PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

Regresión beta usando cópulas gaussianas para analizar series de tiempo

TESIS PARA OPTAR POR EL GRADO DE MAGÍSTER EN ESTADÍSTICA

Presentado por:

Ana Rosa Cajavilca Gonzales

Asesora: Dra. Zaida Jesús Quiroz Cornejo

Miembros del jurado:
Dr. Luis Hilmar Valdivieso Serrano
Dr. Luis Enrique Benites Sánchez
Dra. Zaida Jesús Quiroz Cornejo

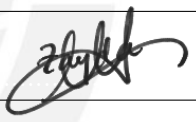
Lima, 2022

Informe de Similitud

Yo Zaida Jesús Quiroz Cornejo docente de la Escuela de Posgrado de la Pontificia Universidad Católica del Perú, asesora de la tesis titulada *Regresión beta usando cópulas gaussianas para analizar series de tiempo*, de la autora Ana Rosa Cajavilca Gonzales, dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 22 %. Así lo consigna el reporte de similitud emitido por el software Turnitin el 11/08/2020.
- He revisado con detalle dicho reporte y la Tesis, y no se advierte indicios de plagio.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lima, 15 de diciembre de 2022

Apellidos y nombres de la asesora: Quiroz Cornejo Zaida Jesús	
DNI: 43704124	Firma: 
ORCID: https://orcid.org/0000-0003-3821-0815	

Agradecimentos

Expreso mis agradecimientos a la profesora Zaida Quiroz por la orientación, paciencia y compromiso para la elaboración de este trabajo.

A los profesores por su retroalimentación en cada fase presentada.

A mis amigos de la maestría y a mis familiares por el acompañamiento y comprensión en el transcurso de este aprendizaje.

Resumen

Este trabajo presenta una alternativa para analizar series de tiempo que se encuentran restringidas al intervalo $(0, 1)$. Se detalla el modelo propuesto Masarotto y Varin (2012) y Guolo y Varin (2014), el cual permite capturar los efectos producidos por covariables a través de una regresión beta y adicionalmente, con el empleo de cópulas permite modelar la dependencia temporal mediante un proceso de autorregresivo de medias móviles. Como ventaja de la aplicación de este modelo se tiene que evita la necesidad de transformar la variable dependiente, así como también evita someterla al cumplimiento de diversos supuestos como los de normalidad y estacionariedad. Además, permite diferenciar los efectos de las covariables y de la dependencia temporal, lo cual coadyuva a mejorar el análisis de los resultados. Se realizó una aplicación a la tasa de desempleo desde enero de 2003 hasta octubre de 2019 en Lima Metropolitana y debido a la distribución que presenta esta variable se usó un modelo de regresión beta usando cópulas gaussianas. Para la estimación se incluyó el logaritmo del índice del PBI, así como un componente de estacionalidad anual como covariables y para tomar en cuenta la dependencia temporal se incorporó un proceso autorregresivo de medias móviles ARMA$(1, 1)$ a través de una cópula gaussiana.

Palabras-clave: ARMA, cópulas gaussianas, regresión beta, series de tiempo, tasa de desempleo.

Abstract

This work presents an alternative to analyze time series restricted to the interval $(0, 1)$. This model was proposed by Masarotto y Varin (2012) and Guolo y Varin (2014), which allows to capture covariates effects through a beta regression and additionally allows to model the temporal dependence by copulas through an autoregressive moving averages process. As an advantage of the application of this model, it is not necessary to transform the dependent variable or subject to compliance the assumptions such as normality and stationarity. Also, it allows to differentiate the effects of the covariates and of the temporal dependence, which helps to improve the analysis of results. An application to the unemployment rate from January 2003 to October 2019 in Metropolitan Lima was implemented and due to the distribution presented by this variable it was used a gaussian copula beta regression model. The model includes the logarithm of the GDP index and annual seasonality component as covariates, and to take into account the temporal dependence it was included an autoregressive moving averages process ARMA$(1, 1)$ through a gaussian copula.

Key-words: ARMA, gaussian copula, beta regression, time series, unemployment rate.

Índice general

Capítulo 1

Introducción

1.1. Consideraciones preliminares

Las variables de series de tiempo que tienen restricciones en el intervalo $(0, 1)$ (por ejemplo, tasas o proporciones) usualmente reciben tratamientos que consisten en la transformación de las mismas para obtener series estacionarias; así como también la aplicación de una serie de supuestos (por ejemplo, normalidad), los cuales en muchos casos no se cumplen para el conjunto de datos. En este contexto, las metodologías empleadas pueden no ser las más adecuadas al brindar resultados limitados.

Los modelos dinámicos bajo inferencia bayesiana representan una alternativa para el análisis de series de tiempo. Sin embargo, la desventaja de estos enfoques es que la inferencia y predicción es complicada debido a la presencia de variables latentes correlacionadas. En particular, en este proyecto se propone analizar una serie de tiempo de proporciones o tasas, a través de un modelo de regresión beta incorporando cópulas gaussianas, que representa una alternativa práctica para este tipo de análisis.

Una cópula describe la estructura de dependencia entre variables aleatorias a través de una función de distribución acumulada, evitando satisfacer supuestos que muchas conjuntos de datos no superan y recurrir a transformaciones de la serie. Así también, se debe tener en cuenta que, el empleo de un tipo de cópula específico resulta independiente a las trasformaciones monotónicas que se pueden realizar a la serie, lo cual ayuda a la interpretación de los resultados. En este contexto, la importancia de emplear el modelo propuesto radica en el uso de la serie en su forma original y cabe precisar que, el empleo de cópulas para determinar la dependencia temporal no afecta la elección de distribuciones marginales consideradas ni los resultados obtenidos de estas.

Para fines de este proyecto se analizará la estructura de dependencia temporal de la tasa de desempleo, cuyos valores se encuentran en el intervalo $(0, 1)$, La tasa de desempleo es un indicador importante de las condiciones del mercado laboral. Según el INEI, hasta octubre de 2019, esta variable ha superado marginalmente el diez por ciento de la Población Económicamente Activa (PEA) para Lima Metropolitana; sin embargo, con la identificación de sus determinantes, su comportamiento y su tendencia, es posible lograr la intervención adecuada que asegure mejores condiciones económicas y sociales.

1.2. Objetivos de la tesis

El objetivo general es aplicar un modelo de regresión beta usando cópulas gaussianas que permita modelar series de tiempo restringidas en el intervalo $(0, 1)$.

Objetivos específicos

- Revisar la literatura de modelos para el análisis de series de tiempo, donde la variable respuesta es una variable continua que se encuentra restringida al intervalo $(0, 1)$.
- Revisar y estudiar métodos de inferencia aplicados al modelo de regresión con cópulas gaussianas que permiten el uso de variables en su forma original.
- Implementar métodos de simulación para conocer el desempeño del modelo propuesto.
- Aplicar el modelo al conjunto de datos de la tasa de desempleo en Lima Metropolitana.

1.3. Revisión de la literatura

Los modelos de regresión se emplean para determinar la relación que hay entre una variable dependiente y una o más variables explicativas. Si bien uno de los modelos más empleados es el de regresión lineal, existen conjuntos de datos con particularidades, como las variables respuesta que se encuentran restringidas en el intervalo $(0, 1)$, donde para la utilización del modelo mencionado se requiere transformaciones a la variable o la aplicación de una serie de supuestos, lo cual puede presentar implicancias en los resultados de la estimación y predicción, al generar valores ajustados que excedan la restricción del intervalo.

Al respecto, Kieschnick y McCullough (2003) proponen un modelo de regresión paramétrica basado en la distribución beta, la cual permite modelar una variable cuyo soporte se encuentra en el intervalo $(0, 1)$ y se caracteriza por la flexibilidad para modelar proporciones, ya que su densidad puede adoptar diferentes formas en función a los valores de los parámetros.

Ferrari y Cribari-Neto (2004) muestran que transformar la variable dependiente conlleva a inconvenientes para la interpretación de parámetros en términos de la variable respuesta original, puesto que este tipo de variables respuesta suelen presentar asimetría, y por lo tanto la base de inferencia bajo el supuesto de normalidad no resulta apropiado. Ante eso, propusieron un modelo de regresión beta que permite modelar la media de la variable respuesta mediante funciones de enlace logit y considerar un parámetro de precisión constante. La estimación de los parámetros del modelo es realizada mediante máxima verosimilitud y al ser un modelo heterocedástico permite adoptar las ventajas de la flexibilidad de la distribución beta. De esta manera, los parámetros resultan interpretables en términos de la media de la variable respuesta. Posteriormente en el trabajo de Simas et al. (2010), extendieron el modelo anterior a uno más general asumiendo que el parámetro de precisión ya no es constante, sino modelado mediante una función de enlace.

Se debe tener en cuenta que la dependencia puede surgir de varias formas como, por ejemplo, mediciones repetidas sobre la misma unidad, observaciones recopiladas secuencialmente en el tiempo o por datos georeferenciados. Para el análisis de la dependencia temporal de este tipo de variables respuesta, Rocha y Cribari-Neto (2009) propusieron un modelo dinámico para las variables aleatorias observadas que presentan distribución beta a lo largo del tiempo.

Se trata de un modelo de media móvil autorregresivo (ARMA) que se adapta a las asimetrías de los datos y también a la dispersión no constante. Para esto, consideraron que tanto para la regresión como para el análisis de series de tiempo es más conveniente trabajar con la respuesta media y también con un parámetro de precisión. Asimismo, para el análisis de serie de tiempo, en la regresión incorporaron regresores y un componente sistemático ARMA de forma aditiva. El modelo fue aplicado al análisis y predicción de la tasa de desempleo, donde al igual que los modelos anteriores, la estimación se hizo por máxima verosimilitud y eligieron el mejor modelo con base en los criterios de información Akaike (AIC) y Bayesiano (BIC). Por otro lado, Bayer et al. (2018) propusieron modelos de regresión beta autorregresivo de media móvil estacional (SARMA) para modelar y pronosticar datos de series temporales que asumen valores en el intervalo $(0, 1)$ y están sujetas a fluctuaciones estacionales.

Los modelos dinámicos bajo inferencia bayesiana también representan una alternativa para el análisis de series de tiempo, por ejemplo Da-Silva et al. (2011) desarrollaron un modelo de regresión beta bayesiano dinámico para modelar y pronosticar series temporales de proporciones limitadas al intervalo $(0, 1)$, el cual se caracteriza por incluir que el parámetro de la media del modelo podía variar con el tiempo. Posteriormente, Da-Silva y Migon (2011) desarrollaron un modelo beta bayesiano dinámico jerárquico para modelar el mismo tipo de serie de proporciones, donde a diferencia del anterior, incluyeron ajuste dinámico a la media y los parámetros de dispersión mediante funciones de enlace. Así también, Jara et al. (2013) propusieron un modelo autorregresivo donde cada respuesta se encuentra marginalmente distribuida con una distribución beta, donde la dependencia se explica gracias a la introducción de variables latentes con una especificación jerárquica, este modelo permite su aplicación a series estacionarias y no estacionarias.

Según Masarotto y Varin (2012), la desventaja del empleo de modelos jerárquicos se encuentra en que la inferencia y predicción es complicada debido a la presencia de variables latentes correlacionadas. También indica que, bajo un contexto de independencia entre las variables, los estimadores que obtienen son consistentes. Sin embargo, en caso de existir dependencia y si esta es apreciable, los estimadores pierden eficiencia y las predicciones no son las más apropiadas. Al respecto, para modelar la dependencia temporal de una variable restringida en el intervalo $(0, 1)$ sugiere el empleo de cópulas gaussianas. Así también, Guolo y Varin (2014) analiza las series temporales a través de un modelo de regresión beta, el cual permite la interpretación directa de los parámetros de regresión en la escala de la variable respuesta original. La dependencia temporal está modelada por una cópula gaussiana, con una matriz de correlación correspondiente a un proceso estacionario autorregresivo de media móvil ARMA. Los autores señalan que la inferencia y la predicción son directos debido a que este modelo permite separar los efectos de la dependencia temporal de las covariables en la variable respuesta. Kurowicka y Cooke (2006) y Patton (2012), indican que la noción de cópula se introdujo para separar el efecto de la dependencia del efecto de las distribuciones marginales en una distribución conjunta.

Las cópulas se han utilizado en el análisis univariado de series de tiempo continuo para caracterizar la dependencia en una secuencia de observaciones extendiéndose a los análisis multivariados (Joe, 1997). Loaiza Maya et al. (2017), indican que los modelos basados en

cópulas proporcionan una gran flexibilidad en el modelado de distribuciones multivariadas, lo que permite especificar los modelos para las distribuciones marginales por separado de la estructura de dependencia (cópula) que las vincula para formar una distribución conjunta. Además de la flexibilidad, esto a menudo también facilita la estimación debido a la simplicidad con la que incorporan estimaciones complejas.

Cabe precisar que, la aplicación de cópulas para modelar la dependencia temporal se puede realizar para variables con distribuciones continuas, así como para distribuciones discretas. De acuerdo a los trabajos realizados existe una amplia aplicación para capturar la dependencia en las variables financieras. Al respecto, Loaiza Maya et al. (2017) mostraron que los modelos de cópula resultantes pueden capturar sus distribuciones marginales con mayor precisión que los modelos de heteroscedasticidad condicional autorregresiva generalizada univariada y multivariada y que estos producen pronósticos más precisos.

1.4. Organización del trabajo

En el Capítulo 2 se presenta una revisión de la literatura estadística, donde se incluyen los principales conceptos de series de tiempo y cópulas gaussianas. En el Capítulo 3 se explica en detalle el modelo de regresión beta con cópula gaussianas y dependencia temporal propuesto por Masarotto y Varin (2012) y Guolo y Varin (2014), su inferencia, método de estimación y análisis de residuos. En el Capítulo 4 se describen los resultados obtenidos de la simulación del modelo presentado sobre diferentes escenarios. En el Capítulo 5 se describen los resultados de la aplicación a la tasa de desempleo en el Perú y finalmente, el Capítulo 6 presenta las conclusiones de la aplicación del modelo abordado.

Capítulo 2

Marco Teórico

2.1. Series de tiempo

Una serie tiempo $\{Y_t, \forall\ t \in \mathbb{Z}\}$ es una secuencia de variables aleatorias medidas en determinados momentos del tiempo. Sin embargo, usualmente se observa que la serie en un periodo específico en el tiempo $\mathbf{Y} = (Y_1, \ldots, Y_T)$, la cual se encuentra indexada según el orden en que se obtienen. Este tipo de serie presenta dependencia entre sí, donde un valor de un punto específico de la serie depende de sus valores pasados (Shumway y Stoffer, 2011).

Cabe precisar que, cuando se observa una serie de tiempo, generalmente se asume que es una realización de un proceso estocástico (estacionario o no estacionario) y resulta importante conocer la dependencia de la variable respecto de sus valores pasados.

2.1.1. Principales definiciones

A continuación se presentan definiciones necesarias para proponer un modelo de serie temporal:

Definición 1. *(Ruido blanco). Es una secuencia de variables aleatorias no correlacionadas* $\{w_t, t \in \mathbb{Z}\}$ *con media cero y varianza constante.*

Es decir se cumple:

i) $\mathrm{E}(w_t) = 0, \quad \forall\ t;$

ii) $\mathrm{Var}(w_t) = \sigma^2, \quad \forall\ t;$

iii) $\mathrm{Cov}(w_t, w_s) = \mathrm{E}(w_t w_s) = 0, \ \ \forall\ t \neq s.$

Si w_t es un ruido blanco normal, entonces se cumple que: $w_t \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$.

Definición 2. *(Estacionariedad estricta). Una serie de tiempo* $\{Y_t, \forall\ t \in \mathbb{Z}\}$ *es estrictamente estacionaria cuando la distribución del conjunto de variables* $\{Y_{t_1}, Y_{t_2}, \ldots, Y_{t_k}\}$ *tiene la misma distribución que* $\{Y_{t_1+h}, Y_{t_2+h}, \ldots, Y_{t_k+h}\}$, $\forall\ h \in \mathbb{Z}$ *, es decir,*

$$P(Y_{t_1} \leq c_1, \ldots, Y_{t_k} \leq c_k) = P(Y_{t_1+h} \leq c_1, \ldots, Y_{t_k+h} \leq c_k),$$

para todo $k = 1, 2, \ldots, T$ *y todo* $h = 0, \pm 1, \pm 2, \ldots$.

Definición 3. *(Estacionariedad débil). Una serie de tiempo* $\{Y_t, \forall\ t \in \mathbb{Z}\}$ *con estacionaredad débil, es un proceso de varianza finita tal que*

i) La media, $E(Y_t) = \mu_t = \mu$, es constante y no depende del tiempo t.

ii) La varianza, $Var(Y_t) = \sigma^2$, es constante,

iii) La función de autocovarianzas, $\gamma(t, t+h)$, depende solo de h , es decir

$$Cov(Y_t, Y_{t+h}) = \gamma_h, \ \forall \ h \in \mathbb{Z}.$$

El concepto de estacionariedad débil forma la base de gran parte del análisis realizado con series de tiempo. Para fines de esta tesis, el término **estacionariedad** hará referencia a la estacionariedad débil. Si un proceso es estacionario en el sentido estricto, usaremos el término estrictamente estacionario.

Definición 4. *(Función de autocovarianza). Para un proceso estacionario $\{Y_t\}$ con $Var(Y_t) < \infty$, la función de autocovarianza es*

$$Cov(Y_t, Y_{t+k}) = Cov(Y_{t+j}, Y_{t+j+k}) = \gamma_k, \forall \ t.$$

Definición 5. *(Función de autocorrelación). Para un proceso estacionario $\{Y_t\}$, la función de autocorrelación es el conjunto de los valores de los coeficientes de autocorrelación ρ_k, donde $corr(Y_t, Y_{t+k}) = corr(Y_{t+j}, Y_{t+j+k}) = \rho_k, \forall \ t, j, k$*

Esto significa que la función de autocorrelación entre dos momentos de la variable que distan k periodos de tiempo, es la misma para otros momentos que disten el mismo tiempo k de periodos. El coeficiente de correlación de orden k es:

$$\rho_k = \frac{\gamma_k}{\gamma_o}, \forall \ k \in \mathbb{Z}.$$

Definición 6. *(Proceso lineal general). Se dice que una serie de tiempo $\{Y_t\}$ es un proceso lineal si este puede ser expresado en la forma*

$$Y_t = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i},$$

donde ϵ_t es un ruido blanco, ψ_i es una secuencia dada de constantes y se cumple que $\sum_{i=0}^{\infty} \psi_i^2 < \infty$.

Definición 7. *(Operador de rezago). Se denota por L y se define por $L(Y_t) = Y_{t-1}$, donde el operador de rezago L opera sobre una serie de tiempo, rezagándola un periodo previo. De igual forma $L(Y_{t-1}) = Y_{t-2}$, entonces $L(L(Y_t)) = L^2(Y_t) = Y_{t-2}$ y en general $L^p(Y_t) = Y_{t-p}$. Se define $L^0 = I$, como el operador identidad.*

Un polinomio de grado p en el operador L se define como el operador formado por una combinación lineal de potencias de L:

$$B_p(L) = \beta_0 + \beta_1 L + \beta_2 L^2 + \cdots + \beta_p L^p,$$

tal que

$$\begin{aligned} B_p(L)(Y_t) &= \left(\beta_0 + \beta_1 L + \beta_2 L^2 + \cdots + \beta_p L^p\right) Y_t, \\ &= \sum_{j=0}^{p} \beta_j L^j Y_t \\ &= \sum_{j=0}^{p} \beta_j Y_{t-j}, \\ &= \beta_0 Y_t + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p}. \end{aligned}$$

Las definiciones antes presentadas son necesarias para la identificación y estimación de modelos de series temporales correspondientes a procesos estacionarios. Para fines de esta tesis, los procesos estacionarios que se presentarán son los procesos autorregresivos AR (p) de medias móviles, MA (q) y autorregresivos de medias móviles ARMA (p, q).

2.1.2. Procesos Autoregresivos de Orden p, AR(p)

Definición 8. *(Proceso autoregresivo). Un proceso $\{Y_t, \forall\ t \in \mathbb{Z}\}$ sigue un proceso autorregresivo, AR(p), de orden p cuando:*

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + w_t, \tag{2.1}$$

donde Y_t es estacionario, $\phi_1, \phi_2, \ldots, \phi_p$ son constantes $(\phi_p \neq 0)$ y w_t es ruido blanco normal con media cero y varianza σ^2.

Así, el modelo AR(1) es definido por:

$$Y_t = c + \phi_1 Y_{t-1} + w_t,$$

donde

i) $E(Y_t) = \dfrac{c}{1-\phi_1}$

ii) $\text{Var}(Y_t) = \gamma_0 = \dfrac{\sigma^2}{1-\phi^2}$

iii) $\text{Cov}(Y_t, Y_{t-1}) = E(Y_t, Y_{t-1}) = \gamma_1 = \phi_1 \gamma_0$

iv) El coeficiente de correlación es

$$\rho_1 = \frac{\gamma_1}{\gamma_0} = \phi_1$$

Generalizando el modelo AR(p) definido en (2.1) se tienen las siguientes propiedades:

i) Si la media de Y_t, μ, es decir, es diferente de cero, es reemplazado Y_t es reemlazado por $Y_t - \mu$ en (2.1),

$$Y_t - \mu = \phi_1 (Y_{t-1} - \mu) + \phi_2 (Y_{t-2} - \mu) + \cdots + \phi_p (Y_{t-p} - \mu) + w_t,$$

o escribir

$$Y_t = \alpha + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + w_t,$$

donde $\alpha = \mu\,(1 - \phi_1 . \cdots - \phi_p)$.

ii) La autocovarianza de orden k es:

$$\gamma_k = \phi_1 \gamma_{k-1} + \ldots \phi_p \gamma_{k-p}.$$

iii) La autocorrelación de orden k es:

$$\rho_k = \phi_1 \rho_{k-1} \phi_2 \rho_{k-2} + \cdots + \phi_p \rho_{k-p}, \tag{2.2}$$

para $k = 1, 2, \ldots, T$. Luego se satisface el sistema de p ecuaciones:

$$\begin{array}{lllllllllll} \rho_1 & = & \phi_1 & + & \phi_2\rho_1 & + & \phi_3\rho_2 & + & \cdots & \phi_{p-1}\rho_{p-2} & + & \phi_p\rho_{p-1} \\ \rho_2 & = & \phi_1\rho_1 & + & \phi_2 & + & \phi_3\rho_1 & + & \cdots & \phi_{p-1}\rho_{p-3} & + & \phi_p\rho_{p-2} \\ \vdots & = & \vdots & & \vdots & & \vdots & & \vdots & \ddots & & \vdots \\ \rho_p & = & \phi_1\rho_{p-1} & + & \phi_2\rho_{p-2} & + & \phi_3\rho_{p-3} & + & \cdots & \phi_{p-1}\rho_1 & + & \phi_p. \end{array} \tag{2.3}$$

Por propiedad $\gamma_{_k} = \gamma_{_{-k}}$, luego se tiene que $\rho_{_k} = \rho_{_{-k}}$. Así el sistema de ecuaciones (2.3) se puede solucionar bajo la metodología Yule-Walker:

$$\begin{bmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \\ \vdots \\ \hat{\phi}_p \end{bmatrix} = \begin{bmatrix} 1 & \hat{\rho}_1 & \hat{\rho}_2 & \cdots & \hat{\rho}_{p-1} \\ \hat{\rho}_1 & 1 & \hat{\rho}_1 & \cdots & \hat{\rho}_{p-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{\rho}_{p-1} & \hat{\rho}_{p-2} & \hat{\rho}_{p-3} & \cdots & \hat{\rho}_1 \end{bmatrix} \begin{bmatrix} \hat{\rho}_{_1} \\ \hat{\rho}_{_2} \\ \vdots \\ \hat{\rho}_{_p} \end{bmatrix}$$

La solución permitirá obtener los coeficientes ϕ y de esa manera obtener la autocorrelación AR(p) propuesta en (2.2).

2.1.3. Procesos de Medias Móviles de orden q, MA(q)

Definición 9. *(Proceso de medias móviles). Se dice que una serie $\{Y_t, \forall\ t \in \mathbb{Z}\}$ sigue un proceso de media móvil de orden q si se cumple que:*

$$Y_t = \epsilon_t + \lambda_1 \epsilon_{t-1} + \cdots + \lambda_q \epsilon_{t-q},\ \ \forall\ t \in \mathbb{Z}, \tag{2.4}$$

donde ϵ_t, sigue un proceso ruido blanco. La expresión con el operador L según la definición 7 define el polinomio:

$$\lambda_q\,(L) = 1 + \lambda_1 L + \cdots + \lambda_q L^q.$$

Así, la ecuación (2.4) se expresa como

$$Y_t = \lambda_q\,(L)\,(\epsilon_t)\,.$$

El modelo MA(1) viene definido por:

$$Y_t = \mu + \epsilon_t - \lambda_1 \epsilon_{t-1},$$

donde

i) $\mathrm{E}(Y_t) = \mu$,

ii) $\mathrm{Var}(Y_t) = \gamma_0 = \left(1 + \lambda_1^2\right)\sigma^2$,

iii) La autocovarianza es:

$$\gamma_1 = -\lambda_1\sigma^2$$

$$\gamma_k = 0, \;\; k > 1,$$

iv) La autocorrelación es:

$$\rho_k = \begin{cases} -\frac{\lambda_1}{1+\lambda_1^2} & k = 1 \\ 0 & k > 1. \end{cases}$$

Generalizando el modelo MA(q) se tienen las siguiente propiedades:
i)Esperanza

$$\mathrm{E}(Y_t) = \mu.$$

ii) Varianza

$$\mathrm{Var}(Y_t) = \gamma_0 = \left(1 + \lambda_1^2 + \cdots + \lambda_q^2\right)\sigma^2.$$

iii) Covarianza

$$\gamma_k = \mathrm{Cov}(Y_t,\, Y_{t-k}) = \left(\lambda_k + \lambda_{k+1}\lambda_1 + \cdots + \lambda_q\lambda_{q-k}\right)\sigma^2.$$

iv) Funcion de autocorrelación

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{\left(\lambda_k + \lambda_{k+1}\lambda_1 + \cdots + \lambda_q\lambda_{q-k}\right)}{\sum_{i_1}^{q}\lambda_i^2}.$$

2.1.4. Procesos Autorregresivos de Medias Móviles, ARMA(p, q)

Definición 10. *(Proceso autorregresivo de medias móviles). Un modelo autorregresivo de medias móviles de orden p, q engloba los conceptos de los modelos AR y MA y es de la forma:*

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \epsilon_t + \lambda_1\epsilon_{t-1} + \cdots + \lambda_q\epsilon_{t-q},$$

donde $\phi_q \neq 0$, $\lambda_q \neq 0$, $\sigma_\epsilon^2 > 0$ y ϵ_t sigue un proceso ruido blanco.

El modelo en términos del operador de retardos queda de la siguiente manera:

$$\left(1-\phi_1 L-\cdots-\phi_p L^p\right) Y_t = \left(1-\lambda_1 L-\cdots \lambda_q L^q\right) \epsilon_t$$
$$\phi_p\left(L\right) Y_t = \lambda_q\left(L\right) \epsilon_t,$$

donde $\phi_p(L)$ es el polinomio autoregresivo y $\lambda_q(L)$ es el polinomio de medias móviles.

Esto implica que:

$$Y_t = \phi(L)\epsilon_t,$$

donde $\phi(L) = \frac{\lambda_q(L)}{\phi_p(L)}$.

El proceso ARMA es estacionario si $\sum_{j=1}^{\infty} |\phi_j| < \infty$ y $E[Y_t] = \phi(L)E[\epsilon_t] = 0$.

El proceso ARMA es invertible si puede escribirse como un proceso AR(∞). Es decir si existe:

$$\pi(L) = \pi_0 + \pi_1 L + \pi_2 L^2 + ...,$$

con $\sum_{j=1}^{\infty} |\pi_j| < \infty$ y $\epsilon_t = \pi(L)X_t$.

El proceso ARMA es causal si puede escribirse como un proceso MA(∞). Es decir si existe

$$\psi(L) = \psi_0 + \psi_1 L + \psi_2 L^2 + ...,$$

con $\sum_{j=1}^{\infty} |\psi_j| < \infty$ y $Y_t = \psi(L)\epsilon_t$.

El modelo (p, q) presenta las características de los modelos AR(p) y MA(q) debido a que contiene ambas estructuras a la vez.

Para el caso ARMA$(1, 1)$, este se define por:

$$Y_t = c + \phi_1 Y_{t-1} + \epsilon_t - \lambda_1 \epsilon_{t-1},$$

el cual es estacionario si se cumple que $|\phi| < 1$. Así también, presenta las siguiente propiedades:

i) $E[Y_t] = \dfrac{c}{1-\phi} = \mu$,

ii) $\gamma_0 = \mathrm{Var}(Y_t) = \dfrac{\sigma^2(\lambda^2 - 2\phi\lambda + 1)}{a - \phi^2}$,

iii)La función de autocovarianza es $\gamma_1 = Cov(Y_t, Y_{t-1}) = \sigma\dfrac{(\phi-\lambda)(1-\phi\lambda)}{1-\phi^2}$
y $\gamma_k = \phi\gamma_{k-1}, \forall\ k > 1$,

iv) La función de autocorrelación es:
$\rho_1 = Corr\left[Y_t, Y_{t-1}\right] = \dfrac{\left(\phi-\lambda\right)\left(1-\phi\lambda\right)}{1+\lambda^2-2\lambda\phi}$ y $\rho_k = Corr\left[Y_t, Y_{t-k}\right] = \phi\rho_{k-1}$, para $k > 1$.

La estimación de procesos ARMA(p, q) se basa por lo usual en el supuesto que el vector $Y = (Y_1, \ldots, Y_n)'$ tiene distribución normal multivariada con media μ, y matriz de covarianzas $\Sigma = \left[Cov\left(Y_i, Y_j\right)\right]_{n\times n}$. Como un proceso estacionario cumple que $Cov\left(Y_i, Y_j\right) = \gamma\left(j-i\right)$, donde $\gamma(k)$ es la función de autocovarianzas de Y_t, la forma de Σ es la de una matriz tipo Toeplitz:

$$\Sigma = \begin{bmatrix} R(0) & R(1) & \cdots & R(n-1) \\ R(1) & R(0) & \cdots & R(n-2) \\ R(2) & R(1) & \cdots & R(n-3) \\ \vdots & \vdots & \ddots & \vdots \\ R(n-1) & R(n-2) & \cdots & R(0) \end{bmatrix}.$$

Por ejemplo, para un AR(p), $R(k)$ se calcula mediante las ecuaciones Yule Walker, $R(k) = \mu + \sum_{j=1}^{p} \phi_j R(k-j)$. Por tanto, definiendo $\beta = \left(\mu, \sigma^2, \phi, \ldots, \phi_p, \lambda_1, \ldots, \lambda_1, \ldots, \lambda_q\right)'$, la matriz Σ depende del vector de parámetros β, y se escribe $\Sigma(\beta)$. Este supuesto permite implementar la estimación por máxima verosimilitud mediante la densidad Normal Multivariada:

$$f(y, \beta) = \frac{1}{(2\pi)^{n/2}\sqrt{|(\Sigma(\beta)|)}} \exp\left(-\frac{1}{2}(y-\mu)\Sigma(\beta)^{-1}(y-\mu)'\right), \tag{2.5}$$

donde $\mu = (\mu, \ldots, \mu)' \in \mathbb{R}^n$. La función de log-verosimilitud se define a partir del logaritmo de la densidad (2.5).

2.2. Regresión beta para el análisis de series de tiempo

2.2.1. Distribución beta

Una variable aleatoria Y sigue una distribución beta si su función de densidad está dada por,

$$f_Y(y; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, \quad 0 < y < 1, \tag{2.6}$$

donde $\alpha > 0$, $\beta > 0$.

Para esta distribución la media y varianza están expresadas por:

$$\mathrm{E}(Y) = \frac{\alpha}{\alpha+\beta},$$

$$\mathrm{Var}(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

2.2.2. Regresión beta

La distribución beta se caracteriza por la flexibilidad para modelar proporciones, ya que su densidad puede adoptar diferentes formas en función a los valores de los parámetros. En este contexto, se han desarrollado modelos de regresión asociados a la distribución mencionada.

Ferrari y Cribari-Neto (2004) proponen un modelo de regresión beta, donde considera la utilidad de modelar la media. Para esto emplean una reparametrización correspondiente a la distribución definida en (2.6). Así, considerando la media $\mu = \frac{\alpha}{\alpha+\beta}$ y un parámetro de precisión $\phi = \alpha + \beta$. La función de densidad reparametrizada de Y es dada por:

$$f_Y(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0 < y < 1, \tag{2.7}$$

donde

$$\mathrm{E}(y) = \mu,$$

y

$$\mathrm{Var}(y) = \frac{V(\mu)}{1+\phi},$$

y además, $0 < \mu < 1$ y $\phi > 0$.

En este contexto, si se tiene que $Y_1, Y_2, ..., Y_n$, es un conjunto de variables aleatorias independientes con una distribución beta reparametrizada definida en (2.7), entonces el modelo de regresión beta propuesto por los autores mencionados es

$$g(\mu_i) = \sum_{j=1}^{k} x_{ij}\beta_j = \eta_i, \tag{2.8}$$

donde $\beta = (\beta_1, ..., \beta_k)^\top$ es el vector de parámetros de regresión, $\mathbf{x_{ij}} = (x_{i1}, ..., x_{ik})^\top$ es el vector de k covariables, η_t es predictor lineal y $g(\cdot)$ es una función de enlace estrictamente monotóna y dos veces diferenciable. Las funciones de enlace pueden ser logística, logaritmo, clog-log, probit, entre otras.

2.2.3. Inferencia en el modelo de regresión beta

Kieschnick y McCullough (2003) sugieren que se utilice un modelo de regresión paramétrica basado en la distribución beta para analizar variables restringidas al intervalo $(0, 1)$. Ferrari y Cribari-Neto (2004) sostienen que emplear un modelo de regresión lineal para proporciones no es el más apropiado ya que puede generar que las predicciones de la variable de interés exceda sus límites superior e inferior. En ese contexto, la estimación de los parámetros de una regresión beta es definida mediante máxima verosimilitud. Estos parámetros son interpretables en términos de la media de la respuesta a diferencia de los parámetros de una regresión lineal que emplea una respuesta transformada.

La función de log-verosimilitud para estimar los parámetros de la regresión definida en (2.8) de una muestra de T observaciones con el parámetro de precisión constante es

$$\ell(\beta, \phi) \sum_{i=1}^{T} \ell_i\left(\mu_i, \phi\right), \tag{2.9}$$

donde

$$\ell(\mu_i, \phi) = \log\Gamma(\phi) - \log\Gamma(\mu_i\phi) - \log((1-\mu_i)\phi) + (\mu_i\phi - 1)\log y_i + ((1-\mu_i)\phi - 1)\log(1-y_i).$$

La función de Score corresponde a la derivada de la ecuación (2.9). Con dichas ecuaciones se resuelve el sistema de ecuaciones y se obtienen los parámetros estimados $\hat{\beta}$ y $\hat{\phi}$, donde de acuerdo a la complejidad de las ecuaciones se emplean métodos numéricos.

El modelo permite realizar pruebas de diagnóstico como la bondad de ajuste, análisis de valores influyentes. Así también, el contraste de hipótesis para el análisis de los parámetros y el uso de los criterios de AIC y BIC para la elección de modelos.

2.2.4. Extensiones para el análisis de series de tiempo

El modelo de regresión beta descrito en la sección 2.2.2 es una referencia para analizar la dependencia temporal. Los modelos dinámicos trabajados por Rocha y Cribari-Neto (2009) y Bayer et al. (2018), a diferencia del modelo definido en (2.8), consideran la inclusión de un componente ARMA y un componente estacional de manera aditiva.

Para esto consideran la distribución beta definida en (2.7) con la diferencia en que asumen que el valor de la variable Y_t depende de información pasada. Así, la función de densidad de un conjunto de variables aleatorias Y_t es dada por

$$Y_t \sim beta(\mu_t, \phi),\ t = 1, 2, \ldots, T.$$

Por ejemplo, el modelo de regresión está dado por:

$$g(\mu_t) = \eta_t = x_t^\top \beta + \tau_t, \tag{2.10}$$

donde τ_t es el componente sistemático ARMA(p, q).

Cabe precisar que el método de estimación para los modelos propuestos ha sido el de máxima verosimilitud expuesta en la sección anterior.

Otra forma de modelar dependencia temporal es incorporando la estimación de los parámetros asociados al proceso ARMA mediante cópulas. En la siguiente sección se aborda los conceptos y aplicaciones de este tipo de modelos.

2.3. Cópulas gausianas para el análisis de dependencia y serie de tiempo

La dependencia entre dos o más variables puede surgir de varias formas como, por ejemplo, mediciones repetidas sobre la misma unidad u observaciones recopiladas secuencialmente en el tiempo o en el espacio.

Existen varias formas de analizar la dependencia entre variables. Una de estas es mediante una cópula, la cual permite incorporar la dependencia entre las variables aleatorias a partir de sus distribuciones acumuladas. De esta manera, es posible separar los efectos de la dependencia y el de las distribuciones marginales en una distribución conjunta (Kurowicka y Cooke, 2006).

Así también, las cópulas se han utilizado en el análisis univariado de series de tiempo continuo para caracterizar la dependencia en una secuencia de observaciones Joe (1997). La ventaja de emplear este enfoque es que se logra especificar por separado los efectos de las distribuciones marginales de las variables aleatorias y los efectos de la dependencia temporal. Cabe precisar que las cópulas pueden usarse para construir modelos multivariados para variables de respuesta dependientes de cualquier tipo; continuas, discretas o mixtas.

2.3.1. Cópulas

Definición 11. *(Cópula). Una cópula p-dimensional es una función de distribución multivariada* $C : [0, 1]^n \to [0, 1]$*, cuyas distribuciones marginales se definen por* $U_i \sim U(0, 1), i = 1, 2, ..., n$.

Según Nelsen (2006) y Joe (1997), esta función tiene las siguientes propiedades:

i. $C(u_1, ..., u_n)$ *es creciente para cada componente* u_i

ii. $C(1..,1,u_i,1,...,1) = u_i,\ i = 1,...,n$

iii. Para todo $(u_{11},...,u_{n1}),(u_{12},...,u_{n2}) \in [0,1]^n$ *con* $u_{i1} \leq u_{i2}\, \forall i = 1,\ldots,n$, *se tiene que:*

$$\sum_{i1=1}^{2} ... \sum_{in=1}^{2} (-1)^{i_1+...+i_n} C(u_{1i_1},...,u_{ni_n}) \geq 0$$

De las propiedades, con la segunda se establece el requisito para tener distribuciones marginales uniformes, las otras dos son condiciones para cualquier distribución. Según la definición, una cópula es una distribución conjunta de variables aleatorias $U_1 \ldots, U_n$, donde cada una es marginalmente distribuida con una distribución uniforme $U(0,1)$. En particular, en la definición citada, el término cópula es usado para la función de distribución acumulada de dicha función:

$$C(u_1,\ldots,u_n) = P(U_1 \leq u_1,\ldots,U_n \leq u_n).$$

Para comprender mejor los alcances de las cópulas en el análisis de dependencia se debe tener en cuenta el teorema de Sklar, el cual indica que toda función de distribución multivariada de variables aleatorias tiene una cópula asociada.

Teorema 1. *(Teorema de Sklar). Para un conjunto de variables aleatorias* $X_1, X_2, ..., X_n$ *con función de distribución acumulada conjunta*

$$F(x_1,\ldots,x_n) = P(X_1 \leq x_1,\ldots,X_n \leq x_n),$$

y función de distribución acumulada marginal:

$$F_j(x) = P(X_j \leq x), j = 1,2,\ldots,n$$

existe una cópula $C : [0,1]^n \to [0,1]$ *tal que*

$$F(x_1,\ldots,x_n) = C(F_1(x_1),\ldots,F_p(x_n)). \tag{2.11}$$

Si cada distribución marginal $F_j(x)$ *es continua, entonces* C *es única.*

Así, si C es una cópula y $F_1(x_1),\ldots,F_n(x_n)$ son funciones de distribución acumuladas univariadas, entonces $F(x_1,\ldots,x_n)$ definida previamente es una función de distribución conjunta con marginales $F_j(x)$.

Como se puede ver del Teorema de Sklar, además de indicar que una cópula es una función de distribución conjunta, dicha distribución a su vez puede escribirse en términos de las distribuciones marginales y una cópula (Becerra y Melo, 2008); es decir, la cópula enlaza las funciones de distribución marginales para formar la distribución conjunta. De esta manera, para el modelamiento, el teorema de Sklar permite separar la distribución marginal de la estructura de dependencia.

Para la demostración del Teorema 1, cuando las funciones de distribución marginales $F_j(x)$ son continuas, estas presentan una función de distribución inversa dada por $F_j^{-1}(x)$ tal que (Joe, 2014):

$$F_j\left(F_j^{-1}(u)\right) = u,\ 0 \leq u \leq 1, \tag{2.12}$$

Así, $U_j = F_j(X_j) \sim U(0,1)$, pues

$$\begin{aligned} P(U_j \leq u) &= P(F_j(X_j) \leq u) \\ &= P(X_j \leq F_j^{-1}(u)) \\ &= F_j[F_j^{-1}(u)] = u. \end{aligned}$$

Luego,

$$\begin{aligned} F(x_1, ..., x_p) &= P(X_1 \leq x_1, ..., X_n \leq x_n) \\ &= P(F_1^{-1}(U_1) \leq x_1, ..., F_n^{-1}(U_n) \leq x_n) \\ &= P(U_1 \leq F_1(x_1), ..., U_n \leq F_n(x_n)) \\ &= C(F_1(x_1), ..., F_p(x_n)). \end{aligned}$$

Por el teorema de Sklar se sabe de la existencia de una cópula para cualquier distribución conjunta que se puede escribir en términos de las inversas de las distribuciones marginales como

$$C(u_1, ..., u_n) = F\left(F_1^{-1}(u_1), ..., F_n^{-1}(u_n)\right),\ 0 \leq u \leq 1 \tag{2.13}$$

donde C es única.

Proposición 1. *Si las funciones $F(\cdot)$ y $C(\cdot)$ son diferenciables, entonces la siguiente ecuación*

$$F(x_1, ..., x_n) = C(F_1(x_1), ..., F_n(x_n)),$$

implica que

$$\frac{f(x_1, \ldots, x_n)}{f_1(x_1), \ldots, f_n(x_n)} = c(F_1(x_1), ..., F_n(x_n)),$$

donde c es la función de densidad de la cópula.

Demostración:

Como

$$\frac{\partial^n F\left(x_1, \ldots, x_n\right)}{\partial x_1, \ldots \partial x_n} = \frac{\partial^n C\left(F_1(x_1), \ldots, F_n(x_n)\right)}{\partial F_1(x_1), \ldots, \partial F_n(x_n)} \times \frac{\partial F_1(x_1)}{\partial x_1} \cdots \frac{\partial F_p(x_n)}{\partial x_n},$$

entonces

$$f\left(x_1, \ldots, x_n\right) = \frac{\partial^n C\left(F_1(x_1), \ldots, F_n(x_n)\right)}{\partial F_1(x_1), \ldots, F_n(x_n)} f_1(x_1), \ldots, f_n(x_n).$$

Así,

$$\frac{f\left(x_1 \ldots, x_n\right)}{f_1(x_1), \ldots, f_p(x_n)} = \frac{\partial^n C\left(F_1(x_1), \ldots, F_n(x_n)\right)}{\partial F_1(x_1), \ldots, F_n(x_n)}$$
$$\frac{f\left(x_1, , \ldots, x_n\right)}{f_1(x_1), \ldots, f_p(x_n)} = c\left(F_1(x_1), \ldots, F_n(x_n)\right).$$

Por este resultado, la función de densidad conjunta puede ser expresada por

$$f(x_1, \ldots, x_n) = c(F_1(x_1) \ldots, F_n(x_n)) \prod_{i=1}^{n} f_i(x_i). \tag{2.14}$$

2.3.2. Cópulas gaussianas

Las cópulas gaussianas proporcionan un marco general flexible para modelar respuestas dependientes de cualquier tipo (Song, 2000; Embrechts, 2009). Sin embargo, la regresión de la cópula gaussiana todavía presenta un uso limitado para las respuestas dependientes no continuas, donde las estimaciones requieren la aproximación de integrales de alta dimensión.

Con base en el teorema de Sklar y a partir de las ecuaciones (2.12) y (2.13), una cópula gaussiana se puede representar de la siguiente manera:

$$C(u_1, ..., u_n) = \Phi_\Sigma(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_n)), \tag{2.15}$$

donde Φ^{-1} es la inversa de la función de distribución normal estándar y Φ_Σ es la función de distribución normal acumulada n-variada con media 0 y matriz de correlación $\mathbf{\Sigma}$. Así, la función de densidad de una cópula gaussiana es:

$$\begin{aligned} c(u_1, u_2, ..., u_n) &= \tfrac{\partial^n}{\partial u_1 \partial u_2 ... \partial u_n} C\left(u_1, u_2, \ldots, u_n\right) \\ &= \tfrac{\partial^n}{\partial u_1 \partial u_2 ... \partial u_n} \Phi_\Sigma(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_n)) \\ &= \tfrac{\phi_n\left(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_n)\right)}{\phi(\Phi^{-1}(u_1)) \cdots \phi(\Phi^{-1}(u_1))} \\ &= \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_n))^\top (\Sigma^{-1} - I)(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_n))\right\}, \end{aligned} \tag{2.16}$$

donde $\phi_n(\cdot)$ representa la fdp de una normal multivariada con media cero y matriz de covarianza Σ, y $\phi(\cdot)$ representa la fdp de una normal univariada estándar.

Como se explicó anteriormente, por el teorema de Sklar, una distribución conjunta puede ser construida por las distribuciones marginales y la estructura de dependencia especificada. Sean $X_1, \ldots, X_n$, v.a.s. con funciones de distribución acumulada F_j, se definen las v.a.s $U_j = F_j(X_j) \sim U(0,1) \quad \forall j = 1 \ldots, n$. Luego por las ecuaciones (2.14) y (2.16), la función de densidad conjunta de $X_1, \ldots, X_n$ usando una cópula gaussiana es dada por:

$$f(x_1, ..., x_n) = \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_n))^\top (\Sigma^{-1} - I)(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_n))\right\} \prod_{i=1}^{n} f_i(x_i). \tag{2.17}$$

En este contexto, los parámetros obtenidos de la cópula reflejan el grado de dependencia de las variables. Así por ejemplo, la dependencia en una cópula gausiana se resumen en la matriz de correlación.

Cabe precisar que, para la estimación de modelos con cópulas gausianas la expresión (2.17) es la función de densidad que se emplea para determinar la función de verosimilitud que permite estimar los parámetros (Parsa y Klugman, 2010).

Capítulo 3

Estructura del modelo

El modelo de regresión beta propuesto por Ferrari y Cribari-Neto (2004) es útil cuando la variable de interés es continua y restringida en el intervalo $(0, 1)$. En este trabajo se detalla este modelo de regresión beta para series de tiempo usando cópulas gaussianas como fue propuesto por Masarotto y Varin (2012).

3.1. Regresión beta

Sea $\mathbf{Y} = (Y_1, \ldots, Y_T)^\top$ un vector de variables aleatorias donde las Y_t definen la realización de una serie de tiempo. Se asume que la distribución marginal de Y_t es una distribución beta, es decir, $Y_t \sim \text{beta}(\mu_t, \kappa_t), \forall\ t = 1, \ldots, T$, donde $E(Y_t) = \mu_t$ y κ_t es un parámetro de precisión tal que, $Var(Y_t) = \dfrac{\mu_t(1-\mu_t)}{1+\kappa_t}$. Luego, la función de densidad marginal de Y_t es dada por:

$$f_{Y_t}(y_t) = \frac{\Gamma(\kappa_t)}{\Gamma(\mu_t\kappa_t)\,\Gamma((1-\mu_t)\,\kappa_t)} y_t^{\mu_t\kappa_t - 1}(1-y_t)^{(1-\mu_t)\kappa_t - 1}, \tag{3.1}$$

donde $0 < \mu_t < 1$ y el parámetro de precisión $\kappa_t > 0$.

Se define el modelo de regresión beta a través de funciones de enlace $g_1(\cdot)$ y $g_2(\cdot)$ que asocian la media μ_t y precisión κ_t a sus respectivos predictores lineales. En particular, se asume:

$$g_1(\mu_t) = \text{logit}(\mu_t) = x_t^\top \beta_x,$$

$$g_2(\kappa_t) = \log(\kappa_t) = z_t^\top \beta_z,$$

donde x_t es un vector de covariables de dimensión h, z_t es un vector de covariables de dimensión r y β_x así como β_z son vectores de coeficientes de regresión.

3.2. Regresión beta usando cópulas gaussianas

Usualmente los modelos estadísticos asumen que un vector aleatorio $\mathbf{Y} = (Y_1, \ldots, Y_T)^\top$ sigue una distribución multivariada, tal que su función de densidad conjunta define implícitamente la dependencia, temporal por ejemplo, entre las variables aleatorias correspondientes y como consecuencia se puede proponer un modelo estadístico apropiado a partir de dicha función de densidad conjunta.

En esta tesis se detalla el modelo propuesto por Masarotto y Varin (2012) y Guolo y Varin (2014), que plantean una forma distinta de modelar dependencias temporales. La base para

el modelo planteado, se centra en las cópulas. Por teoría básica de probabilidades sabemos que a partir de una función de densidad conjunta podemos calcular las funciones de densidad marginales; sin embargo, lo opuesto solo ocurre bajo ciertas restricciones. Las cópulas permiten básicamente definir la distribución marginal de cada variable aleatoria y a partir de ella construir la función de distribución acumulada conjunta del vector aleatorio, de la cual es fácil obtener la función de densidad conjunta.

Especificamente, si se define la distribución marginal de cada Y_t; $t = 1, \ldots, T$, y a partir de las funciones de distribución acumuladas marginales, se puede usar la cópula para construir la función de distribución acumulada del vector aleatorio $\mathbf{Y} = (Y_1, \ldots, Y_T)^\top$, y a partir de este la función de densidad conjunta de $\mathbf{Y}$. Cabe resaltar que de existir dependencias temporales, espaciales o de otra índole, se pueden incorporar en la cópula, que es la que básicamente "enlaza" las variables aleatorias según dicha dependencia.

Bajo este contexto, el modelo presentado por Guolo y Varin (2014), asume que cada variable aleatoria $Y_t \sim \text{beta}\,(\mu_t,\, \kappa_t)\,; \forall t = 1, \ldots, T$, tal que $F_{Y_t}\,(y_t) = F_t(y_t), \forall\ t = 1, \ldots, T$ es la función de distribución acumulada de cada $Y_t \sim \text{beta}\,(\mu_t,\ \kappa_t)$, donde se puede definir $g_1(\mu_t) = \text{logit}\,(\mu_t) = x_t^\top \beta_x$ y $\log\,(\kappa_t) = z_t^\top \beta_z$. Para construir la función de distribución acumulada de $\mathbf{Y} = (Y_1, \ldots, Y_T)^\top$ se hace uso de la definición de cópula, dado que toda función de distribución acumulada conjunta puede ser definida por una cópula.

Formalmente, por el Teorema de Sklar (2.11) la función de distribución acumulada conjunta de $\mathbf{Y}$ puede ser definida por una cópula tal que,

$$F_{_Y}(y_1, \ldots, y_{_T}) = C\,(F_1(y_1), \ldots, F_{_T}(y_{_T}))\,.$$

Si asumimos que

$$u_t = F_t(y_t),\ \forall\ t = 1, \ldots, T,$$

entonces la cópula es una función de $u_1, \ldots, u_{_T}$, tal que

$$F_{_Y}(y_1, \ldots, y_{_T}) = C(u_1, \ldots, u_{_T}).$$

En particular, se puede asumir que C es una cópula gaussiana como la que se definió en la ecuación (2.15), luego se tiene que:

$$F_{_Y}(y_1, \ldots, y_{_T}) = \Phi_\Sigma(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_{_T})), \tag{3.2}$$

donde Φ es una función de distribución acumulada univariada normal estándar y Φ_Σ representa la función de distribución acumulada n-variada normal estándar con matriz de correlación Σ. En general esta matriz de correlación tomará en cuenta la autocorrelación entre las variables aleatorias Y_t como se define más adelante. Para ello se definen errores aleatorios

$$\varepsilon_{_t} = \Phi^{-1}\,(u_t)\,,\ \forall t = 1, \ldots, T$$

y $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_{_T})^\top$. Entonces por la ecuación (3.2) la función de distribución acumulada

conjunta de $\mathbf{Y}$ puede ser redefinida por:

$$F_Y(y_1, \ldots, y_T) = \Phi_\Sigma(\varepsilon_1, \ldots, \varepsilon_T) = \Phi_\Sigma(\varepsilon), \tag{3.3}$$

donde $\varepsilon \sim \mathbf{N}(\mathbf{0}, \mathbf{\Sigma})$. Y para terminar de definir la función de distribución acumulada conjunta de $\mathbf{Y}$ se debe definir la estructura de la matriz $\mathbf{\Sigma}$.

Luego, en particular para el modelo en estudio, la matriz de correlación entre los errores aleatorios Σ toma en cuenta la autocorrelación temporal entre las variables aleatorias Y_t. Para ello, se asume que los errores aleatorios ε_t siguen una distribución normal estándar, y es definido por:

$$\varepsilon_t = \sum_{i=1}^{p} \phi_i \varepsilon_{t-i} + \sum_{j=1}^{q} \lambda_j \eta_{t-j} + \eta_t, \ t = 1, \ldots, T, \tag{3.4}$$

donde ϕ_i representa el componente del proceso autorregresivo AR(p), λ_j al de medias móviles MA(q) y $\eta_t \overset{\text{iid}}{\sim} \text{N}(0, \text{Var}(\eta_t))$, con $\text{Var}(\eta_t) = 1$. Luego

$$\varepsilon \sim N(0, \mathbf{\Sigma})$$

y Σ es la matriz de correlación que está definida por un proceso ARMA (p, q) definido en la ecuación (2.5).

De esta definición, por ejemplo, un caso particular de la ecuación (3.4) es el modelo AR(1) para los errores aleatorios, en donde se tiene que,

$$\varepsilon_t = \phi_1 \varepsilon_{t-1} + \eta_t,$$

donde $\text{E}(\eta_t) = 0$ y $\text{Var}(\eta_t) = 1 - \phi_1^2$, con un coeficiente de autocorrelación $\mid \phi_1 \mid < 1$. Luego $\mathbf{\Sigma}$ para un proceso AR(1) está dada por:

$$\Sigma = \begin{pmatrix} 1 & \phi & \phi^2 & \ldots & \phi^{T-1} \\ \phi & 1 & \phi & \ldots & \phi^{T-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{T-1} & \phi^{T-2} & \phi^{T-3} & \ldots & 1 \end{pmatrix}.$$

Como la función de distribución acumulada conjunta de $\mathbf{Y}$ es construida a partir de las funciones de densidad marginales de $Y_t \sim \text{beta}(\mu_t, \kappa_t)$ y de la cópula gaussiana, implícitamente la función de distribución acumulada conjunta también depende de los coeficientes de regresión β_x y β_z.

Es importante notar que la tendencia de la serie temporal es ajustada a través de la media μ_t definida en la distribución beta, y por ello no es necesario desestacionalizar la serie de tiempo, esta una de las ventajas agregadas del modelo planteado.

Finalmente, a partir de la definición de la distribución acumulada conjunta en la ecuación (3.3), se puede obtener la función de densidad conjunta de $\mathbf{Y}$, a partir de las funciones de densidad condicionales de ε_t y de la función de densidad de la cópula, como se muestra en la siguiente sección.

3.3. Inferencia del modelo

Los parámetros a estimar son $\theta = (\beta, \phi, \lambda)$ donde $\beta = \left(\beta_x^\top, \beta_z^\top\right)^\top$, $\phi = (\phi_1, \ldots, \phi_p)$ y $\lambda = (\lambda_1, \ldots, \lambda_q)$. La función de verosimilitud del modelo está dada por

$$\begin{aligned} L(\theta; y) &= f_{_Y}(y_1, \ldots, y_{_T}) \\ &= f_{_{Y_1}}(y_1) f_{_{Y_2}}(y_2 \mid y_1) \times \cdots \times f_{_{Y_T}}(y_{_T} \mid y_{_{T-1}}, \ldots, y_1). \end{aligned} \tag{3.5}$$

Considerando a $f_{_N}(\varepsilon_t)$ **y** $f_{_N}(\varepsilon_t \mid \cdot)$ comos las funciones de densidad marginal y condicional de los errores con distribución normal, respectivamente y considerando el jacobiano de la transformación de $\varepsilon_t = \Phi^{-1}(F_t(y_t))$, se tiene que:

$$f_{_{Y_t}}(y_t \mid y_{t-1}, \ldots, y_1) = f_{_N}(\varepsilon_t \mid \varepsilon_{t-1}, \ldots, \varepsilon_1) \left| \frac{d\varepsilon_t}{dy_t} \right|,$$

donde

$$\begin{aligned} \left| \frac{d\varepsilon_t}{dy_t} \right| &= \left| \frac{d\varepsilon_t}{dF_{_{Y_t}}(y_t)} \frac{dF_{_{Y_t}}(y_t)}{dy_t} \right| \\ &= \left| \frac{d\varepsilon_t}{dF_{_{Y_t}}(y_t)} \frac{dF_{_{Y_t}}(y_t)}{dy_t} \right| \\ &= \left| \frac{d\varepsilon_t}{dF_{_{Y_t}}(y_t)} \right| f_{_{Y_t}}(y_t) \\ &= \left| \frac{1}{\dfrac{dF_{_{Y_t}}(y_t)}{d\varepsilon_t}} \right| f_{_{Y_t}}(y_t) \\ &= \left| \frac{1}{\dfrac{d\Phi(\varepsilon_t)}{d\varepsilon_t}} \right| f_{_{Y_t}}(y_t) \\ &= \frac{1}{f_{_N}(\varepsilon_t)} f_{_{Y_t}}(y_t). \end{aligned}$$

Luego, se tiene que

$$f_{_{Y_t}}(y_t \mid y_{t-1}, \ldots, y_1) = f_{_N}(\varepsilon_t \mid \varepsilon_{t-1}, \ldots, \varepsilon_1) \frac{f_{_{Y_t}}(y_t)}{f_{_N}(\varepsilon_t)} \tag{3.6}$$

y reemplazando las funciones de densidad condicionales en (3.5) se tiene que:

$$
\begin{aligned}
L(\theta; y) &= f_{y_1}(y_1) f_{_N}\left(\varepsilon_2 \mid \varepsilon_1\right) \frac{f_{y_2}(y_2)}{f_{_N}\left(\varepsilon_2\right)} \times \cdots \times f_{_N}(\varepsilon_{_T} \mid \varepsilon_{_{T-1}}, \ldots, \varepsilon_1) \frac{f_{y_T}(y_{_T})}{f_{_N}\left(\varepsilon_{_T}\right)} \\
&= \prod_{t=1}^{T} f_{_{Y_t}}(y_t) \prod_{t=2}^{T} f_{_N}\left(\varepsilon_t \mid \varepsilon_{t-1}, \ldots, \varepsilon_1\right) \prod_{t=2}^{T} \frac{1}{f_{_N}\left(\varepsilon_t\right)} \\
&= \prod_{t=1}^{T} \frac{f_{_{Y_t}}(y_t)}{f_{_N}\left(\varepsilon_t\right)} f_{_N}\left(\varepsilon_t\right) \prod_{t=1}^{T} f_{_N}\left(\varepsilon_t \mid \varepsilon_{t-1}, \ldots, \varepsilon_1\right) \\
&= \prod_{t=1}^{T} \frac{f_{_{Y_t}}(y_t)}{f_{_N}\left(\varepsilon_t\right)} f_{_N}\left(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{_T}; \Sigma\right), \qquad (3.7)
\end{aligned}
$$

donde $f_{_N}(\cdot; \Sigma)$ representa la función de densidad de una distribución normal T-variada con matriz de covarianza Σ. Cabe resaltar que la función de verosimilitud definida en (3.5) es el producto de un término proveniente de la regresión beta marginal y un término que modela la dependencia temporal de los errores aleatorios.

Para estimar los parámetros $\theta = (\beta, \psi, \lambda)$ se empleará la de verosimilitud definida en la ecuación (3.7), la cual es definida específicamente como

$$
\begin{aligned}
L(\theta; y) &= \left[\prod_{t=1}^{T} f_{_{Y_t}}\left(y_t\right)\right] \frac{f_{_N}\left(\varepsilon_1, \ldots, \varepsilon_{_T}; \Sigma\right)}{f_{_N}\left(\varepsilon_{_1}\right) f_{_N}\left(\varepsilon_{_2}\right), \ldots, f_{_N}\left(\varepsilon_{_T}\right)} \\
&= \left[\prod_{t=1}^{T} f_{_{Y_t}}\left(y_t\right)\right] \mathrm{q}\left(\varepsilon; \theta\right),
\end{aligned}
$$

donde $\mathrm{q}\left(\varepsilon; \theta\right)$ representa la función de densidad de la cópula gaussiana definida en la ecuación (2.16), luego

$$
L(\theta; y) = \left[\prod_{t=1}^{T} f_{_{Y_t}}\left(y_t\right)\right] \frac{1}{|\Sigma^{1/2}|} \exp\left\{-\frac{1}{2}\left(\varepsilon_1, \ldots, \varepsilon_{_T}\right)^{\top}\left(\Sigma^{-1} - I\right)\left(\varepsilon_1, \ldots, \varepsilon_{_T}\right)\right\}.
$$

La función de log-verosimilitud es definida por

$$
\ell(\theta, y) = -\frac{1}{2}\log|\Sigma| + \sum_{t=1}^{T} \log f_{_{Y_t}}\left(y_t\right) - \frac{1}{2}\left(\varepsilon_1, \ldots, \varepsilon_{_T}\right)^{\top}\left(\Sigma^{-1} - I\right)\left(\varepsilon_1, \ldots, \varepsilon_{_T}\right),
$$

donde $\varepsilon_t = \Phi^{-1}\left(F_t\left(y_t\right)\right)$.

El estimador por máxima verosimilitud (EMV) de θ es dado por

$$
\hat{\theta} = \mathrm{argmax}_{_\theta}\, \ell\left(\theta; y\right).
$$

Para obtener el EMV de θ se puede maximizar $\ell\left(\theta; y\right)$ mediante métodos numéricos como el de Broyden Fletcher Goldfarb Shanno (BFGS), el cual es un método de optimización directa.

3.4. Análisis de los residuos

Masarotto y Varin (2012) recomiendan validar un modelo de regresión con cópulas gaussianas con variables de respuesta continua mediante el comportamiento de los residuos:

$$r_t = \Phi^{-1}\left(F\left(y_t \mid y_{t-1}\dots y_1; \hat{\theta}\right)\right),$$

donde $\hat{\theta}$ representa a los parámetros estimados por máxima verosimilitud de θ. Este tipo de residuos condicionales son importantes porque involucran la distribución condicional de Y_t dadas las observaciones anteriores. Así, el modelo presentará mejores resultados en la medida que los residuos presenten una distribución normal estándar.

Adicionalmente, cabe precisar que, es posible considerar versiones marginales de los residuos; sin embargo, solo serían útiles para verificar los supuestos sobre el componente marginal del modelo y no serían informativos sobre la cópula gaussiana.

3.5. Criterio de información de Akaike - AIC

Para evaluar el desempeño de los modelos es posible usar el criterio de información Akaike (AIC) que se define como:

$$\text{AIC} = -2\hat{\ell} + 2k,$$

donde $\hat{\ell} = \ell(\hat{\theta}, y)$ y k es la dimensión de los parámetros en el modelo estimado. Un mejor desempeño se indica en la medida que este valor sea menor.

Capítulo 4

Simulación

En este capítulo se desarrollará el estudio de simulación donde se evaluará el desempeño de los estimadores obtenidos por máxima verosimilitud para el modelo definido en el Capítulo 3.

Para esto se considerarán 2 escenarios, los cuales se diferencian principalmente por la cantidad de covariables y el modelo de correlación serial elegido mediante un proceso ARMA (p, q).

4.1. Escenario 1

Se considera una covariable y un proceso ARMA (p, q). La implementación del estudio de simulación se realiza de la siguiente manera:

1) Establecer el tamaño de la muestra T y definir cada proceso ARMA (p, q) para $p = 0, 1, 2$ y $q = 0, 1, 2$ según se muestra en el Cuadro 4.1:

Cuadro 4.1: Modelos ajustados según los procesos ARMA(p, q) , definición del error aleatorio correspondiente y sus parámetros.

Proceso ARMA(p, q)	Errores aleatorios	Parámetros
ARMA$(1, 0)$	$\varepsilon_t = \phi_1 \varepsilon_{t-1} + \eta_t$	ϕ_1
ARMA$(0, 1)$	$\varepsilon_t = \lambda_1 \eta_{t-1} + \eta_t$	λ_1
ARMA$(1, 1)$	$\varepsilon_t = \phi_1 \varepsilon_{t-1} + \lambda_1 \eta_{t-1} + \eta_t$	ϕ_1, λ_1
ARMA$(1, 2)$	$\varepsilon_t = \phi_1 \varepsilon_{t-1} + \lambda_1 \eta_{t-1} + \lambda_2 \eta_{t-2} + \eta_t$	ϕ_1, λ_1, λ_2
ARMA$(2, 1)$	$\varepsilon_t = \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \lambda_1 \eta_{t-1} + \eta_t$	ϕ_1, ϕ_2, λ_1

2) Generar T errores aleatorios ϵ_t, $t = 1, \ldots, T$, a partir de la cópula gaussiana cada proceso ARMA (p, q) elegido. Es decir se generan T errores aleatorios a partir de una distribución acumulada normal, $\epsilon \sim \mathrm{N}(0, \Sigma)$ donde la covarianza depende de cada proceso ARMA (p, q) elegido.

3) Se define la covariable. Para este caso se va a considerar el tiempo estandarizado $\widehat{t} = \dfrac{t - \overline{t}}{\sqrt{\mathrm{var}(t)}}$, donde $\overline{t} = \frac{\sum_{t=1}^{T} t}{T}$ y $\mathrm{var}(t) = \frac{\sum_{t=1}^{T} (t - \overline{t})^2}{T-1}$.

4) Se generan μ_t y κ_t, para $t = 1, \ldots, 500$, en función de los parámetros según cada proceso

ARMA(p, q) propuesto:

$$\text{logit}(\mu_t) = \beta_{0X} + \beta_{1X}\widehat{t},$$
$$\log(\kappa_t) = \beta_{0Z}\widehat{t}.$$

5) Se generan los valores observados de la variable aleatoria Y_t, para $t = 1, \ldots, 500$, mediante la inversa de la función de distribución acumulada de una distribución beta $Y = F_Y^{-1}(\Phi_\Sigma)$, para $Y = (Y_1, \ldots, Y_n)^\top$.

6) Estimar los parámetros $\theta_i = (\beta_{0x}, \beta_{1x}, \beta_{0z}, \phi_1, \phi_2, \lambda_1, \lambda_2)$ de acuerdo a cada proceso ARMA (p, q), $i = 1, 2, 3, 4, 5$ (ver Cuadro 4.2), mediante máxima verosimilitud.

Cuadro 4.2: Modelos del escenario 1 según los procesos ARMA(p, q) y sus parámetros

Proceso ARMA(p, q)	Parámetros
ARMA$(1, 0)$	$\theta_1 = (\beta_{0x}, \beta_{1x}, \beta_{0z}, \phi_1)$
ARMA$(0, 1)$	$\theta_2 = (\beta_{0x}, \beta_{1x}, \beta_{0z}, \lambda_1)$
ARMA$(1, 1)$	$\theta_3 = (\beta_{0x}, \beta_{1x}, \beta_{0z}, \phi_1, \lambda_1)$
ARMA$(1, 2)$	$\theta_4 = (\beta_{0x}, \beta_{1x}, \beta_{0z}, \phi_1, \lambda_1, \lambda_2)$
ARMA$(2, 1)$	$\theta_5 = (\beta_{0x}, \beta_{1x}, \beta_{0z}, \phi_1, \phi_2, \lambda_1)$

En la Figura 4.1, se muestran las series simuladas de tamaño $T = 500$ de acuerdo a cada uno de los procesos ARMA (p, q) y los valores originales de los parámetros, respectivamente. En general, se observa que la serie temporal presenta una ligera tendencia decreciente.

Los resultados obtenidos EMVs e intervalos al 95 % de confianza para el escenario 1 son presentados en el Cuadro 4.3. Observamos que, en general, los EMVs presentan resultados próximos a los valores originales que generaron los datos.

Por otro lado, en el Cuadro 4.4 se muestran algunos resultados respecto del sesgo para $M = 500$ simulaciones para tres diferentes tamaños de muestra, tales como $T = 100, 250$, y 500. El objetivo es evaluar la consistencia de los estimadores analizando el sesgo de las estimaciones de los coeficientes de regresión, parámetro de precisión y parámetros asociados a los tres primeros modelos con procesos ARMA$(1, 0)$, ARMA$(0, 1)$ y ARMA$(1, 1)$. Cabe precisar que, para el modelo con proceso ARMA$(1, 1)$, con mayor cantidad de parámetros,solo se consideró la muestra de tamaño $T = 500$. Según estos resultados, el sesgo calculado en base a los resultados de EMV para la variable respuesta fueron similares en todos los modelos; así, se tiene que el sesgo de los coeficientes de regresión de la media es mínimo y disminuye en la medida que el tamaño de la muestra aumenta. También, el sesgo de los coeficientes asociados al parámetro de precisión es pequeño, aunque mayor en comparación a los otros parámetros y finalmente, el sesgo para el parámetro asociado al proceso autoregresivo disminuye en la medida que el tamaño de la muestra aumenta y el sesgo para el parámetro asociado al proceso de medias móviles es mínimo conforme el el tamaño de la muestra aumenta.

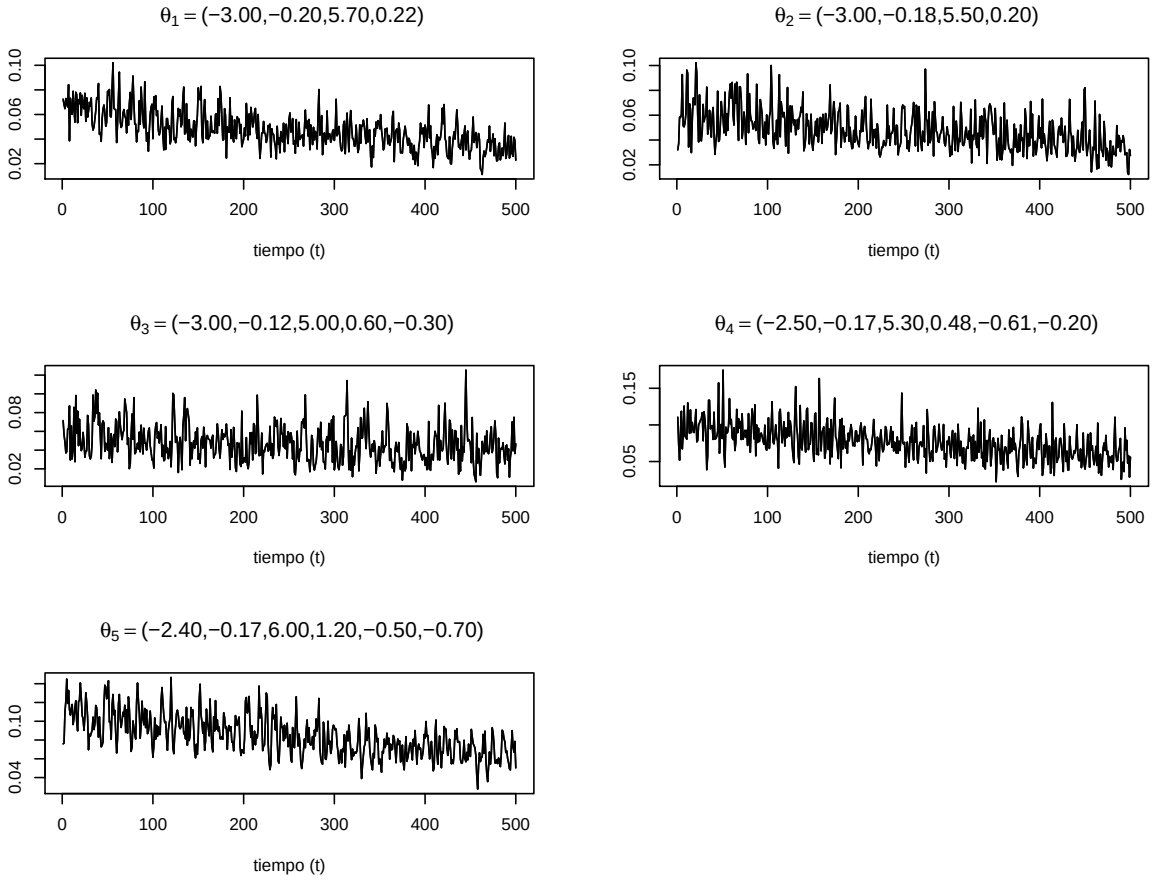


Figura 4.1: Series de tiempo simuladas usando el modelo propuesto usando cópulas gaussianas y procesos ARMA(p, q). El modelo ARMA$(1, 0)$ tiene parámetros θ_1, el modelo ARMA$(0, 1)$ tiene parámetros θ_2, el modelo ARM$A(1, 1)$ tiene parámetros θ_3, el modelo ARMA$(1, 2)$ tiene parámetros θ_4y el modelo ARMA$(2, 1)$ tiene parámetros θ_5.

4.2. Escenario 2

Se consideran dos covariables (tiempo estandarizado y una variable que presenta una distribución normal) y un proceso ARMA (p, q). A diferencia del escenario anterior, se considerará una covariable adicional solo para el caso de modelar el parámetro de la media, mas no el de la precisión. Así, el punto 4). se modificará a:

$$\begin{aligned} \text{logit}(\mu_t) &= \beta_{0X} + \beta_{1X}\widehat{t} + \beta_{2X}x_t, \\ \log(\kappa_t) &= \beta_{0Z}, \end{aligned}$$

donde x_t se genera a partir de una distribución normal estándar, $t = 1, \ldots, 500$.

Para este escenario se considerarán los vectores de parámetros que se muestran en el Cuadro 4.5. El Cuadro 4.6 muestra los resultados obtenidos para este escenario. Los resultados sugieren que los modelos consiguieron ajustar de manera satisfactoria la estructura adoptada en la generación de los datos. Se observa que los EMVs para los parámetros en general fueron similares a los valores originales del verdadero modelo. Los coeficientes de regresión para la media son los que se aproximan más a los verdaderos valores de los parámetros. Mientras que

Cuadro 4.3: Resultados del escenario 1: Estimación (EMV) e intervalo al 95 % de confianza.

	Parámetros	Valor original	Estimación	IC
Media	β_{0x}	-3.00	-2.99***	$(-3{,}02, -2{,}96)$
	β_{1x}	-0.20	-0.19***	$(-0{,}22, -0{,}17)$
Precisión	β_{0z}	5.70	5.56***	$(4{,}42, 5{,}70)$
Error: ARMA$(1,0)$	ϕ_1	0.22	0.22***	$(0{,}13, 0{,}30)$
Media	β_{0x}	-3.00	-2.98***	$(-3{,}00, -2{,}95)$
	β_{1x}	-0.18	-0.17***	$(-0{,}20, -0{,}14)$
Precisión	β_{0z}	5.50	5.42***	$(5{,}30, 5{,}55)$
Error: ARMA$(0,1)$	λ_1	0.20	0.15***	$(0{,}07, 0{,}24)$
Media	β_{0x}	-5.00	-2.99***	$(-3{,}00, -2{,}93)$
	β_{1x}	-0.12	-0.12***	$(-0{,}18, -0{,}07)$
Precisión	β_{0z}	5.00	4.96***	$(4{,}80, 5{,}11)$
Error:	ϕ_1	0.60	0.67***	$(0{,}52, 0{,}82)$
ARMA$(1,1)$	λ_1	-0.30	-0.35***	$(-0{,}54, 0{,}16)$
Media	β_{0x}	-2.50	-2.50***	$(-2{,}51, -2{,}49)$
	β_{1x}	-0.17	-0.17***	$(-0{,}18, -0{,}16)$
Precisión	β_{0z}	5.30	5.28***	$(5{,}14, 5{,}4)$
Error:	ϕ_1	0.48	0.50***	$(0{,}31, 0{,}70)$
ARMA$(1,2)$	λ_1	-0.61	-0.62***	$(-0{,}83, -0{,}41)$
	λ_2	-0.20	-0.19***	$(-0{,}31, -0{,}06)$
Media	β_{0x}	-2.40	-2.39***	$(-2{,}41, -2{,}37)$
	β_{1x}	-0.17	-0.15***	$(-0{,}18, -0{,}12)$
Precisión	β_{0z}	6.00	5.79***	$(5{,}40, 6{,}16)$
Error:	ϕ_1	1.20	1.30***	$(1{,}00, 1{,}60)$
ARMA$(2,1)$	ϕ_2	-0.50	-0.56***	$(-0{,}72, -0{,}39)$
	λ_1	-0.70	-0.84***	$(-1{,}21, -0{,}48)$
Significancia: '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Cuadro 4.4: Resultados del escenario 1 considerando 500 réplicas para diferentes tamaños de muestras.

			$T = 100$		$T = 250$		
	Parámetros	Valor original	Sesgo	ECM	Sesgo	ECM	Ses
Media	β_{0x}	-3.00	0.00		0.00		0.0
	β_{1x}	-0.20	0.00		0.00		0.0
Precisión	β_{0z}	5.70	-0.00	0.00	-0.03	0.00	-0.0
Error: ARMA$(1,0)$	ϕ_1	-0.22	-0.03		0.01		0.0
Media	β_{0x}	-3.00	0.00		0.00		0.0
	β_{1x}	-0.18	0.00		0.00		0.0
Precisión	β_{0z}	5.50	-0.01	0.00	-0.03	0.00	-0.0
Error: ARMA$(0,1)$	λ_1	-0.22	-0.02		-0.01		0.0
Media	β_{0x}	-3.00					0.0
	β_{1x}	-0.12					0.0
Precisión	β_{0z}	5.00					0.1
Error:	ϕ_1	0.60					0.0
ARMA$(1,1)$	λ_1	-0.30					0.0

el coeficiente de regresión asociado al parámetro de precisión tiene un mayor sesgo, respecto a los otros parámetros.

Cuadro 4.5: Modelos para el escenario 2 según los procesos ARMA(p, q) y sus parámetros.

Proceso ARMA(p, q)	Parámetros
ARMA $(1, 0)$	$\theta_1 = (\beta_{0x}, \beta_{1x}, \beta_{2x}, \beta_{0z}, \phi_1)$
ARMA $(0, 1)$	$\theta_2 = (\beta_{0x}, \beta_{1x}, \beta_{2x}, \beta_{0z}, \lambda_1)$
ARMA $(1, 1)$	$\theta_3 = (\beta_{0x}, \beta_{1x}, \beta_{2x}, \beta_{0z}, \phi_1, \lambda_1)$
ARMA $(1, 2)$	$\theta_4 = (\beta_{0x}, \beta_{1x}, \beta_{2x}, \beta_{0z}, \phi_1, \lambda_1, \lambda_2)$
ARMA $(2, 1)$	$\theta_5 = (\beta_{0x}, \beta_{1x}, \beta_{2x}, \beta_{0z}, \phi_1, \phi_2, \lambda_1)$

Cuadro 4.6: Resultados del escenario 2: Estimación (EMV) e intervalos al 95 % de confianza.

			T=500	
	Parámetros	Valor original	Estimación	IC
Media	β_{0x}	-3.00	-3.01***	$(-3{,}10, -2{,}92)$
	β_{1x}	-0.18	-0.20***	$(-0{,}25, -0{,}16)$
	β_{2x}	-0.20	-0.19***	$(-0{,}21, -0{,}17)$
Precisión	β_{0z}	5.30	5.65***	$(5{,}52, 5{,}79)$
Error: ARMA $(1, 0)$	ϕ_1	0.20	0.26***	$(0{,}18, 0{,}34)$
Media	β_{0x}	-3.20	-3.23***	$(-3{,}31, -3{,}12)$
	β_{1x}	-0.14	-0.14***	$(-0{,}18, -0{,}11)$
	β_{2x}	-0.12	-0.11***	$(-0{,}13, -0{,}09)$
Precisión	β_{0z}	6.00	5.95***	$(5{,}82, 6{,}08)$
Error: ARMA $(0, 1)$	λ_1	-0.18	0.20***	$(0{,}12, 0{,}29)$
Media	β_{0x}	-3.00	-3.01***	$(-3{,}10, -2{,}93)$
	β_{1x}	-0.14	-0.15***	$(-0{,}21, -0{,}08)$
	β_{2x}	-0.12	-0.12***	$(0{,}14, -0{,}11)$
Precisión	β_{0z}	5.75	5.72****	$(5{,}57, 5{,}87)$
Error:	ϕ_1	0.90	0.91***	$(0{,}84, 0{,}98)$
ARMA $(1, 1)$	λ_1	-0.75	-0.80***	$(-0{,}89, -0{,}71)$
Media	β_{0x}	-2.38	-2.38***	$(-2{,}44, -2{,}32)$
	β_{1x}	-0.13	-0.12***	$(-0{,}17, -0{,}05)$
	β_{2x}	-0.11	-0.10***	$(-0{,}11, -0{,}10)$
Precisión	β_{0z}	6.4	6.18***	$(5{,}95, 6{,}41)$
Error:	ϕ_1	0.96	0.96***	$(0{,}91, 1{,}00)$
ARMA $(1, 2)$	λ_1	-0.61	-0.54***	$(-0{,}63, -0{,}44)$
	λ_2	-0.24	-0.30***	$(-0{,}39, -0{,}22)$
Media	β_{0x}	-2.42	-2.37***	$(-2{,}44 - 2{,}32)$
	β_{1x}	-0.10	-0.10***	$(-0{,}12, -0{,}09)$
	β_{2x}	-0.13	-0.14***	$(-0{,}15, 0{,}13)$
Precisión	β_{0z}	5.80	5.72***	$(5{,}58, 5{,}86)$
Error:	ϕ_1	0.90	0.91***	$(0{,}75, 1{,}07)$
ARMA $(2, 1)$	ϕ_2	-0.30	-0.34***	$(-0{,}43, -0{,}26)$
	λ_1	-0.80	-0.74***	$(-0{,}89, -0{,}58)$
Significancia: '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Por otro lado, en el Cuadro 4.7 se muestran algunos resultados respecto del sesgo y error cuadrático medio (ECM) para $M = 500$ réplicas correspondiente a un tamaño de la muestra $T = 500$. Al respecto, se observa que los sesgos son mínimos para todos los valores considerados, principalmente, en para el parámetro de la media. Para el caso del ECM, se

observa que el modelo presenta un buen desempeño, pues se encuentran alrededor de 0,00.

Cuadro 4.7: Resultados del escenario 2 considerando 500 réplicas para diferentes tamaños de muestras.

	Parámetros	Valor original	Sesgo ($T = 500$)	ECM
Media	β_{0x}	-3.00	0.00	
	β_{1x}	-0.18	0.00	
	β_{2x}	-0.20	0.00	
Precisión	β_{0z}	5.80	-0.03	0.00
Error: ARMA$(1, 0)$	ϕ_1	-0.20	-0.01	
Media	β_{0x}	-3.2	0.00	
	β_{1x}	-0.14	0.00	
	β_{2x}	-0.12	0.00	
Precisión	β_{0z}	6.00	0.01	0.00
Error: ARMA$(0, 1)$	λ_1	-0.15	0.00	
Media	β_{0x}	-2.38	0.00	
	β_{1x}	-0.10	0.00	
	β_{2x}	-0.20	0.00	
Precisión	β_{0z}	5.10	0.08	0.00
Error:	ϕ_1	0.60	0.03	
ARMA$(1, 1)$	λ_1	0.20	0.03	

Capítulo 5

Aplicación a la tasa de desempleo en Lima Metropolitana

En esta sección se realizará la aplicación del modelo y la metodología propuesta a la tasa de desempleo mensual en Lima Metropolitana para el periodo que abarca desde enero de 2003 a octubre de 2019, puesto que se trata de una variable cuyos valores se encuentran en el intervalo $(0, 1)$. La importancia de la aplicación a este conjunto de datos radica en que los valores serán analizados en su forma original, evitando la transformación de la serie así como su desestacionalización.

5.1. Descripción de los datos

El conjunto de datos que se usa en esta tesis es referente a la tasa de desempleo mensual en Lima Metropolitana desde enero de 2003 a octubre de 2019. Esta tasa es calculada por el Instituto Nacional de Estadística e Informática (INEI).

Se define $Y_t \in (0, 1)$ como una variable aleatoria que representa la tasa de desempleo en el tiempo $t = 1, \ldots, T$, donde $T = 202$. La Figura 5.1 muestra que los datos presentan una asimetría a la derecha. Asimismo, en la Figura 5.1 se observa la serie temporal correspondiente a la tasa de desempleo y se evidencia que tiene una tendencia decreciente.

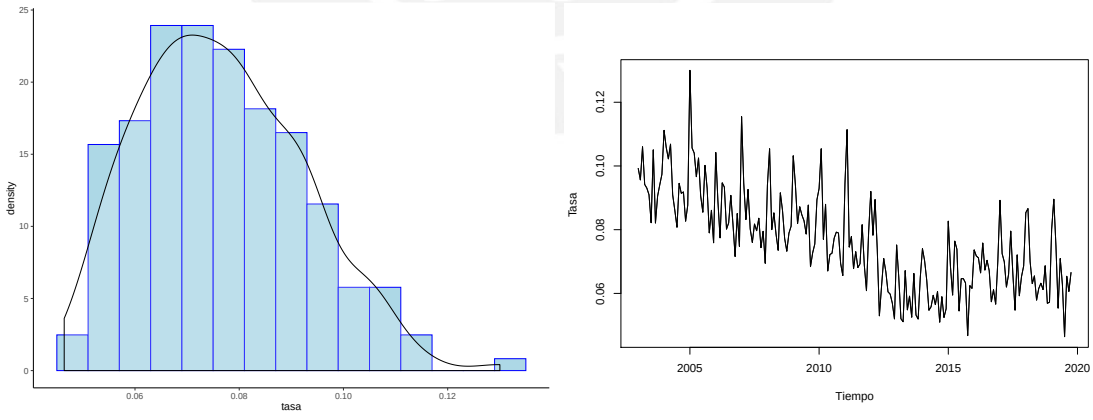


Figura 5.1: Izquierda: Histograma de la Tasa de desempleo en Lima Metropolitana (enero 2003 - octubre 2019). Derecha: Serie de tiempo de la tasa de desempleo en Lima Metropolitana (enero 2003 - octubre 2019).

Como potenciales covariables se consieran el índice del Producto Bruto Interno (PBI) obtenido del Banco Central de Reserva del Perú y el año correspondiente a la tasa de desempleo. La Figura 5.2 muestra una asociación lineal inversa entre el logaritmo del PBI y la tasa de

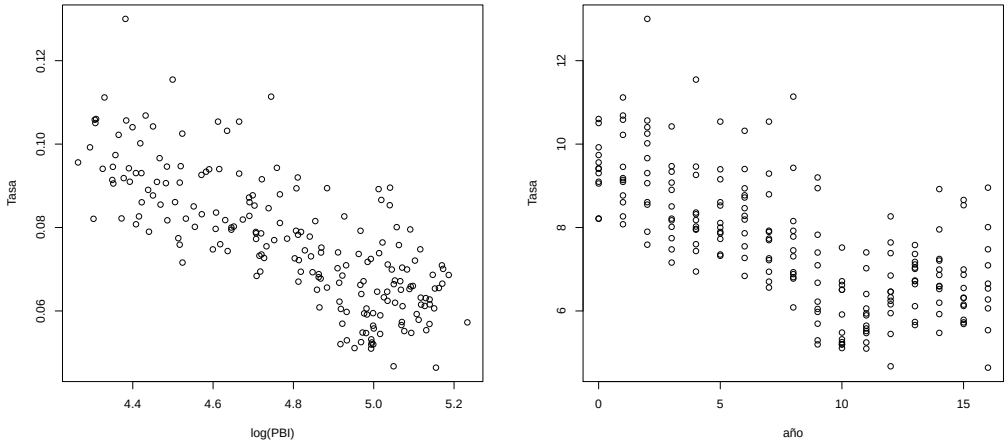


Figura 5.2: Izquierda: El logaritmo del PBI vs Tasa de desempleo en Lima Metropolitana (enero 2003 - octubre 2019). Derecha: El año vs. Tasa de desempleo, en Lima Metropolitana (enero 2003 - octubre 2019).

desempleo. La correlación entre ambas variables es -0.76. De la misma forma, la Figura 5.2 muestra una posible asociación cuadrática entre el el año y la tasa de desempleo.

5.2. Modelo de regresión beta

Para determinar el proceso autorregresivo y de medias móviles se consideraron los residuos que se obtuvieron de una regresión beta aplicada a la tasa de desempleo.

Sea $\mathbf{Y} = (Y_1, \ldots, Y_{202})^\top$ un vector de variables aleatorias donde se asume que la distribución marginal de Y_t es una distribución beta, $Y_t \sim \text{beta}\,(\mu_t,\, \kappa_t)\,, \forall\ t = 1, \ldots, 202$; es decir, la función de densidad marginal de Y_t es dada por:

$$f_{Y_t}\,(y_t) = \frac{\Gamma\,(\kappa_t\,)}{\Gamma\,(\mu_t\kappa_t)\,\Gamma\,((1-\mu_t)\,\kappa_t)} y_t^{\mu_t\kappa_t - 1}\,(1-y_t)^{(1-\mu_t)\kappa_t - 1}\,,$$

donde $0 < \mu_t < 1$ y parámetro de precisión $\kappa_t > 0$.

Se define el modelo de regresión beta a través de funciones de enlace $g_1\,(\cdot)$ y $g_2\,(\cdot)$ que asocian la media μ_t y precisión κ_t a sus respectivos predictores lineales. En particular, se asume:

$$g_1(\mu_t) = \text{logit}\,(\mu_t) = \beta_{0x} + \beta_{1x} x_{1t},$$

$$g_2(\kappa_t) = \log\,(\kappa_t) = \beta_{z0},$$

donde x_{1t} representa a la covariable logaritmo del PBI. Además para la precisión solo se toma en cuenta el intercepto β_{z0}, debido a que de acuerdo a un análisis exploratorio, en los modelos donde se incluyó una covariable, los respectivos parámetros para la precisión resultaron ser no significativos.

En particular, no estamos interesados en las estimaciones de este modelo de regresión beta porque no considera la dependencia temporal entre las observaciones. El interés, se centra analizar los residuales $r_t = y_t - \hat{y}_t$, de dicho modelo. A través de estos residuales se muestra

que la tendencia es modelada por las covariables que evidencien la dependencia temporal en los datos (ver Figura 5.3).

En la Figura 5.3 también se observan los correlogramas de autocorrelación (ACF) y autocorrelación parcial (PACF) de los residuales obtenidos de la regresión beta ajustada. Estos resultados son útiles para decidir qué proceso ARMA(p, q) debe considerarse para ajustar la dependencia temporal de la tasa de desemeplo. Según la figura de autocorrelación parcial, puede considerarse un proceso AR(1) y según la figura de autocorrelación se puede considerar hasta un proceso MA(1). Finalmente, el ACF y PACF de los residuales también muestran una ligera estacionalidad, más evidente en el ACF.

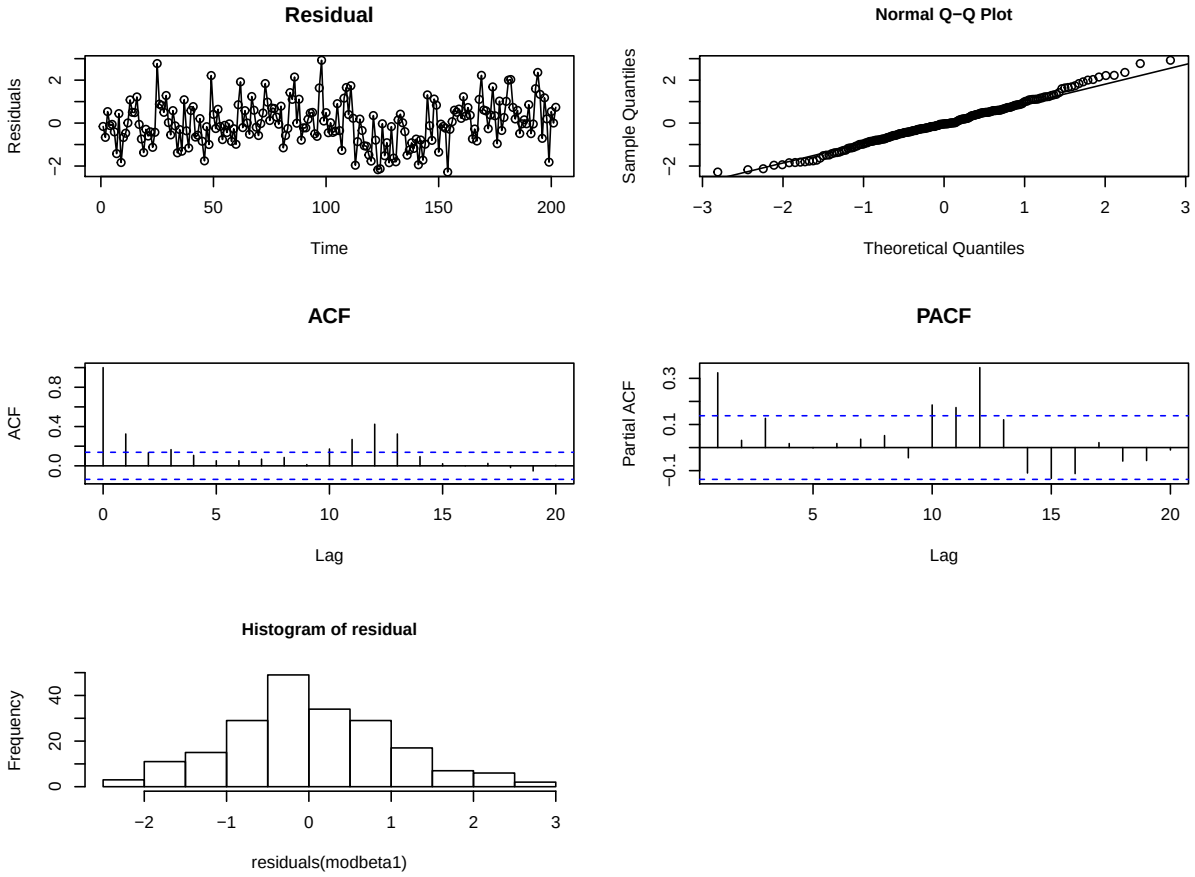


Figura 5.3: Análisis de errores de una regresión beta sin considerar una dependencia temporal para la tasa de desempleo Lima Metropolitana (enero 2003 - octubre 2019)

5.3. Modelo de regresión beta usando cópulas gaussianas

Sea $\mathbf{Y} = (Y_1, \ldots, Y_{202})^\top$ un vector de variables aleatorias donde las Y_t se asume que la distribución marginal de Y_t es una distribución beta, es decir, $Y_t \sim \text{beta}(\mu_t, \kappa_t), \forall\ t = 1, \ldots, 202$. Se define el modelo marginal de regresión beta a través de funciones de enlace $g_1(\cdot)$ y $g_2(\cdot)$ que asocian la media μ_t y precisión κ_t a sus respectivos predictores lineales. En particular, se asume:

$$g_1(\mu_t) = \text{logit}\,(\mu_t) = \beta_{0x} + \beta_{1x}x_t + \beta_{2x}sin(2*\pi/365*t) + \beta_{3x}cos(2*\pi/365*t) + \beta_{4x}\text{año}_t,$$

$$g_2(\kappa_t) = \log(\kappa_t) = \beta_{z0},$$

donde x_t representa a la covariable logaritmo del PBI. La estacionalidad anual del modelo se modela a través de las funciones seno y coseno. Y la tendencia de la tasa de desempleo anual se modela a través de la covariable año_t que corresponde a la observación en el tiempo t. Además para la precisión solo se toma en cuenta el intercepto β_{z0}, debido a que de acuerdo a un análisis exploratorio, en los modelos donde se incluyó una covariable, los respectivos parámetros para la precisión resultaron ser no significativos.

Para incorporar la dependencia temporal entre las tasas de desempleo Y_t y $Y_{t'}$, se define el error aleatorio ε_t tal como fue definida en la ecuación 3.4:

$$\varepsilon_t = \sum_{i=1}^{p} \phi_i \varepsilon_{t-i} + \sum_{j=1}^{q} \lambda_j \eta_{t-j} + \eta_t,\ t = 1, \ldots, 202, \tag{5.1}$$

donde ϕ_i representa el componente del proceso autorregresivo AR(p) y λ_j al proceso de medias móviles MA(q). Luego, se asume que los errores $\varepsilon_t \sim N(0, \Sigma)$ son autocorrelacionados temporalmente a través de la definición de la función de distribución acumulada usando una cópula gaussiana, tal como se define en la ecuación 3.3:

$$F_Y(y_1, \ldots, y_{202}) = \Phi_\Sigma(\varepsilon_1, \ldots, \varepsilon_{202}) = \Phi_\Sigma(\varepsilon),$$

donde Σ es la matriz de correlación del proceso ARMA(p, q).

Una vez definidas las distribuciones marginales de Y_t, la función de densidad conjunta de $\mathbf{Y}$ es definida usando la cópula gaussiana, tal que La función de verosimilitud la cual es definida específicamente como:

$$L(\theta; y) \;=\; \left[\prod_{t=1}^{202} f_{Y_t}(y_t)\right] \text{q}(\varepsilon; \theta),$$

donde $\text{q}(\varepsilon; \theta)$ representa la función de densidad de la cópula gaussiana definida en la ecuación (2.16), luego

$$\begin{aligned} L(\theta; y) &= \left[\prod_{t=1}^{202} f_{Y_t}(y_t)\right] \frac{1}{|\Sigma^{1/2}|} \exp\left\{-\frac{1}{2}(\varepsilon_1, \ldots, \varepsilon_{202})^\top \left(\Sigma^{-1} - I\right)(\varepsilon_1, \ldots, \varepsilon_{202})\right\}, \\ &= \left[\prod_{t=1}^{202} (y_t) = \frac{\Gamma(\kappa_t)}{\Gamma(\mu_t\kappa_t)\Gamma\{(1-\mu_t)\kappa_t\}} y_t^{\mu_t\kappa_t - 1}(1-y_t)^{(1-\mu_t)\kappa_t - 1}(y_t)\right] \\ &\quad \frac{1}{|\Sigma^{1/2}|} \exp\left\{-\frac{1}{2}(\varepsilon_1, \ldots, \varepsilon_{202})^\top \left(\Sigma^{-1} - I\right)(\varepsilon_1, \ldots, \varepsilon_{202})\right\}, \end{aligned}$$

donde $\varepsilon_t = \Phi^{-1}(F_t(y_t))$.

Para estimar los posibles parámetros $\theta = (\beta_{0x}, \beta_{1x}, \beta_{2x}, \beta_{0z}, \lambda_1, \lambda_2, \phi_1, \phi_2)$ se emplea el paquete *gcrm* publicado por *Journal of Statistical Software* y disponible en el software libre *R* (Masarotto y Varin, 2017). El método de estimación es por máxima verosimilitud, usando

métodos numéricos.

En ese contexto, se procedió a ajustar diversos modelos para capturar la autocorrelación temporal en la tasa de desempleo mensual. Estos modelos consisten básicamente en variaciones del proceso ARMA(p, q) definido en la ecuación 5.1. Solo se considera procesos ARMA hasta de orden 2, tomando en cuenta el modelo de regresión beta ajustado en la sección previa, y que de acuerdo a análisis no presentados en la tesis, para valores de p y q superiores a dos, los parámetros no resultan significativos. En el Cuadro 5.1 se muestra los modelos ajustados según los procesos ARMA(p, q) correspondientes, y sus respectivos valores de AIC. Se puede observar que el modelo que presentó menor AIC consideró un proceso temporal ARMA$(1, 1)$. Este resultado concuerda con el resultado de analizar los residuales del modelo de regresión beta sin efecto temporal ajustado en la sección previa.

Cuadro 5.1: Modelos ajustados según los procesos ARMA(p, q) , definición del error aleatorio correspondiente y valores de AIC de cada modelo ajustado.

Proceso ARMA(p, q)	Errores aleatorios	AIC
$p = 0$, $q = 1$	$\varepsilon_t = \lambda_1 \eta_{t-1} + \eta_t$	-1353.2
$p = 1$, $q = 0$	$\varepsilon_t = \phi_1 \varepsilon_{t-1} + \eta_t$	-1353.5
$\mathbf{p = 1, q = 1}$	$\boldsymbol{\varepsilon_t = \phi_1 \varepsilon_{t-1} + \lambda_1 \eta_{t-1} + \eta_t}$	**-1356.2**
$p = 1$, $q = 2$	$\varepsilon_t = \phi_1 \varepsilon_{t-1} + \lambda_1 \eta_{t-1} + \lambda_2 \eta_{t-2} + \eta_t$	-1354.8
$p = 2$, $q = 1$	$\varepsilon_t = \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \lambda_1 \eta_{t-1} + \eta_t$	-1354.6

Masarotto y Varin (2012) recomiendan validar el modelo mediante el comportamiento de los residuos:

$$r_t = \Phi^{-1}\left(F\left(y_t \mid y_{t-1} \dots y_1; \hat{\theta}\right)\right),$$

donde $\hat{\theta}$ representa la estimación por máxima verosimilitud del vector de parámetros θ y $F\left(y_t \mid y_{t-1} \dots y_1; \hat{\theta}\right)$ representa la función de distribución acumulada condicional de Y_t dadas las observaciones anteriores. Este tipo de residuos condicionales toman en cuenta la cópula gaussiana usada para incorporar la dependencia temporal.

La figura 5.4 muestra resúmenes para estos residuales r_t, en particular para el modelo considerando un proceso temporal ARMA$(1, 1)$. Se observa los residuales no presentan ningún patrón, por lo cual podemos concluir que la tendencia es modelada adecuadamente por las covariables y que no se evidencia más la dependencia temporal entre los errores aleatorios. También se observa el correlograma de autocorrelación (ACF) de los residuales obtenidos del modelo seleccionado. Estos resultados indican que el modelo seleccionado ajusta la dependencia temporal de la tasa de desempleo adecuadamente. Además, muestran que no hay más picos en el lag 12 y 13. Adicionalmente, los residuos se distribuyen normalmente en la medida que los valores se encuentran dentro de la banda. Las figuras S.1, S.2, S.3, S.4 en el anexo A se muestran resumenes para los residuales r_t, para los otros modelos, se observa que los residuales presentan aún tendencia y además están autocorrelacionados en los modelos que consideran el procesos ARMA$(1, 0)$ y ARMA$(0, 1)$. Mientras que los modelos con procesos ARMA$(1, 2)$ y ARMA$(2, 1)$ presentan resultados similares a los expuesto en el modelo con proceso ARMA$(1, 1)$.

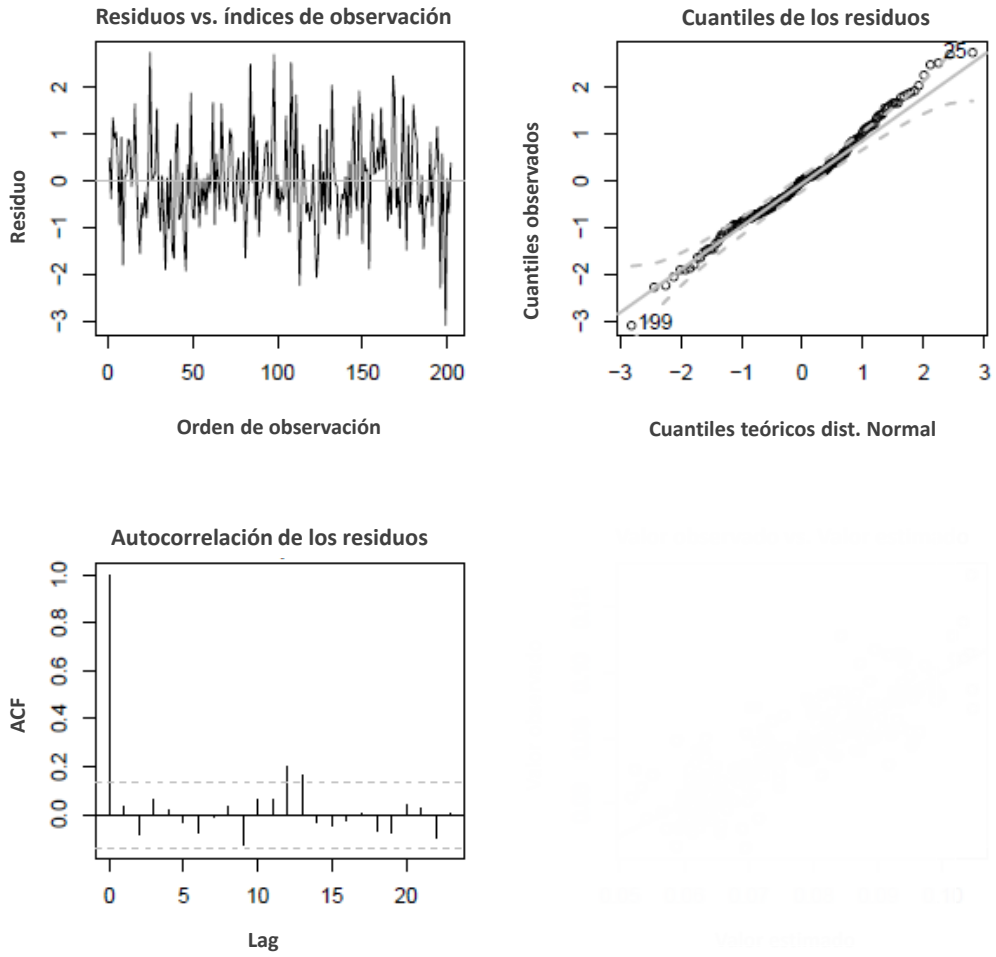


Figura 5.4: Análisis de residuos del modelo marginal de Regresión beta con cópulas gaussianas considerando un modelo ARMA $(1, 1)$.

Las estimaciones para todos los modelos ajustados se encuentran en el Cuadro del anexo A. En términos generales vemos que los coeficientes de regresión para la media y precisión son similares para los modelos Sin embargo, como el modelo ARMA$(1, 1)$ presenta el menor AIC y el análisis de sus residuales es adecuado, se selecciona este modelo para continuar con el análisis de los parámetros estimados. Los resultados de las estimaciones de los parámetros del modelo ARMA$(1, 1)$ y sus errores estándar respectivos, se encuentran en el Cuadro 5.2. Como se observa, todas las estimaciones de los parámetros son significativas.

Respecto del coeficiente β_{1x} se tiene que dado exp$(-1{,}35)$=0.26, ante el incremento de una unidad en el logaritmo del índice del PBI, el odds de la tasa de desempleo se reduce en 74 %. Respecto del coeficiente β_{4x} se tiene que dado exp$(0{,}11)$=1.12, el odds de la tasa de desempleo aumenta en 12 % por cada año adicional, esta es una tendencia que se observó desde el año 2015.

Respecto a los parámetros que definen el proceso $ARMA$, el parámetro del proceso autoregresivo, $\phi_1 = 0{,}91$, indica que hay una dependencia temporal fuerte entre la tasa de

Cuadro 5.2: Resultados de la aplicación a datos de desempleo: Estimación (EMV) y errores estándar para el modelo seleccionado usando un proceso ARMA$(1,1)$.

		Proceso ARMA (1,1)
Parámetros	Estimación puntual	Error estándar
β_{0x}	3.10***	0.94
β_{1x}	-1.35***	0.19
β_{2x}	0.11*	0.05
β_{3x}	0.46**	0.17
β_{4x}	0.11***	0.02
β_{0z}	6.87***	0.13
ϕ_1	0.91***	0.07
λ_1	-0.78***	0.09
Significancia:	'***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1	

desempleo en el mes t y el mes $t-1$, especificamente, el efecto temporal de un mes t es 0.91 veces el efecto temporal del mes anterior $t-1$. El parámetro del proceso de medias móviles, $\lambda_1 = -0{,}78$ indica que la variabilidad restante en el efecto temporal de un mes t se reduce en 0.78 veces la variabilidad restante del mes anterior $t-1$. Finalmente, el parámetro de precisión $\kappa_t = \exp(6{,}87) = 962{,}95$, por lo tanto la varianza marginal de la tasa de desempleo $Var\left(Y_t\right) = \dfrac{\mu_t\left(1-\mu_t\right)}{1+\kappa_t}$ es pequeña, esto indica que la variabilidad restante en el modelo luego de incluir el proceso temporal $ARMA$ es mínima.

Finalmente, en la Figura 5.5 muestra la serie temporal real de la tasa de desempleo versus la serie temporal estimada de esta variable con el modelo propuesto, se puede observar un ajuste bastante satisfactorio.

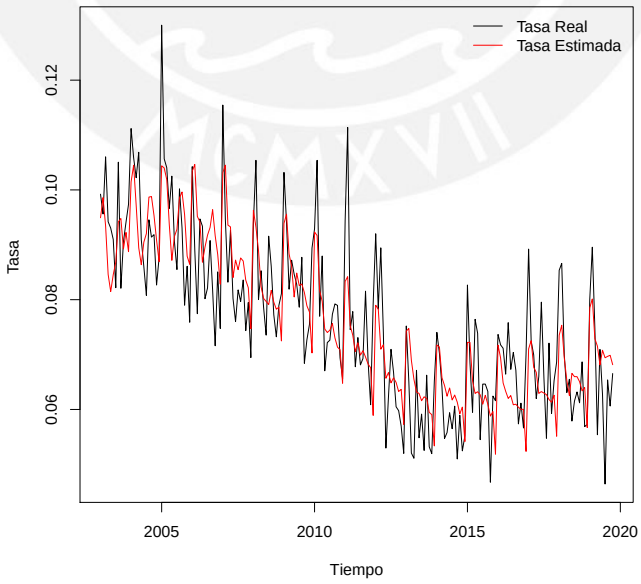


Figura 5.5: Tasa de desempleo real vs. la tasa de desempleo estimada usando el modelo de regresión beta usando cópulas gaussianas a través de un proceso temporal ARMA (1,1).

Capítulo 6

Conclusiones

En el presente trabajo se ha presentado el modelo de regresión beta con cópulas gaussianas para series de tiempo. Este modelo presenta una forma particular para analizar la dependencia temporal de las distribuciones marginales de series que encuentran restringidas al intervalo $(0, 1)$ sin realizar transformaciones a la misma.

Mediante el modelo propuesto es posible analizar los efectos que presentan otras covariables diferenciando los efectos ocasionados por la dependencia temporal. Dicha dependencia es modelada mediante una cópula gaussiana, lo cual es posible debido al comportamiento de los errores que se obtienen de aplicar una regresión beta al conjunto de datos.

El estudio de simulación permite evidenciar que el desempeño de los parámetros mejora en la medida que se cuente con una muestra más grande y que en esos casos se pueden emplear procesos autorregresivos de medias móviles de órdenes superiores. Asimismo, con los resulados del error cuadrático medio, se observa el buen desempeño del modelo.

De la aplicación a los datos de la tasa de desempleo en Lima metropolitana, desde enero de 2003 hasta octubre de 2019, se observa que a pesar de que la serie no es estacionaria no fue necesario desestacionalizarla, debido a que el modelo permite modelar la tendencia de la serie temporal a través de la media marginal de la distribución beta, la cual se modeló incluyendo el logaritmo del PBI como covariable, el año y un efecto que ajusta la estacionalidad anual del modelo. Por otro lado, la dependencia temporal es incorporada en el modelo de forma muy elegante a través de una cópula gaussiana, la cual depende de una matriz de correlación cuya estructura es dada por un proceso ARMA(p, q).

Respecto del coeficiente β_{1x} se tiene que dado exp(−1,35)=0.26, ante el incremento de una unidad en el logaritmo del índice del PBI, el odds de la tasa de desempleo se reduce en 74 %; respecto del coeficiente β_{4x} se tiene que dado exp(0,11)=1.12, el odds de la tasa de desempleo aumenta en 12 % por cada año adicional, esta es una tendencia que se observó desde el año 2015. Asimismo, la tasa de desempleo presenta dependencia temporal con sus valores pasados medido a través del proceso autorregresivo de medias móviles ARMA$(1, 1)$. Finalmente, se estimó la tasa de desempleo y los resultados del ajuste del modelo fueron muy satisfactorios. Los códigos de la aplicación se presentan en el anexo B.

Bibliografía

Bayer, F. M., Cintra, R. J. y Cribari-Neto, F. (2018). Beta seasonal autoregressive moving average models, *Journal of Statistical Computation and Simulation* **88**: 2961–2981.

Becerra, O. y Melo, L. F. (2008). Medidas de riesgo financiero usando cópulas: teoría y aplicaciones, *Borradores de Economia* .

Calsaverini, R. S. y Vicente, R. (2009). An information-theoretic approach to statistical dependence: Copula information, *EPL (Europhysics Letters)* **88**(6): 68003.

Chen, X. y Fan, Y. (2006). Estimation of copula-based semiparametric time series models, *Journal of Econometrics* **130**(2): 307–335.

Da-Silva, C. Q. y Migon, H. S. (2011). Hierarchical dynamic beta model, *REVSTAT – Statistical Journal* **14**(1): 49–73.

Da-Silva, C. Q., Migon, H. S. y Correia, L. T. (2011). Dynamic Bayesian beta models, *Computational Statistics & Data Analysis* **55**(6): 2074–2089.

Embrechts, P. (2009). Copulas: A Personal View, *Journal of Risk and Insurance* **76**(3): 639–650.

Ferrari, S. y Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions, *Journal of Applied Statistics* **31**(7): 799–815.

Guolo, A. y Varin, C. (2014). Beta regression for time series analysis of bounded data, with application to Canada Google Flu Trends.

Jara, A., Nieto-Barajas, L. E. y Quintana, F. (2013). A time series model for responses on the unit interval, *Bayesian Analysis*. (3): 723–740.

Joe, H. (1997). *Multivariate models and dependence concepts*, Chapman & Hall.

Joe, H. (2014). *Dependence modeling with copulas*, Chapman & Hall/CRC, Boca Raton, FL.

Kieschnick, R. y McCullough, B. D. (2003). Regression analysis of variates observed on (0, 1): percentages, proportions and fractions, *Statistical Modelling* **3**(3): 193–213.

Kurowicka, D. y Cooke, R. (2006). *Uncertainty Analysis with High Dimensional Dependence Modelling*, Wiley Series in Probability and Statistics, John Wiley & Sons, Ltd, Chichester, UK.

Lennon, H. (2016). Gaussian copula modelling for integer-valued time series, *Technical report*.

Loaiza Maya, R., Smith, M. y Maneesoonthorn, W. (2017). Time series copulas for heteroskedastic data, *Journal of Applied Econometrics* **33**.

Masarotto, G. y Varin, C. (2012). Gaussian copula marginal regression, *Electronic Journal of Statistics* **6**(0): 1517–1549.

Masarotto, G. y Varin, C. (2017). Gaussian Copula Regression in R, *Journal of Statistical Software* **77**(8): 1–26.

Nelsen, R. B. (2006). *An introduction to copulas*, Springer.

Parsa, R. A. y Klugman, S. A. (2010). Copula Regression, *Technical report*, Variance Advancing the Science of Risk.

Patton, A. J. (2009). Copula–Based Models for Financial Time Series, *Handbook of Financial Time Series*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 767–785.

Patton, A. J. (2012). A review of copula models for economic time series, *Journal of Multivariate Analysis* **110**: 4–18.

Renard, B. y Lang, M. (2007). Use of a Gaussian copula for multivariate extreme value analysis: Some case studies in hydrology, *Advances in Water Resources* **30**(4): 897–912.

Rocha, A. y Cribari-Neto, F. (2009). Beta autoregressive moving average models, *TEST* **18**: 529–545.

Shumway, R. H. y Stoffer, D. S. (2011). *Time Series Analysis and Its Applications*, Springer Texts in Statistics, Springer New York, New York, NY.

Simas, A., Barreto-Souza, W. y Rocha, A. (2010). Improved estimators for a general class of beta regression models, *Computational Statistics and Data Analysis* **54**(2): 348–366.

Song, P. X.-K. (2000). Multivariate Dispersion Models Generated from Gaussian Copula.

Apéndice A

Anexo A

Resultados adicionales de la aplicación a la tasa de desempleo en Lima Metropolitana

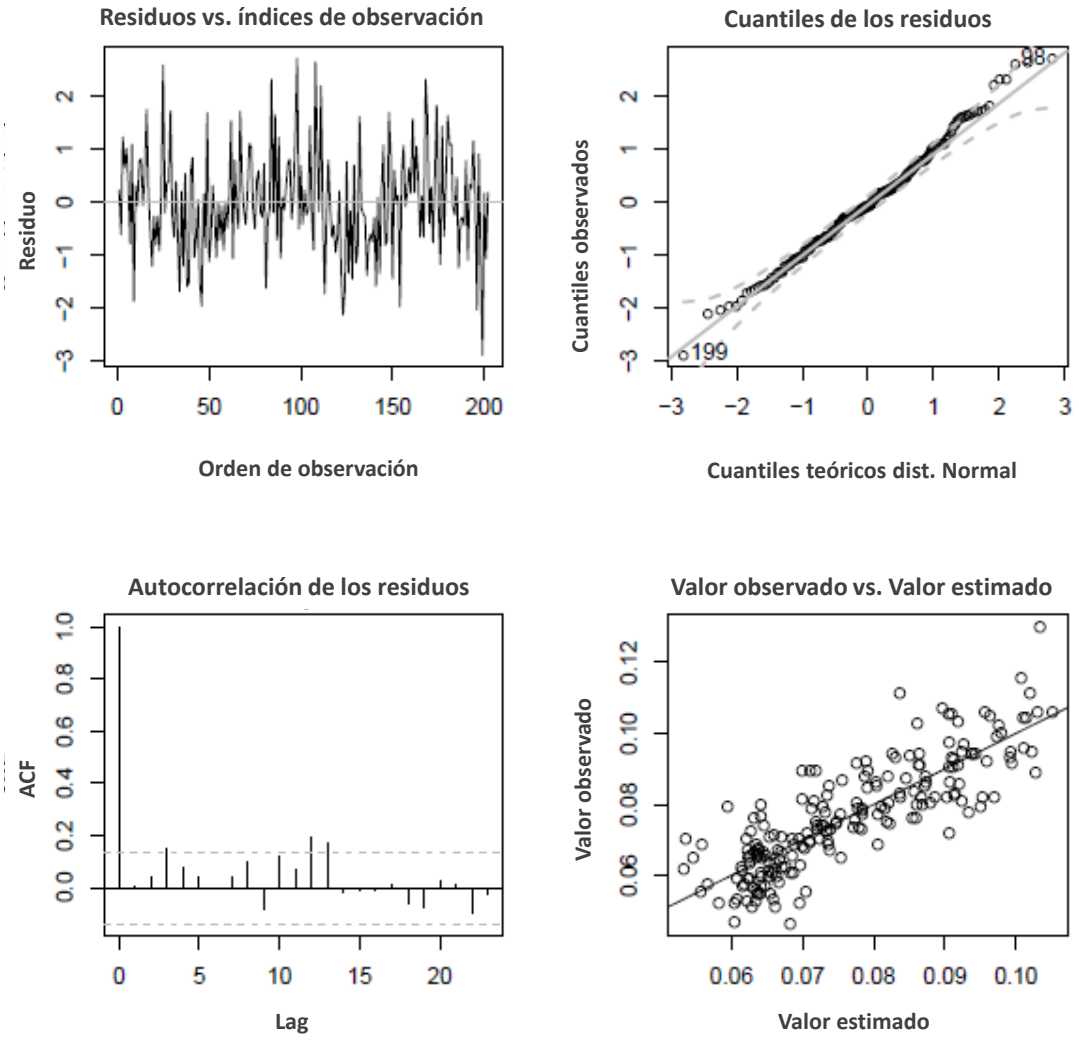


Figura S.1: Análisis de residuos Regresión beta y modelo ARMA (0,1).

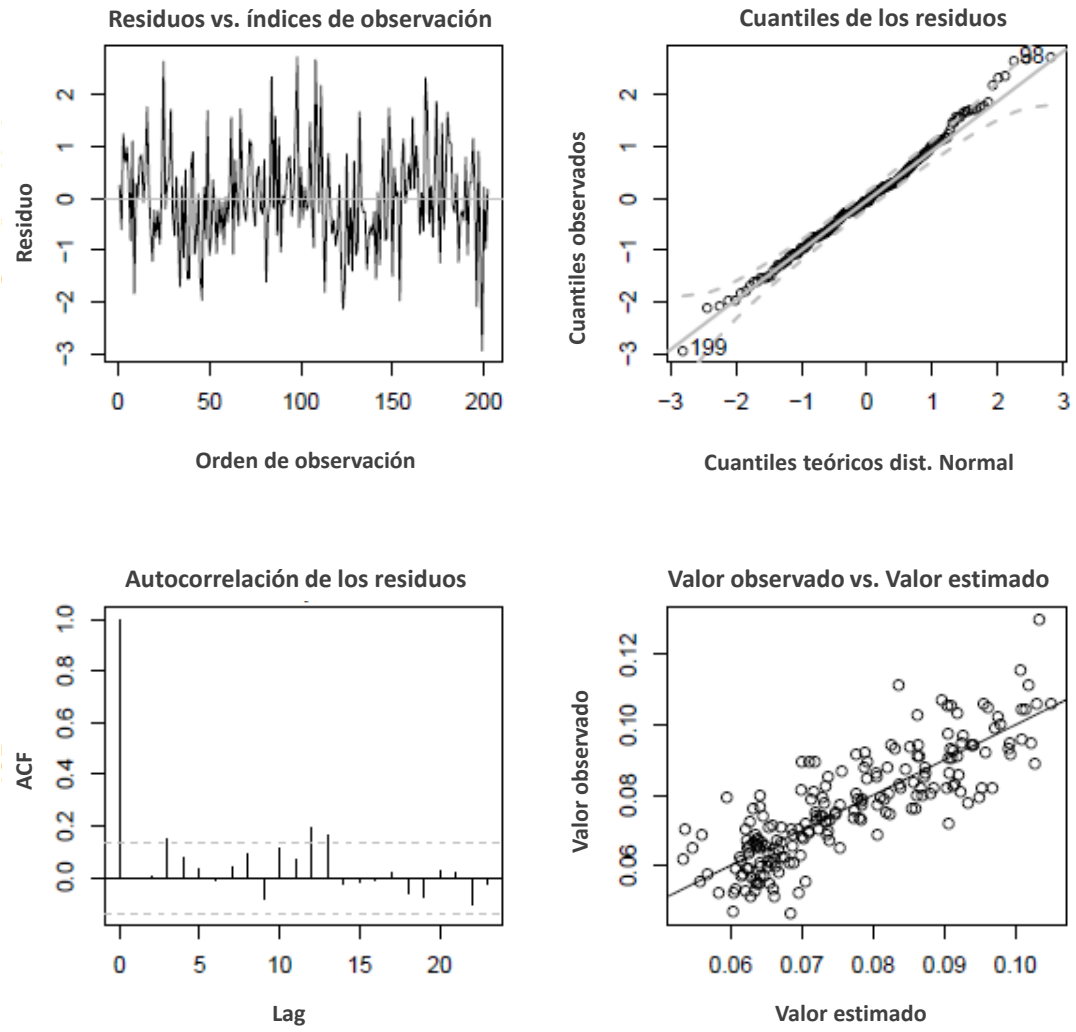


Figura S.2: Análisis de residuos Regresión beta y modelo ARMA (1,0).

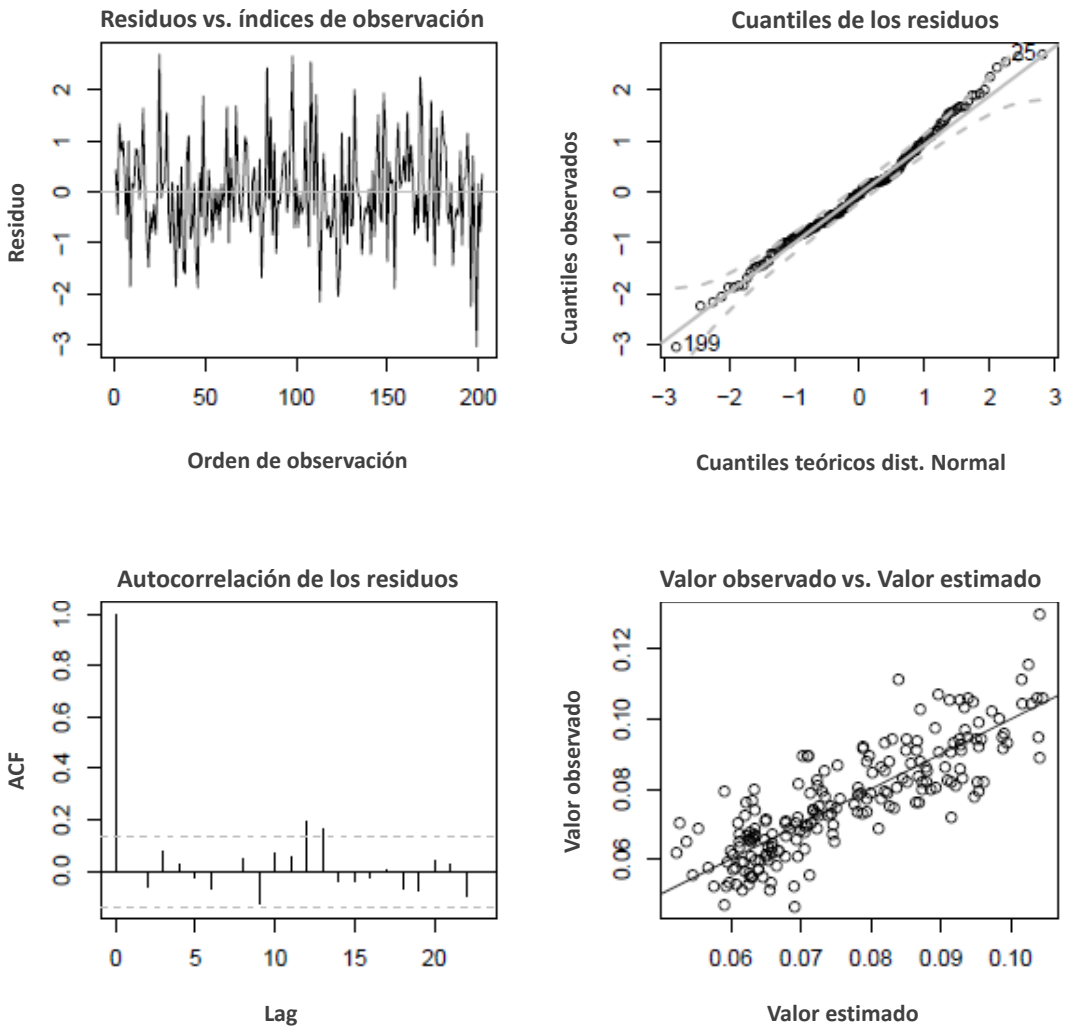


Figura S.3: Análisis de residuos Regresión beta y modelo ARMA (1,2).

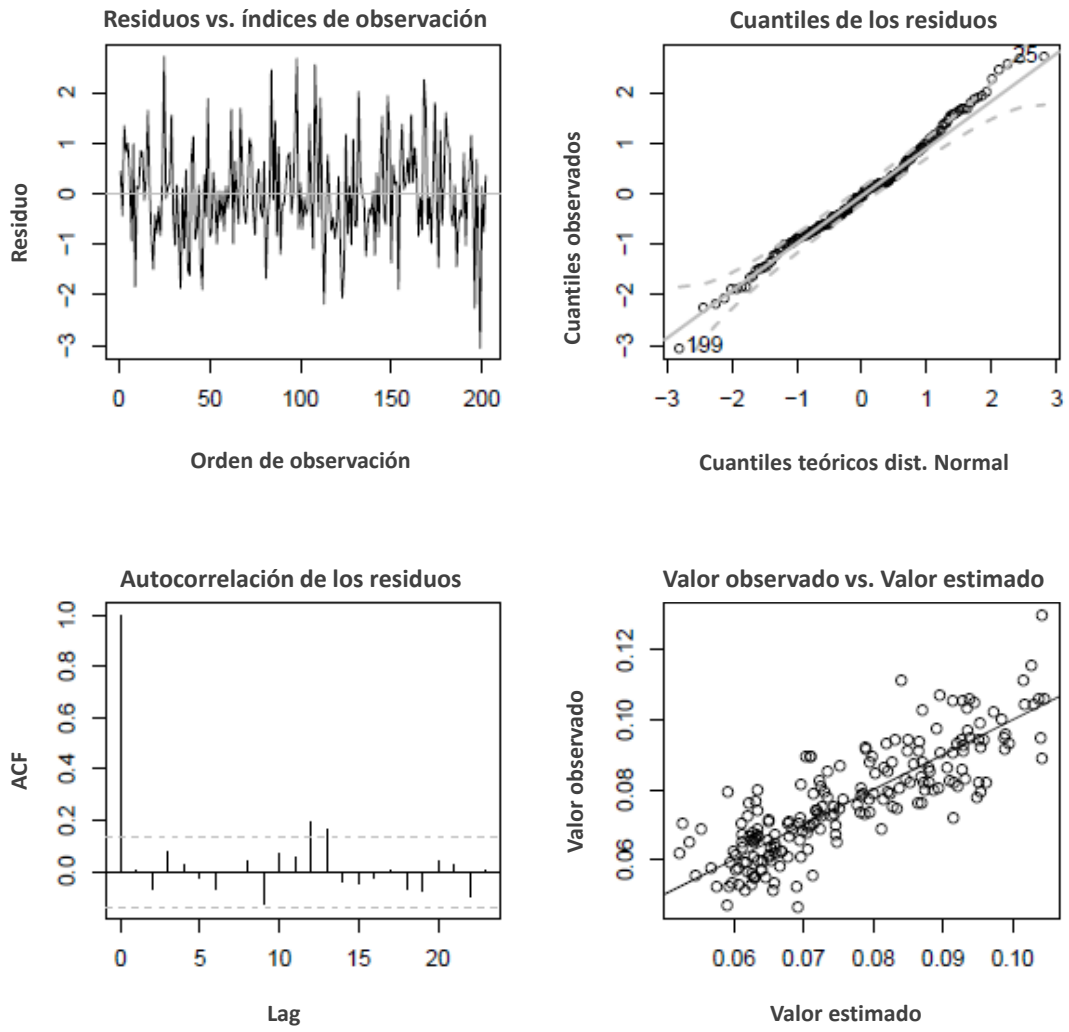


Figura S.4: Análisis de residuos Regresión beta y modelo ARMA (2,1).

Cuadro A.1: Resultados de la aplicación a datos de desempleo: Estimación (EMV) y errores estándar para cada modelo ajustado.

	Parámetros	Estimación	Error estándar
Media	β_{0x}	3.42***	0.83
	β_{1x}	-1.39***	0.18
	β_{2x}	0.08***	0.03
	β_{3x}	0.32**	0.08
	β_{4x}	0.09***	0.01
Precisión	β_{0z}	6.89***	0.10
Error: ARMA$(0,1)$	ϕ_1	0.18***	0.07
Media	β_{0x}	3.37***	0.83
	β_{1x}	-1.38***	0.18
	β_{2x}	0.08**	0.03
	β_{3x}	0.33***	0.09
	β_{4x}	0.09***	0.01
Precisión	β_{0z}	6.89***	0.10
Error: ARMA$(1,0)$	λ_1	0.19**	0.07
Media	β_{0x}	3.09***	0.93
	β_{1x}	-1.34***	0.18
	β_{2x}	0.10*	0.05
	β_{3x}	0.43**	0.16
	β_{4x}	0.10***	0.02
Precisión	β_{0z}	6.88***	0.12
Error:	ϕ_1	0.91***	0.05
ARMA$(1,2)$	λ_1	-0.74***	0.10
	λ_2	-0.06	0.08
Media	β_{0x}	3.10***	0.90
	β_{1x}	-1.35***	0.18
	β_{2x}	0.10*	0.05
	β_{3x}	0.44**	0.16
	β_{4x}	0.10***	0.02
Precisión	β_{0z}	6.88***	0.12
Error:	ϕ_1	0.97***	0.10
ARMA $(2,1)$	ϕ_2	-0.05	0.07
	λ_1	-0.81***	0.08
Significancia: '***' 0.001 '**' 0.01 '*' 0.05 '·' 0.1 ' ' 1			

Apéndice B

Anexo B

```
##############################
## aplicación
##############################
install.packages( "gcmr" )
install.packages( "betareg" )
install.packages( "tseries" )
install.packages( "forecast" )
install.packages( "LMtest" )
install.packages( "zoo" )
install.packages("strucchange")
install.packages("ggplot2")
library(strucchange)
library(gcmr)
library(betareg)
library(tseries)
library(forecast)
library(lmtest)
library(zoo)
library(ggplot2)
library(dplyr)
library(tseries)



setwd('/home/Downloads/')
dir.save <- getwd()
datos <-  read.csv(paste(dir.save,"/","desempleo2.csv",sep=""), sep=",")

#Preparación de los datos
datos$tasa <- datos$desempleo/100
plot(datos$tasa)
```

```
lines(datos$tasa)

#Tiempo estandarizado

trend <- 1:NROW(datos$tasa)
trend <- scale(trend)

#Histograma

datos$tasa=as.numeric(datos$tasa)

ggplot(datos,aes(x=tasa)) +
  geom_histogram(aes(y=..density..),binwidth = 0.006,color="blue",fill="lightblue") +
  geom_density(alpha=.2, color="black",fill="white")+theme_classic()

#Gráfico Serie de tiempo

serie=datos$tasa
serie=ts(serie,frequency =12, start=c(2003,1))
plot(serie,ylab = "Tasa ", xlab = "Tiempo" )
lines(serie)

#Modelo Regresión beta para analizar errores

modbeta1=betareg(datos$tasa ~ log(datos$pbi))
res=residuals(modbeta1)

#Gráficos de los residuos
par(mfrow = c(3, 2))
plot(residuals(modbeta1), xlab="Time",ylab="Residuals",main="Residual")
lines(residuals(modbeta1))
qqnorm(residuals(modbeta1),cex.main=1)
qqline(residuals(modbeta1))
acf(residuals(modbeta1),main="ACF",cex.main=0.5,lag=20)
pacf(residuals(modbeta1),main="PACF",cex.main=1,lag=20)
hist(residuals(modbeta1), main="Histogram of residual",cex.main=1)
#
adf.test(res)


#Regresión beta+cópulas
```

```
#2. Modelo usando cópulas

datos$time <- 1:NROW(datos$tasa)

datos$y2=(datos$year)^2

#Modelos
peru.xmas1 <- gcmr( datos$tasa ~log(datos$pbi)+
sin(2*pi/365*datos$time)+cos(2*pi/365*datos$time) +
datos$year | 1,  marginal = beta.marg, cormat =
                      arma.cormat( p=1 ,q=1 ) )
summary(peru.xmas1)

peru.xmas2 <- gcmr( datos$tasa ~log(datos$pbi)+
sin(2*pi/365*datos$time)+cos(2*pi/365*datos$time) +
datos$year| 1,  marginal = beta.marg, cormat =
                      arma.cormat( p=0 ,q=1 ) )
summary(peru.xmas2)


peru.xmas3 <- gcmr( datos$tasa ~log(datos$pbi)+
sin(2*pi/365*datos$time)+cos(2*pi/365*datos$time) +
datos$year| 1,  marginal = beta.marg, cormat =
                      arma.cormat( p=1 ,q=0 ) )
summary(peru.xmas3)


peru.xmas4 <- gcmr( datos$tasa ~log(datos$pbi)+
sin(2*pi/365*datos$time)+cos(2*pi/365*datos$time) +
datos$year| 1, marginal = beta.marg, cormat =
                      arma.cormat( p=1 ,q=2 ) )
summary(peru.xmas4)

peru.xmas5 <- gcmr( datos$tasa ~log(datos$pbi)+
sin(2*pi/365*datos$time)+cos(2*pi/365*datos$time) +
datos$year| 1, marginal = beta.marg, cormat =
                      arma.cormat( p=2 ,q=1 ) )
summary(peru.xmas5)



#Gráficos
```

```
pdf("pred.pdf")
serie=datos$tasa
serie=ts(serie,frequency =12, start=c(2003,1))
plot(serie,ylab = "Tasa ", xlab = "Tiempo",  )
serie1=fitted(peru.xmas1)
serie1=ts(serie1,frequency =12, start=c(2003,1))
lines(serie1,col="red")
legend("topright", bty="n", lty=c(1,1), col=c("black","red"),
       legend=c("Tasa Real ", "Tasa Estimada "))
dev.off()


#Gráficos de residuos
pdf("rplotarma11.pdf")
par(mfrow = c(2, 2))
plot(peru.xmas1,1 )
#qqnorm(residuals(peru.xmas1),cex.main=1)
#qqline(residuals(peru.xmas1))
plot(peru.xmas1,3)
plot(peru.xmas1,  5)
plot(fitted(peru.xmas1), datos$tasa,xlab="valor
estimado",ylab="valor observado",
cex.main=1,main="Valor estimado vs. valor predicho")
abline(0,1)
dev.off()
#



pdf("rplotarma01.pdf")
par(mfrow = c(2, 2))
plot(peru.xmas2,1 )
#qqnorm(residuals(peru.xmas1),cex.main=1)
#qqline(residuals(peru.xmas1))
plot(peru.xmas2,3)
plot(peru.xmas2,  5)
plot(fitted(peru.xmas2), datos$tasa,xlab="valor
estimado",ylab="valor observado",
cex.main=1,main="Valor estimado vs. valor predicho")
abline(0,1)
#
dev.off()
```

```
pdf("rplotarma10.pdf")
par(mfrow = c(2, 2))
plot(peru.xmas3,1 )
#qqnorm(residuals(peru.xmas1),cex.main=1)
#qqline(residuals(peru.xmas1))
plot(peru.xmas3,3)
plot(peru.xmas3,  5)
plot(fitted(peru.xmas3), datos$tasa,xlab="valor
estimado",ylab="valor observado",
cex.main=1,main="Valor estimado vs. valor predicho")
abline(0,1)
#
dev.off()



pdf("rplotarma12.pdf")
par(mfrow = c(2, 2))
plot(peru.xmas4,1 )
#qqnorm(residuals(peru.xmas1),cex.main=1)
#qqline(residuals(peru.xmas1))
plot(peru.xmas4,3)
plot(peru.xmas4,  5)
plot(fitted(peru.xmas4), datos$tasa,xlab="valor
estimado",ylab="valor observado",
cex.main=1,main="Valor estimado vs. valor predicho")
abline(0,1)
#
dev.off()



pdf("rplotarma21.pdf")
par(mfrow = c(2, 2))
plot(peru.xmas5,1 )
#qqnorm(residuals(peru.xmas5),cex.main=1)
#qqline(residuals(peru.xmas1))
plot(peru.xmas5,3)
plot(peru.xmas5,  5)
```

```
plot(fitted(peru.xmas5), datos$tasa,xlab="valor
estimado",ylab="valor observado",
cex.main=1,main="Valor estimado vs. valor predicho")
abline(0,1)
#
dev.off()
```