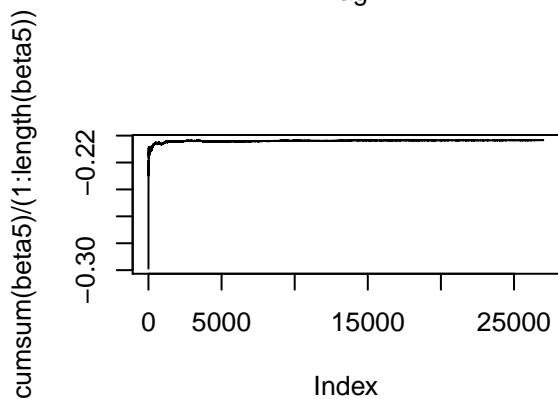
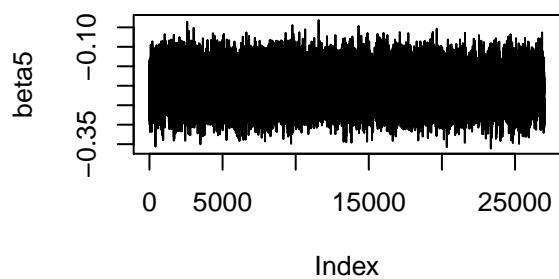
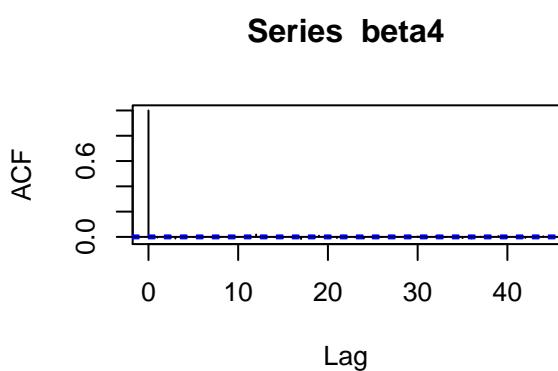
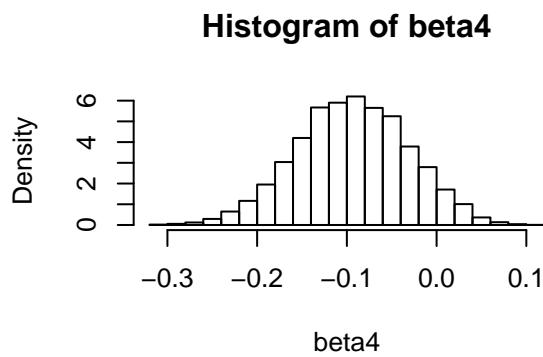
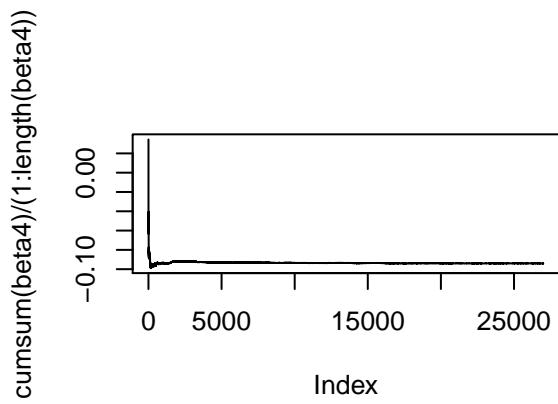
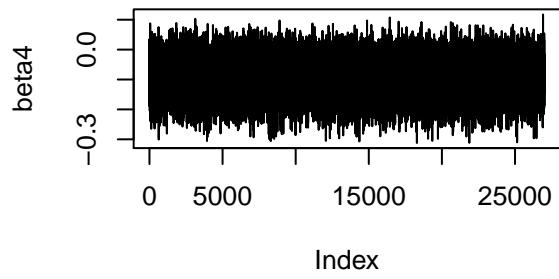
Histogram of beta1

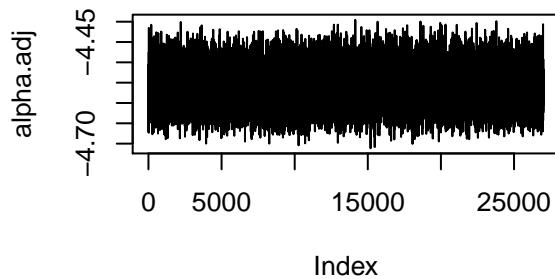


Series beta1

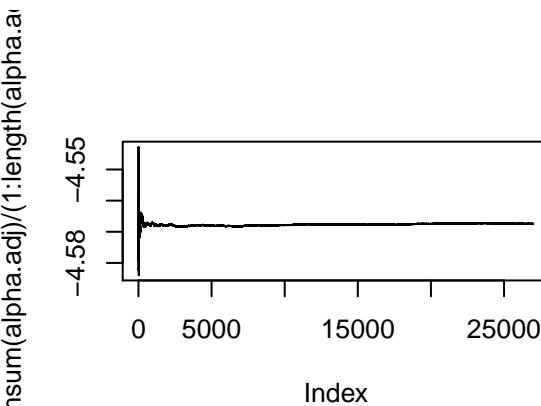
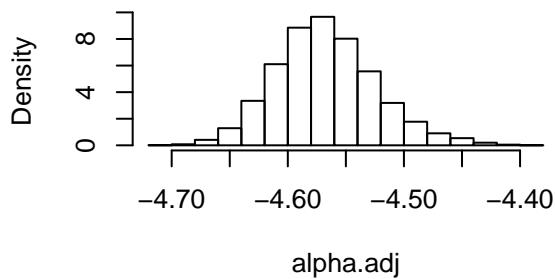




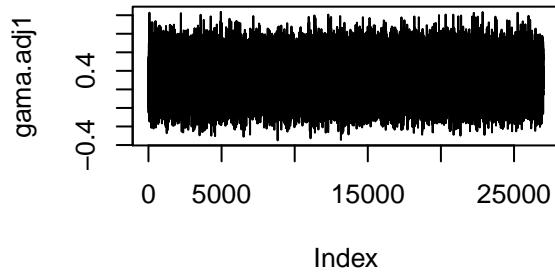
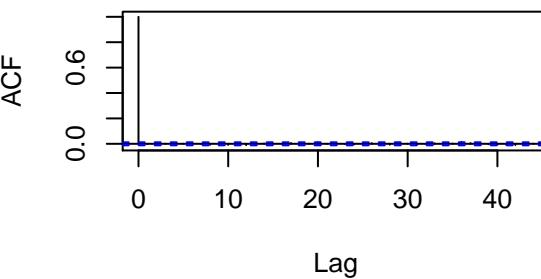




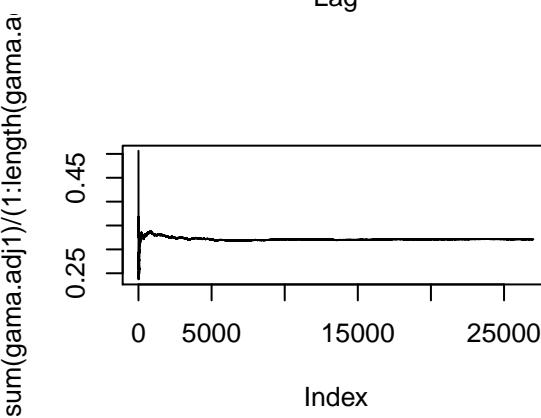
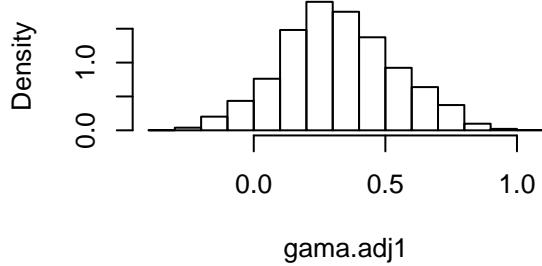
Histogram of alpha.adj



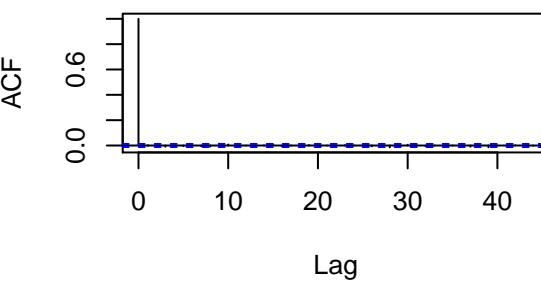
Series alpha.adj

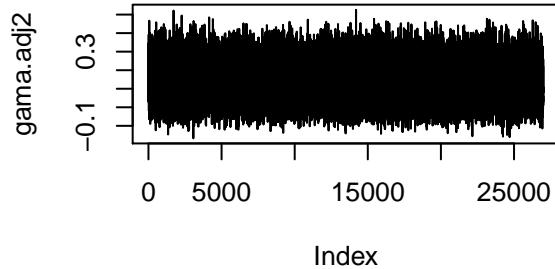


Histogram of gama.adj1

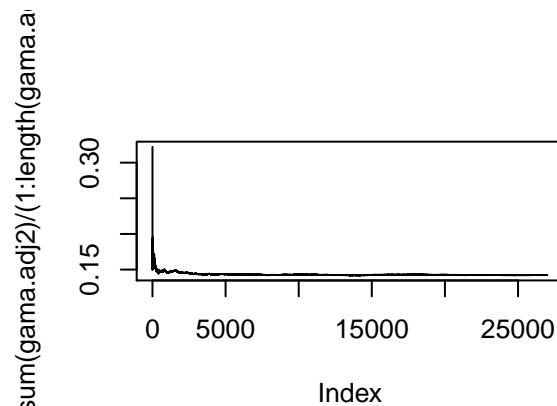
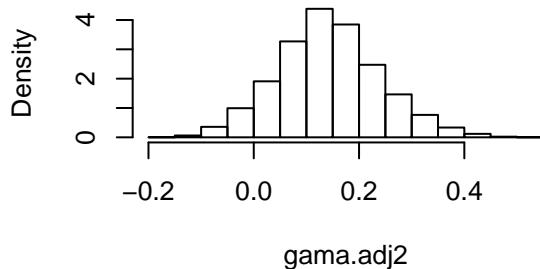


Series gama.adj1

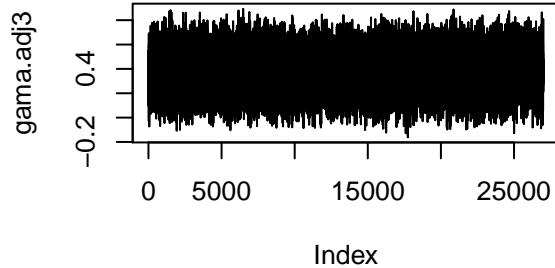
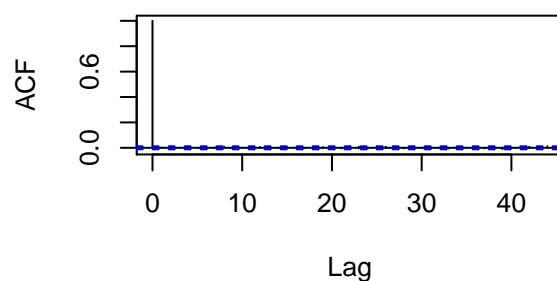




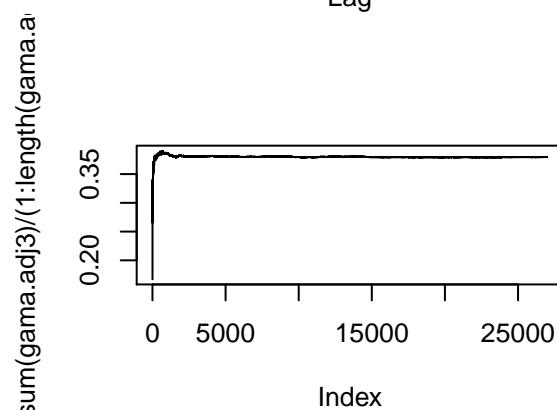
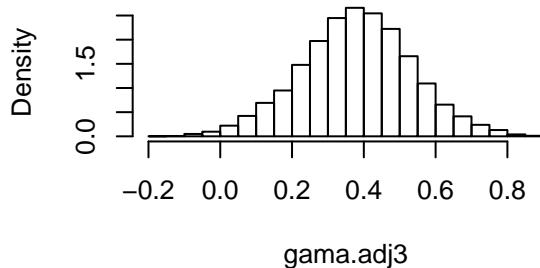
Histogram of gama.adj2



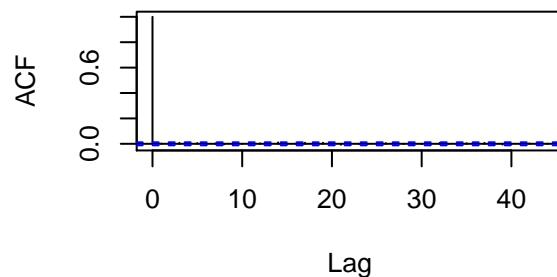
Series gama.adj2

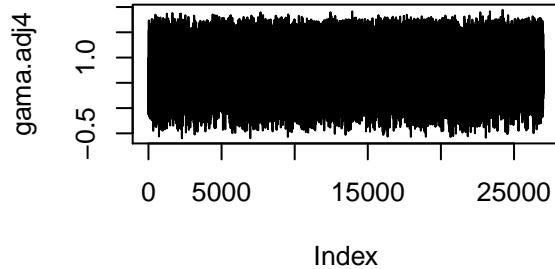


Histogram of gama.adj3

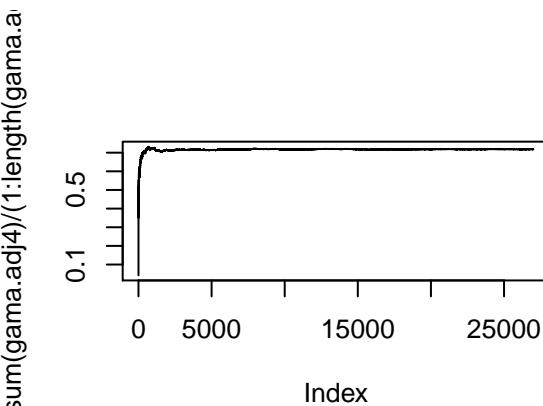
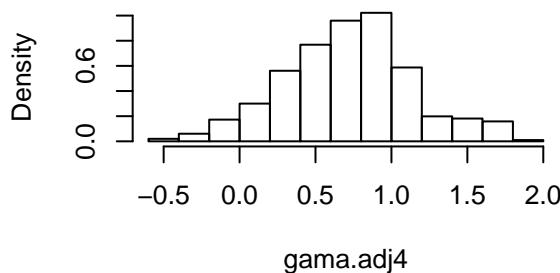


Series gama.adj3

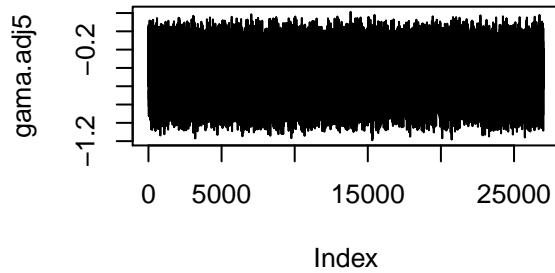
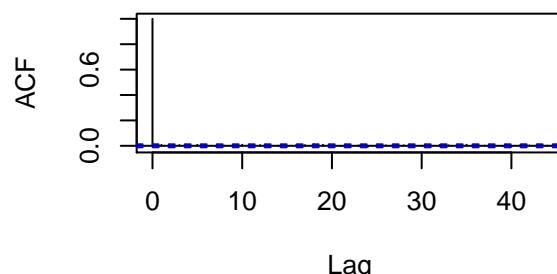




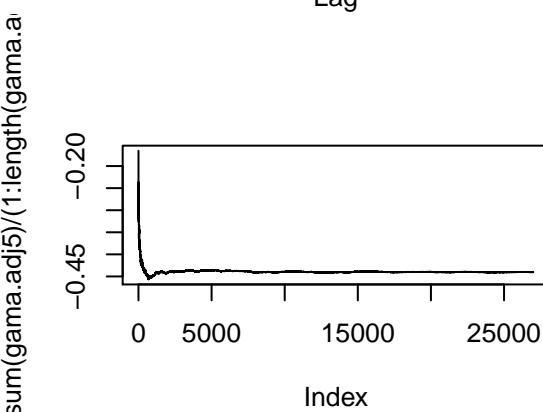
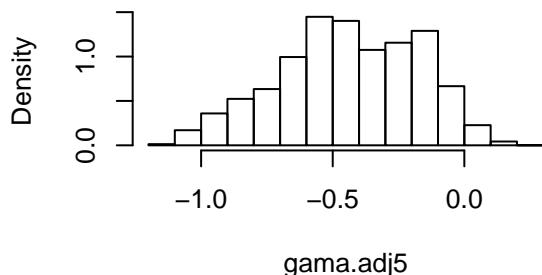
Histogram of gama.adj4



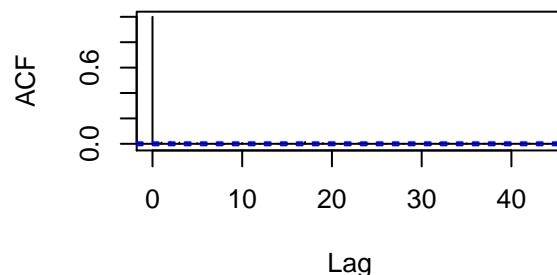
Series gama.adj4

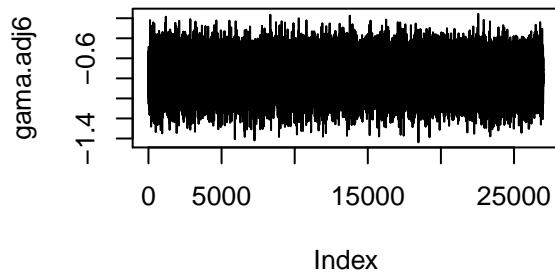


Histogram of gama.adj5

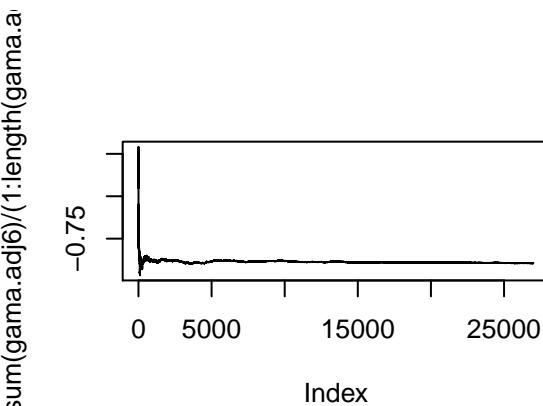
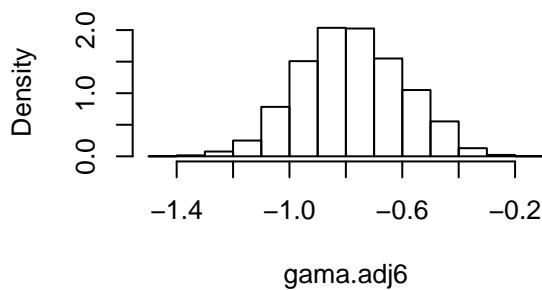


Series gama.adj5

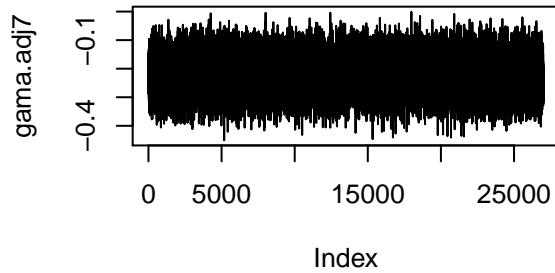
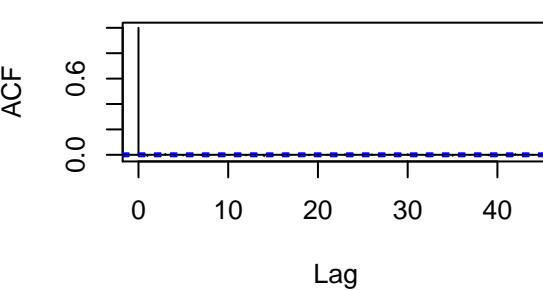




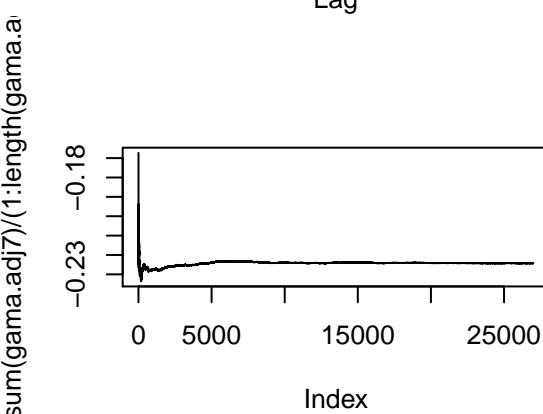
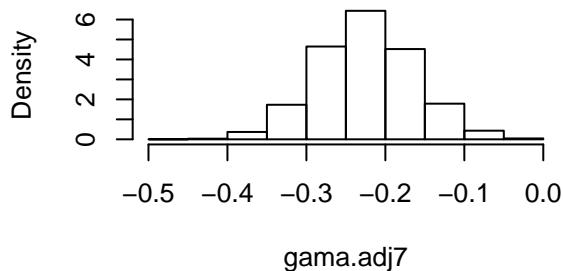
Histogram of gama.adj6



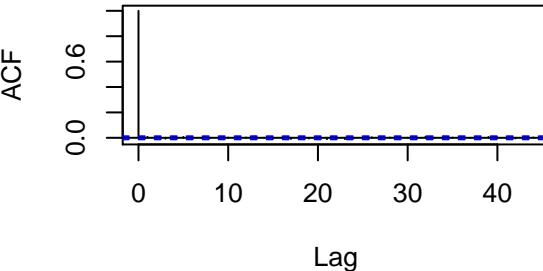
Series gama.adj6

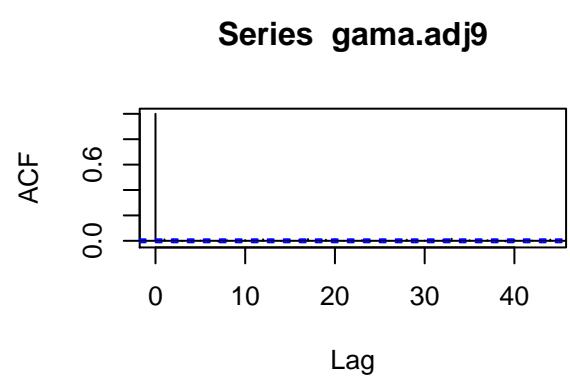
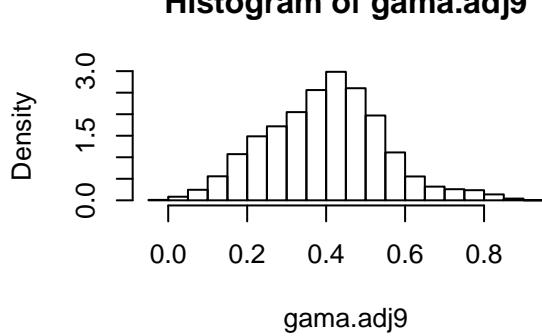
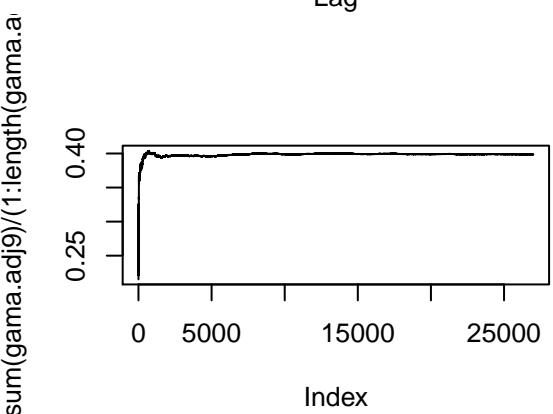
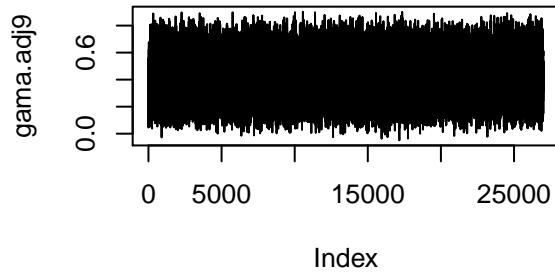
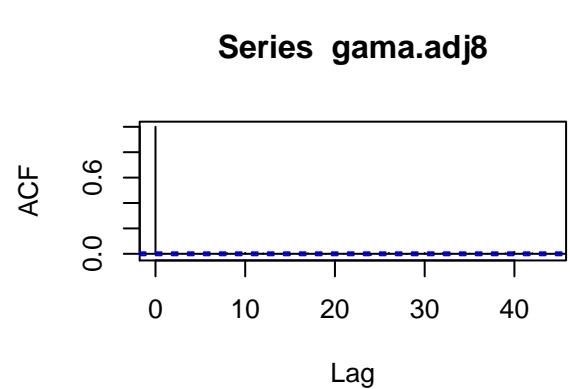
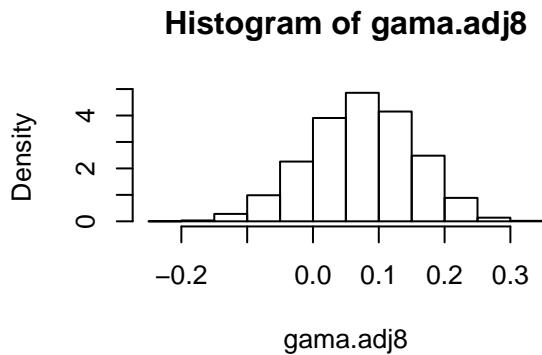
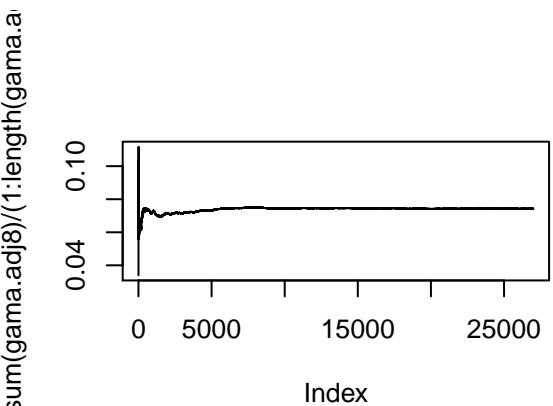
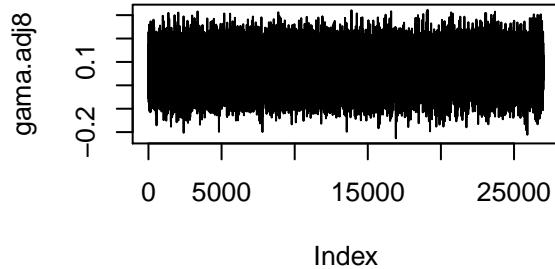


Histogram of gama.adj7



Series gama.adj7





Código JAGS

```
##### "estatico_regresion_normal.txt"
model
{
  #Likelihood
  for (i in 1:n) {
    y[i] ~ dnorm(mu[i],tau)
    mu[i]<-alpha+gama[x1[i]]+beta[1]*x2[i]+beta[2]*x3[i]+beta[3]*x4[i]+beta[4]*x5[i]+beta[5]*x6[i]
  }
  #Priors
  alpha ~ dnorm(0,0.001)

  for (j in 1:9) {
    gama[j] ~ dnorm(0,0.001)
  }

  for (k in 1:5){
    beta[k] ~ dnorm(0,0.001)
  }

  tau ~ dgamma(0.001,0.001)

  #Prediction 1
  for (i in 1:n) { yf1[i] ~ dnorm(mu[i],tau) }

  #Estimable quantities
  alpha.adj <- alpha+mean(gama[])

  for (j in 1:9) {
    gama.adj[j] <- gama[j]-mean(gama[])
  }

}

#### "estatico_glm_poisson.txt"
model
{
  #Likelihood
  for (i in 1:n) {
    y[i] ~ dpois(mu[i])
    mu[i]<-ne[i]*lambda[i]
    log(lambda[i])<-alpha+gama[x1[i]]+beta[1]*x2[i]+beta[2]*x3[i]+beta[3]*x4[i]+beta[4]*x5[i]+beta[5]*x6[i]
  }
  #Priors
  alpha ~ dnorm(0,0.001)

  for (j in 1:9) {
    gama[j] ~ dnorm(0,0.001)
  }

  for (k in 1:5){
    beta[k] ~ dnorm(0,0.001)
  }
}
```

```

#Prediction 1
for (i in 1:n) {
  yf1[i] ~ dpois(mu[i])
}

#Estimable quantities
alpha.adj <- alpha+mean(gama[])
for (j in 1:9) {
  gama.adj[j] <- gama[j]-mean(gama[])
}
}

##### "poisson_dinamico.txt"
model
{
  #Likelihood
  for (i in 1:n) {
    y[i] ~ dpois(mu[i])
    mu[i]<-ne[i]*lambda[i]
    log(lambda[i])<-alpha+gama[x1[i]]+beta1[i]*x2[i]+beta2[i]*x3[i]+beta3[i]*x4[i]+beta4[i]*x5[i]+beta5[i]
  }

  for (s in (10):n) {
    beta1[s] ~ dnorm(beta1[s-9], tau)
    beta2[s] ~ dnorm(beta2[s-9], tau)
    beta3[s] ~ dnorm(beta3[s-9], tau)
    beta4[s] ~ dnorm(beta4[s-9], tau)
    beta5[s] ~ dnorm(beta5[s-9], tau)
  }

  for (w in 1:9){
    beta1[w] ~ dnorm(0,0.001)
    beta2[w] ~ dnorm(0,0.001)
    beta3[w] ~ dnorm(0,0.001)
    beta4[w] ~ dnorm(0,0.001)
    beta5[w] ~ dnorm(0,0.001)
  }
  tau ~ dgamma(100, 1)

  #Priors
  alpha ~ dnorm(0,0.001)

  for (j in 1:9) {
    gama[j] ~ dnorm(0,0.001)
  }

#Prediction 1
for (i in 1:n) {
  yf1[i] ~ dpois(mu[i])
}

```

```
#Estimable quantities
alpha.adj <- alpha+mean(gama[])
for (j in 1:9) {
  gama.adj[j] <- gama[j]-mean(gama[])
}
}
```