Modelamiento de series de tiempo de conteo: Un caso de estudio para la venta de aves sacrificadas

Juan Diego Sepulveda Ballesteros

Universidad Industrial de Santander
Facultad de Ciencias básicas
Matemáticas
Bucaramanga
2023

Modelamiento de series de tiempo de conteo: Un caso de estudio para la venta de aves sacrificadas

Juan Diego Sepulveda Ballesteros

Trabajo de grado para optar al titulo de Matemático

Asesor Técnico

Asesor Metodológico
Andrés Sebastián Ríos Gutiérrez

Universidad Industrial de Santander
Facultad de Ciencias básicas
Matemáticas
Bucaramanga
2023

Tabla de Contenido

Resumen

Título: Modelamiento de series de tiempo de conteo: Un caso de estudio para la venta de aves sacrificadas.
Autor: Juan Diego Sepulveda Ballesteros.
Palabras clave: Series de tiempo, R, Pronósticos, Estadística.
Descripción:
Este estudio presenta un modelo de series de tiempo de conteo para analizar las ventas diarias de aves sacrificadas. Se emplearon las distribuciones de Poisson y binomial negativa para modelar el número de aves vendidas, considerando la estacionalidad y las tendencias presentes en los datos. Los resultados mostraron que la distribución binomial negativa proporcionó un mejor ajuste al conjunto de datos, lo que sugiere una mayor variabilidad en las ventas de lo esperado bajo un modelo de Poisson. Adicionalmente, se identificaron patrones estacionales y tendencias de decrecimiento en las ventas a lo largo del tiempo. Las implicaciones de este estudio son relevantes para la toma de decisiones en el negocio, ya que permite realizar pronósticos más precisos de la demanda, optimizar los niveles de inventario y ajustar las estrategias de marketing en función de las fluctuaciones estacionales.

Abstract

Title: Time series counting modeling: A case study for the sale of slaughtered birds.
Author: Juan Diego Sepulveda Ballesteros.
Keywords: Time series, R, Forecasting, Statistics.
Description:
This study presents a count time series model to analyze daily sales of slaughtered birds. Poisson and negative binomial distributions were used to model the number of birds sold, considering the seasonality and trends present in the data. The results showed that the negative binomial distribution provided a better fit to the data set, suggesting greater variability in sales than expected under a Poisson model. Additionally, seasonal patterns and declining trends in sales over time were identified. The implications of this study are relevant to business decision making, as it allows for more accurate demand forecasts, optimization of inventory levels, and adjustment of marketing strategies based on seasonal fluctuations.

Introducción

A lo largo de la historia el ser humano ha tratado de comprender mejor el mundo a su alrededor, saber el sentido de las cosas a partir del conocimiento que esta a su alcance usando la razón, lógica o abstracción para en esta búsqueda encontrar la verdad. En la antigüedad la filosofía encuentra relación con las matemáticas con pensamientos como el de Pitágoras al tener que los números son el elemento del que se componen todas las cosas o el pensamiento platonico donde en el mundo de las ideas, las matemáticas son vistas como una entidad eterna y perfecta que constituye la base de todo lo que se percibe en el mundo físico. Esta área de conocimiento esta implícita en casi todo porque es una disciplina que es esencial para la comprensión del mundo que nos rodea. Una de las ramas que conforman esta disciplina es la estadística que consiste en recopilar, organizar, analizar e interpretar datos. Esta área es empleada para obtener información, identificar tendencias y en particular para predecir resultados futuros. Los datos pueden ser tratados de distintas maneras, depende de las propiedades que estos satisfagan. Por ejemplo, las series de tiempo de conteo que son usadas para analizar datos de un periodo de tiempo que mediante un modelado permite hacer predicciones, describir tendencias o identificar patrones, es usada en general si se desea analizar las ventas en un establecimiento, el comportamiento de una enfermedad o identificar tendencias en los precios de algún producto. Las series de tiempo de conteo nacen en 1662 con “Natural and Political Observations Mentioned in a following Index, and made upon the Bills of Mortality"de John Grant que en su libro utiliza datos estadísticos para estudiar la población de la ciudad de Londres analizando los casos de mortalidad de esta ciudad determinando el crecimiento de la población, la mortalidad y la esperanza de vida (Sutherland, 1963). Continuando con los estudios que han influido en las series de tiempo de conteo estan los relacionados con Sir Francis Galton quien en 1985 con su articulo “Philosophical Transactions of the Royal Society"dio un pimer acercamiento a lo que se conoce como linea de regresión con su sistema de elipses concéntricas que más adelante con un estudio de 1988 define que la pendiente de la recta como medida de correlación. Como último desarrollo importante en esta área se tiene el trabajo realizado por George Box y Gwilym Jenkins acerca de los modelos ARIMA trayendo nuevos modelos y técnicas para el estudio de series de tiempo (Box, Jenkins, y Bacon, 1967). Además, con la llegada de las computadoras que permiten mediante software realizar el análisis a gran cantidad de estos datos en poco tiempo.

Las bondades que ofrece la estadística y en particular las series de tiempo de conteo como una herramienta para considerar la toma de decisiones a partir de los datos recolectados del pasado, permite inferir mediante los modelados que se acercan al comportamiento de los datos porque resulta práctico en ocasiones tener un respaldo al tomar una decisión numéricamente hablando. Teniendo en cuenta la utilidad que tienen las series de tiempo para identificar tendencias en los datos, en particular en un pequeño negocio que se trabaja de manera anónima por sugerencia de los administradores, este negocio trabaja en tres municipios y se hace referencia a estos como

“Municipio 1"donde son distribuidores de los locales en la plaza de mercado, restaurantes, además de un pequeño local, el “Municipio 2çuenta con distribución para los locales de la plaza de mercado y clientes de restaurantes en la carretera de ese trayecto, el “Municipio 3üna ruta que se realiza los días jueves y sábados varias horas más allá del “Municipio 2".El negocio se dedica a la distribución y venta de aves, en su defecto gallinas y por su movimiento se requiere tener una herramienta que permita tener control y saber si se esta sacrificando la cantidad adecuada de aves en el proceso de tal forma que se satisfaga la demanda de producto por parte de los comerciantes en las plazas como los compradores ocasionales del local con el que cuenta este pequeño negocio sin dejar apenás aves a congelar. En este negocio es necesario ir con un camión a cargar las aves en una granja avícola ubicadas a almenos tres o más horas en el camión y ocasionalmente en granjas más cercanas siendo de una hora o dos de trayecto. Es necesario tener clara la cantidad de aves a sacrificar dado que se debe realizar una orden o consignación previamente para que estas se puedan transportar de los galpones de la granja avícola hacía la planta de sacrificio, el funcionamiento más a profundidad del negocio se encuentra el cápitulo 3.

Dentro del negocio se cuenta con datos recopilados a lo largo de un periodo de 1087 días, con el propósito de analizar el comportamiento, los patrones que este pueda tener se plantea una posible solución mediante las series de tiempo detallando las variables que influyen en el comportamiento de cada día de la semana e implementando un modelo con y sin estas, comparando los cálculos, gráficas y las predicciones generadas de cada modelado con el fin de tomar el más adecuado según los parametros y distribuciones apoyado de la herramienta `RStudio`. En el cápitulo **??** se aborda la terminología, las razones y la explicación de los conceptos necesarios para un mejor entendimiento del documento. El cápitulo 2 mostrando el procedimiento que se realiza en caso de tenerse datos faltantes, las funciones del paquete que se emplea para este documento.

1. Planteamiento del problema

1.1. Descripción del problema

El funcionamiento en este negocio consiste en la comercialización de productos cárnicos de gallinas que han sido procesadas para el consumo humano. La carne de las aves sacrificadas es la base para varios platos como lo son: sancocho de gallina, picadas, gallina sudada, en salsa de champiñones, asada o guisada, esto ya queda como decisión del cliente. Las aves son sacrificadas y se encuentran disponibles según las necesidades de vendedores, restaurantes o personas que las necesiten. Los clientes tienen diferentes maneras de pedir las aves sacrificadas, variando algunos factores como son el uso de colorantes para asemejarse a aves criollas, dejar las tripas, quitarle las patas, dejar el pescuezo con sangre, esto ya es decisión del cliente y es algo que no se tendrá en cuenta dado que se busca determinar la cantidad de aves a sacrificar en total y no por municipios o clientes. Para llegar al producto final del ave sacrificada inicialmente se debe considerar la cantidad que es necesaria cargar para el día siguiente, este cargue se realiza en alguna granja que por lo general se encuentra a 90 o más kilómetros de la planta de sacrificio, dichas estimaciones se dan a partir de la experiencia de los administradores en este campo. Sin embargo, puede suceder que dichas estimaciones terminen con aves faltantes para alguna parte del negocio que son: el local en la plaza de mercado, los pedidos en las plazas de los municipios en los que labora y los restaurantes.

Uno de los problemas esta en determinar cuantas aves se requieren para que de esta manera el administrador realice el pago de una factura con la cantidad de aves requeridas. Una vez hecha esta consignación el administrador tiene permitido enviar el camión a cargar las aves con la seguridad que estas son entregadas, ya que sin factura en caso de llegar a la granja pierde el turno frente a otro camión que si tenga realizado el pago y por otro lado cuando el avicultor tiene la factura, este programa con los demás ayudantes de la granja para entregar estas aves en horas de la mañana o mediodía. La cantidad de gallinas a sacrificar no se puede tener con certeza el día anterior o en el instante que el camión va camino a la granja porque los clientes aún no han finalizado su jornada y puede que queden con aves sin vender o hagan pedidos más grandes de lo esperado. Quizá se puede ver como solución sacrificar más aves de las requeridas pero puede suceder que al finalizar la jornada en el local no se vendan el total de aves, esto termina con las aves en un congelador donde caben a lo sumo 50 aves, aumentando gastos al emplearse esta máquina junto al hecho de que la frescura es algo que busca siempre el cliente, desencadenando una disminución de precio drástica. Por otro lado si las gallinas son insuficientes los clientes en los 3 municipios tienen cierta prioridad dado que son pequeños locales en las plazas de mercado y el local que tiene el negocio termina sin aves para la venta o incluso los restaurantes. Es decir, los clientes del local de este negocio o de algunos restaurantes en caso de no alcanzar la cantidad suficiente deben ir a otro local a comprar que es algo que no desea el establecimiento.

Es así que con el fin de tener una estimación guía basada en los rendimientos de tres años anteriores se modelan las cantidades necesarias a sacrificar, teniendo en cuenta los valores que se llegan a extraer de los cuadernos donde anotan los pedidos de cada cliente junto con las ventas del local que tiene el negocio.

1.2. Formulación del problema

Una empresa avícola enfrenta la incertidumbre de determinar la cantidad óptima de aves a sacrificar diariamente, debido a la naturaleza aleatoria y basada en la experiencia de los empleados de esta decisión. Se busca desarrollar un modelo cuantitativo que permita predecir con mayor precisión la demanda diaria de carne de ave, reduciendo así la variabilidad en la producción y optimizando los recursos.

1.3. Justificación

Algunos de los beneficios que da el proyecto modelado con series de tiempo en una Empresa Avícola son los siguientes:

- Optimización de recursos: Reducción de costos asociados a la producción, almacenamiento y desperdicio.

- Mejora de la toma de decisiones: Mayor precisión en la planificación y reducción de la incertidumbre.

- Aumento de la rentabilidad: Mayor eficiencia en la producción y mejor gestión de inventario.

- Satisfacción del cliente: Mayor disponibilidad y calidad del producto.

- Base para futuros análisis: Permite identificar oportunidades de negocio y optimizar procesos.

- Mejora de la imagen de marca: Refuerza la percepción de la empresa como eficiente y confiable.

- En pocas palabras, este proyecto permitirá a la empresa tomar decisiones más informadas, reducir costos, mejorar la eficiencia y aumentar la satisfacción del cliente, lo que se traduce en un mayor crecimiento y competitividad en el mercado.

2. Objetivos

2.1. Objetivo General

Modelizar, utilizando series de tiempo de conteo, el número diario de gallinas vendidas para establecer el valor medio diario de individuos a sacrificar.

2.2. Objetivos específicos

- Modelar el número de gallinas sacrificadas usando un modelo binomial negativo.
- Modelar el número de gallinas sacrificadas usando un modelo de Poisson.

3. Marco Teórico

3.1. Conceptos básicos de probabilidad

Para los modelos lineales generalizados para series de tiempo de conteo se se emplean distribuciones como Poisson y binomial negativa. A continuación se definen

Definición 3.1. (Blanco Castañeda, 2013) **(Distribución Poisson)** Se dice que una variable aleatoria X tiene distribución Poisson de parámetro $\lambda > 0$, si su función de densidad está dada por:

$$f_X(x) = \begin{cases} e^{-\lambda}\frac{\lambda^x}{x!} & \text{si } x = 0, 1, \cdots \\ \\ 0 & \text{en otro caso} \end{cases}$$

Notación: Sea X una variable aleatoria. Se escribe $X \overset{d}{=} \mathscr{P}(\lambda)$ para indicar que X tiene una distribución Poisson de parámetro λ.

Definición 3.2. (Blanco Castañeda, 2013) **(Distribución binomial negativa)** Se dice que una variable aleatoria X tiene distribución binomial negativa de parámetros k y p, si su función densidad está dada por:

$$f_X(x) = \begin{cases} \binom{x-1}{k-1} p^k (1-p)^{x-k} & \text{si } x = k, k+1, \cdots \\ \\ 0 & \text{en otro caso} \end{cases}$$

En el caso especial $k = 1$, se dice que la variable aleatoria tiene distribución geométrica de parámetro p.

Notación: Las expresiones $X \overset{d}{=} \mathscr{B}_N(k, p)$ indica que X tiene distribución binomial negativa de parámetros k y p.

En los modelos se utiliza la notación '$E(Y_t|\mathscr{F}_t)$', donde E denota la esperanza de $X_t|\mathscr{F}_t$. Para ciertos requisitos en las distribuciones es necesario la noción de varianza. Además para el tratamiento de las covariables se emplea el coeficiente de correlación que describe la dependencia entre las variables aleatorias X e Y. Estos conceptos se definen a continuación:

Definición 3.3. (Blanco Castañeda, 2013) Sea X una variable aleatoria real definida sobre el espacio de probabilidad $(\Omega, \mathfrak{J}, P)$.

(i) Si X es una variable aleatoria discreta con valores $x_1, x_2, \ldots$, se dice que tiene esperanza si $\sum_{i=1}^{+\infty} |x_i| P(X = x_i) < +\infty$, caso en el cual la **esperanza** $E(X)$ se define como

$$E(X) = \sum_{i=1}^{+\infty} x_i P(X = x_i)$$

(ii) Si $E(X^2)$ y $E(X)$ existen se define la **varianza** como

$$V(X) = E\left([X - E(X)]^2\right) = E(X^2) - [E(X)]^2$$

(iii) Sean X y Y variables aleatorias sobre $(\Omega, \mathfrak{J}, P)$ tales que $E(X^2) < +\infty$ y $E(Y^2) < +\infty$. Se define la **covarianza** de X y Y como

$$Cov(X,Y) = E\left([X - E(X)][Y - E(Y)]\right) = E(XY) - E(X)E(Y)$$

(iv) Sean X y Y variables aleatorias sobre $(\Omega, \mathfrak{J}, P)$ tales que $E(X^2) < +\infty$ y $E(Y^2) < +\infty$. Se define el **coeficiente de correlación** de X y Y como

$$\rho(X,Y) := \frac{Cov(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}}.$$

Proposición 3.4.(Blanco Castañeda, 2013) Sea X una variable aleatoria con distribución Poisson de parámetro λ. Entonces:
1. $E(X) = \lambda$.
2. $V(X) = \lambda$.

Proposición 3.5. (Blanco Castañeda, 2013) Sea X una variable aleatoria con distribución binomial negativa de parámetros k y p. Entonces:
(i) $E(X) = \frac{k(1-p)}{p}$
(ii) $V(X) = \frac{k(1-p)}{p^2}$

Dado que $E(X_t|\mathscr{F}_t)$ es una esperanza condicional dada la σ-álgebra generada por la variable aleatoria X_t, entonces se procede a definir (1) σ-álgebra generada por una variable aleatoria y (2) esperanza condicional dada una σ-álgebra.

Definición 3.6.(Jacod y Protter, 2004) Sean $(\Omega, \mathfrak{J}, P)$ un espacio de probabilidad $\left(\tilde{\Omega}, \tilde{\mathfrak{J}}\right)$ un espacio medible y $X : \Omega \to \tilde{\Omega}$ una variable aleatoria. Se define la **σ-álgebra generada por la variable aleatoria** X como la $\sigma-$álgebra definida por:

$$\sigma(X) := \bigcap_{\mathfrak{G} \text{ es una } \sigma-\text{álgebra}} \left\{\mathfrak{G} : X^{-1}(B) \in \mathfrak{G} \text{ para todo } B \in \tilde{\mathfrak{J}}\right\}$$

A continuación se define la independencia entre variables aleatorias.

Definición 3.7. (Jacod y Protter, 2004) Sean los espacios medibles $\left(\tilde{\Omega}_1, \tilde{\mathfrak{J}}_1\right), \ldots, \left(\tilde{\Omega}_n, \tilde{\mathfrak{J}}_n\right)$ sobre los cuáles están definidas las variables aleatorias $X_1, \ldots, X_n$, respectivamente. Se dice que $X_1, \ldots, X_n$ son **independientes**, si y sólo si, las σ-álgebras $\sigma(X_1), \sigma(X_2), \ldots, \sigma(X_n)$ son independientes.

Definición 3.8. (Blanco Castañeda, 2013) Sean $X_1, \ldots, X_n$ variables aleatorias reales definidas

todas sobre un espacio de probabilidad $(\Omega, \mathfrak{J}, P)$. La variable aleatoria definida por

$$\begin{aligned} \mathbf{X} : \Omega &\rightarrow \mathbb{R}^n \\ \omega &\rightarrow \mathbf{X}(\omega) = (X_1(\omega), \ldots, {}_n(\omega) \end{aligned}$$

recibe el nombre de **vector aleatorio** n-**dimensional**. La medida de probabilidad definida sobre $\mathfrak{B}(\mathbb{R}^n)$ por

$$P_X(B) := P(X \in B) \text{para todo } B \in \mathfrak{B}(\mathbb{R}^n)$$

se llama **distribución del vector aleatorio X**.

Para poder definir la esperanza condicional dada una variable aleatoria, primero se debe definir (1) la función de densidad de probabilidad condicional y (2) la función indicadora

Definición 3.9. (Blanco Castañeda, 2013) Sean X y Y dos variables aleatorias continuas, definidas sobre el espacio de probabilidad $(\Omega, \mathfrak{J}, P)$, con función de densidad de probabilidad conjunta f y funciones de densidad f_X y f_Y, respectivamente. Se define la **función de densidad de probabilidad condicional de X dado Y** $= y$ como

$$f_{X|Y}(x|y) := \frac{f(x,y)}{f_Y(y)} \text{ para todo } y \text{ tal que } f_Y(y) > 0.$$

Definición 3.10. ((Ríos, 2019)) Sean $(\Omega, \mathfrak{J}, P)$ un espacio de probabilidad y $B \in \mathfrak{J}$. Se define la **función indicadora** de B para todo $\omega \in \Omega$ como

$$\mathbf{1}_B(\omega) = \begin{cases} 1 & \text{si } \omega \in B \\ \\ 0 & \text{en otro caso} \end{cases}$$

El modelo de Poisson se utiliza la notación '$E(Y_t|\mathscr{F}_t)$', donde $E(Y_t|\mathscr{F}_t)$ es una esperanza condicional de Y_t dado $\mathscr{F}_t$. De esta manera, a continuación se da tal definición:

Definición 3.11. ((Ríos, 2019)) Sean X una variable aleatoria real sobre el espacio de probabilidad $(\Omega, \mathfrak{J}, P)$,

(**i**) Si $B \in \mathfrak{J}$ con $P(B) > 0$ y $\mathbb{E}(X\mathbf{1}_B) < +\infty$, se define **la esperanza condicional de** X **dado** B como

$$\mathbb{E}(X|B) := \frac{\mathbb{E}(X\mathbf{1}_B)}{P(B)}.$$

(**ii**) Si $\mathfrak{G}$ es una sub-σ-álgebra de $\mathfrak{J}$ se define la esperanza condicional de X dada $\mathfrak{G}$ como una

variable aleatoria $\mathfrak{G}$-medible denotada por $\mathbb{E}(X|\mathfrak{G})$, tal que

$$\mathbb{E}([X-\mathbb{E}(X|\mathfrak{G})]\mathbf{1}_G = 0, \text{ para todo } G \in \mathfrak{G}.$$

3.2. Conceptos básicos de procesos estocásticos

Se asume que el número de gallinas por día sacrificadas es una serie de tiempo. Como una serie de tiempo es un proceso estocástico, entonces se procede a definir este concepto a continuación
Definición 3.12. (Blanco Castañeda, 2013) Un **proceso estocástico** es una familia de variables aleatorias $\{X_t\}_{t\in T}$, definidas sobre el mismo espacio de probabilidad $(\Omega,\mathfrak{J},P)$ y con valores en un espacio medible $(S,\mathfrak{G})$. El conjunto T es conocido como el **conjunto de índices del proceso** y S como el **espacio de estados.**
Definición 3.13. (Blanco Castañeda, 2013) Sea $\{X_t\}_{t\in T}$ un proceso estocástico definido sobre un espacio de probabilidad $(\mathbb{R},\mathfrak{B}(\mathbb{R}))$ y con valores en $(S,\mathfrak{G})$. La función $(t,\omega)\to X_t(\omega)$ para un $\omega\in\Omega$ se llama la **trayectoria de** $\{X_t\}_{t\in T}$ **asociado a** ω.
La función de autocovarianza permite establecer la dependencia de la variable aleatoria presente con respecto a una variable aleatoria del pasado. A continuación, se define esta función
Definición 3.14. (Brockwell y Davis, 2009) Si $\{X_t, t\in T\}$ es un proceso tal que $V(X_t)<\infty$ para todo $t\in T$, entonces la **la función de autocovarianza** $\gamma_X(\cdot,\cdot)$ de $\{X_t\}$ esta dada por

$$\gamma_X(r,s) = Cov(X_r,X_s) = E[(X_r-EX_r)(X_s-EX_s)], \qquad r,s\in T.$$

Al momento de realizarse la inferencia el programa R asume la estacionariedad que consiste en media, varianza y autocorrelación constantes a lo largo del tiempo. A continuación se define
Definición 3.15. (Brockwell y Davis, 2009) La serie de tiempo $\{X_t, t\in\mathbb{Z}\}$ con indice establecido $\mathbb{Z}=\{0,\pm1,\pm2,\dots\}$, es llamada **estacionaria** si
(i) $E|X_t|^2<\infty$ para todo $t\in\mathbb{Z}$
(ii) $E(X_t)=m$ para todo $t\in\mathbb{Z}$
(iii) $\gamma_X(r,s)=\gamma_X(r+t,s+t)$ para todo $r,s,t\in\mathbb{Z}$
Definición 3.16. (Brockwell y Davis, 2009) La serie de tiempo $\{X_t, t\in T\}$ es llamada **estrictamente estacionaria** si para las distribuciones conjuntas de $(X_{t1},\dots,X_{tk})'$ y de $(X_{t_1+h},\dots,X_{t_k+h})'$ son la misma para todos los $k\in\mathbb{Z}^+$ y para todo $t_1,\dots,t_k,h\in\mathbb{Z}$.
Otra suposición en la inferencia es la propiedad de mantener la misma media a lo largo del tiempo del tiempo o ergodicidad. A continuación se define este concepto
Definición 3.17. (Brockwell y Davis, 2009) Un proceso de series de tiempo $\{X_t, t\in\mathbb{N}\}$ es **ergódico** si para cualquier par de funciones acotadas f, g:

$$Lm_{k\to\infty}|E[f(y_t,y_{t+1},\ldots,Y_{t+a})g(y_{t+k},y_{t+k+1},\ldots,Y_{t+k+b})]| =$$

$$|E[f(y_t,y_{t+1},\ldots,Y_{t+a})]|\cdot|E[g(y_{t+k},y_{t+k+1},\ldots,Y_{t+k+b})]|$$

Para comparar las variables cualitativas que se tienen en el documento se emplea el test chi cuadrado de Pearson, a continuación su definición.

Definición 3.18. ((Anderson y Burnham, 2004)) Considere un modelo de regresión lineal donde se supone que una variable dependiente Y es una función de r variables explicativas (predictoras) $X_j(j=1,2,\ldots,r)$. Aquí se supone que los residuos ε_i de las n observaciones son independientes, normalmente distribuidos con una varianza constante σ^2, y la estructura del modelo se expresa como

$$Y_i=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_rX_r+\varepsilon_i,\quad i=1,\ldots,n.$$

Por consiguiente

$$E\left(Y_i\right)=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_rX_r,\quad i=1,\ldots,n,$$

y $E\left(Y_i\right)$ es una función lineal de parámetros $r+1$. Los residuos conceptuales,

$$\varepsilon_i=Y_i-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_rX_r)=Y_i-E\left(Y_i\right),$$

tienen la distribución de probabilidad conjunta $g(\underline{\varepsilon}\mid\underline{\theta})$, donde $\underline{\theta}$ es un vector de $K=r+2$ parámetros ($\beta_0,\beta_1,\ldots,\beta_r$ y σ). Aquí, correspondiente a la observación i se tiene el modelo

$$g\left(\varepsilon_i\mid\underline{\theta}\right)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{1}{2}\left[\frac{\varepsilon_i}{\sigma}\right]^2}.$$

La probabilidad es simplemente el producto de estas sobre las n observaciones, interpretadas como una función de los parámetros desconocidos, dados los datos, la estructura del modelo lineal y el supuesto de normalidad:

$$L(\underline{\theta}\mid\underline{x})=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{1}{2}\left[\frac{\varepsilon_j}{\sigma}\right]^2}=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n\mathrm{e}^{-\frac{1}{2}\sum_{i=1}^{n}\left[\frac{\varepsilon_j}{\sigma}\right]^2}.$$

Para definir el criterio de información de Akaike es necesaria la estimación de la máxima verosimilitud que se define a continuación.

Definición 3.19. (Casella y Berger, 2021) Recuerde que si $X_1,X_2,\ldots,X_n$ es una muestra de datos independientes e idénticamente distribuidos de una población con función de masa de probabilidad o función de densidad de probabilidad $f(x|\theta_1,\ldots,\theta_k)$, la función de probabilidad se define por

$$L(\theta|x)=L(\theta_1,\ldots,\theta_k|x_1,\ldots,x_n)=\prod_{i=1}^{n}f(x_i|\theta_1,\ldots,\theta_k)$$

Para cada punto de la muestra x, sea $\hat{\theta}$ un parámetro, vale que $L(\theta|x)$ alcanza su máximo como función de θ, con x manteniendose fijo. Una estimación de máxima verosimilitud del parámetro θ basada en una muestra X es $\theta(\hat{X})$.
Para comparar los diferentes modelos se emplea el criterio de información de Akaike, se da su respectiva definición

Definición 3.20. ((Anderson y Burnham, 2004)) " an information criterion"(AIC) se define multiplicando $\log(\mathscr{L}(\theta \mid y)) - K$por -2 ("teniendo en cuenta razones históricas") para conseguir

$$\text{AIC} = -2\log(\mathscr{L}(\hat{\theta} \mid y)) + 2K$$

Donde la expresión $\log(\mathscr{L}(\theta \mid y))$ es el valor numerico de la log-verosimilitud la probabilidad logarítmica en su punto máximo y K el número de parámetros estimables

Definición 3.21. (Gravetter, Wallnau, Forzano, y Witnauer, 2020) Considere $H_0 : \pi_j = \pi_{j0}$, $j = 1,\ldots,c$ donde $\sum_j \pi_{j0} = 1$. Cuando H_0 es verdadera, los valores esperados de $\{n_j\}$, son llamadas frecuencias esperadas, son $\mu_j = n\pi_{j0}$, $j = 1,\ldots,c$. Pearson propone el test estadístico

$$\chi^2 = \sum_j \frac{(n_j - \mu_j)^2}{\mu_j}$$

Donde grandes diferencias $\{n_j - \mu_j\}$ producen χ^2 grandes, para n fijo.

3.3. Modelos de conteo

Denotemos una serie de tiempo de conteo por $\{Y_t : t \in \mathbb{N}\}$. Denotaremos por $\{\mathbf{X}_t : t \in \mathbb{N}\}$ una variable de tiempo vector aleatorio r-dimensional, es decir $\mathbf{X}_t = (X_{t,1},\ldots,X_{t,r})^\top$. Modelamos la media condicional $E(Y_t|\mathscr{F}_{t-1})$ de la serie de tiempo de conteo por un proceso, digamos $\{\lambda_t : t \in \mathbb{N}\}$, tal que $E(Y_t|\mathscr{F}_{t-1}) = \lambda_t$. Denotemos por $\mathscr{F}_t$ la σ-álgebra de los procesos conjuntos $\{Y_t, \lambda_t, \mathbf{X}_{t+1} : t \in \mathbb{N}\}$ hasta el tiempo t incluyendo la información de la covariable en el tiempo $t+1$. La distribución supuesta por Y_t se aborda más adelante. Sea el método

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \tilde{g}(Y_{t-i_k}) + \sum_{\ell=1}^{q} \alpha_\ell g(\lambda_{t-j_\ell}) + \eta^\top \mathbf{X}_t, \tag{3.1}$$

donde $g : \mathbb{R}^+ \to \mathbb{R}$ es una función de enlace y $\tilde{g} : \mathbb{N}_0 \to \mathbb{R}$ es una función transformación. El vector parámetro $X = (X_1,\ldots,X_r)^\top$ corresponde a los efectos de las covariables.
Por ejemplo, sea la serie de tiempo $Y_t :=$ "Precio del café en una tienda en el día t". Se toma que el precio de un capuchino se calcula por Precio del Capuchino$(Y_t) = 4Y_t + 0{,}2$ entonces se utiliza

la siguiente serie de tiempo para modelar el precio del capuchino usando un modelo de conteo

$$\begin{aligned}\text{Precio del Capuchino}(\lambda_t) = \beta_0 \quad &+\sum_{k=1}^{p} \beta_k \text{Precio del Capuchino}(Y_{t-i_k}) \\ &+\sum_{\ell=1}^{q} \alpha_\ell \text{Precio del Capuchino}(\lambda_{t-j_\ell})\end{aligned}$$

En este caso la función de enlace y de transformación coinciden.
$g(\lambda_t)$ se define como el predictor lineal. Para permitir la regresión en las observaciones del pasado arbitrario de la respuesta, definimos un conjunto $P = \{i_1, i_2, \ldots, i_p\}$ y enteros $0 < i_1 < i_2 \ldots < i_p < \infty$, con $p \in \mathbb{N}_0$. Esto nos permite regresar en las observaciones atrasadas $Y_{t-i_1}, Y_{t-i_2}, \ldots, Y_{t-i_p}$. Análogamente, defina un conjunto $Q = \{j_1, j_2, \ldots, j_q\}$, $q \in \mathbb{N}_0$ y enteros $0 < j_1 < j_2 \ldots < j_q < \infty$, para la regresión sobre medias condicionales rezagadas $\lambda_{t-j_1}, \lambda_{t-j_2}, \ldots, \lambda_{t-j_q}$. Este caso es cubierto por la teoría de modelos con $P = \{1, \ldots, p\}$ y $Q = \{1, \ldots, q\}$ eligiendo p, q adecuadamente y configurando algunos parámetros del modelo a cero.
A partir del modelo lineal generalizado dado por la ecuación 3.1 se definen los modelos de las definiciones 1.1 y 1.2, para los cuales no se consideran covariables.
Definición (Heinen [2003], Ferland, Latour, y Oraichi [2006] y Fokianos, Rahbek, y Tjøstheim [2009]) Sean g y $\tilde{g}$ iguales a la identidad, es decir, $g(x) = \tilde{g}(x) = x$. Sea $P = \{1, \ldots, p\}$ y $Q = \{1, \ldots, q\}$. Entonces el modelo

$$\lambda_t = \beta_0 + \sum_{k=1}^{p} \beta_k Y_{t-k} + \sum_{\ell=1}^{q} \alpha_\ell \lambda_{t-\ell} \tag{3.2}$$

Se llama el modelo condicional autorregresivo de Poisson. El modelo $(3,1)$ junto con la suposición de *Poisson*, es decir, $Y_t|\mathscr{F}_{t-1} \sim Poisson(\lambda_t)$, es decir,

$$P(Y_t = y|\mathscr{F}_{t-1}) = \frac{\lambda_t^y exp(-\lambda_t)}{y!} \tag{3.3}$$

Se mantiene $V(Y_t|\mathscr{F}_{t-1}) = E(Y_t|\mathscr{F}_{t-1}) = \lambda_t$. Por lo tanto, en el caso de una respuesta del modelo condicional de Poisson la media condicional es idéntica a la varianza condicional del proceso observado.
Definición(Christou y Fokianos [2014]) Se supone que $Y_t|\mathscr{F}_{t-1} \sim NegBin(\lambda_t, \phi)$, donde la distribución Binomial Negativa es parametrizada en términos de su media con un parámetro de dispersión adicional $\phi \in (0, \infty)$, es decir,

$$P(Y_t = y|\mathscr{F}_{t-1}) = \binom{y-1}{\lambda_t - 1} \phi^{\lambda_t} (1-\phi)^{y-\lambda_t},\ y = 0, 1, \ldots \tag{3.4}$$

En este caso, $V(Y_t|\mathscr{F}_{t-1}) = \lambda_t + \lambda_t^2/\phi$, es decir, la varianza condicional aumenta cuadráticamente con λ_t. La distribución de Poisson es un caso límite de la Binomial Negativa cuando $\phi \to \infty$.

La distribución *Binomial Negativa* permite que una varianza condicional sea mayor que la media λ_t, que a menudo se denomina sobredispersión.

En el modelo 3.1, el efecto de una covariable entra de lleno en la dinámica del proceso y se propaga a observaciones futuras tanto por la regresión sobre observaciones pasadas como por la regresión sobre medias condicionales pasadas. El efecto de dichas covariables puede considerarse como una influencia interna en el proceso de generación de datos, por lo que lo denominamos efecto de covariable interna. También se permite incluir covariables de forma que su efecto solo se propague a futuras observaciones por la regresión en observaciones pasadas pero no directamente por la regresión en medias condicionales pasadas.

Siguiendo a (Liboschik, Kerschke, Fokianos, y Fried, 2016), que hacen esta distinción para el caso de efectos de intervención descritos por covariables deterministas, se refiere al efecto de dichas covariables como un efecto de covariable externa. Sea $\mathbf{e} = (e_1, \ldots, e_r)^\top$ un vector especificado por el usuario con $e_i = 1$ si la i-esima componente del vector covariable tiene un efecto externo y $e_i = 0$ en otro caso, $i = 1, \ldots, r$. Denote por $diag(e)$ una matriz diagonal con elementos en la diagonal dados por $\mathbf{e}$. La generalización de 3.1 teniendo en cuenta los efectos de las covariables internas y externas viene dada por

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \tilde{g}(Y_{t-i_k}) + \sum_{\ell=1}^{q} \alpha_\ell (g(\lambda_{t-j_\ell}) - N^\top diag(e)(\mathbf{X}_{t-j_\ell}) + N^\top \mathbf{X}_t \tag{3.5}$$

Básicamente, el efecto de todas las covariables con un efecto externo se resta en los términos de retroalimentación, de forma que su efecto entra en la dinámica del proceso sólo a través de las observaciones. Para una amplia discusión y comparación de los efectos internos y externos (Liboschik y cols., 2016). La experiencia con estos modelos es que una discriminación empírica entre efectos de covariables internos y externos es difícil y que no es crucial que tipo de efecto de covariable se elija para las aplicaciones.

4. Metodología

4.1. Imputación

En estadística, la imputación se refiere al proceso de completar los valores de datos faltantes con valores estimados basados en otros datos disponibles. Esto se realiza a menudo con el fin de evitar resultados sesgados o para garantizar que los modelos estadísticos se apliquen de forma correcta. Los métodos de imputación pueden ser simples, como usar la media o la mediana de los datos disponibles, o más complejos, como usar modelos de regresión o algoritmos de aprendizaje automático. Para solucionar la ausencia de estos datos en R se emplea el paquete $'imputeTS'$ y la función `nainterpolation` (Moritz y cols., 2019) "Utiliza interpolación lineal, spline para reemplazar los valores faltantes."
Respecto a la interpolación se tiene en cuenta el libro de (Burden, Faires, y Burden, 2015), donde las funciones que mapean el conjunto de números reales en sí mismo. Los polinomios algebraicos y las funciones de la forma

$$EP_n(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

donde $n \in \mathbb{Z}$ y $a_0, \ldots, a_n$ son constantes reales. Una razón de su importancia es que se aproximan de manera uniforme a las funciones continuas. Dada una función, definida y continua sobre un intervalo cerrado y acotado, existe un polinomio que está tan "cerca"de la función como se desee.
Teorema 1. (Teorema de aproximación de Weiertrass)((Burden y cols., 2015)) Suponga que f está definida y es continua en $[a,b]$, para cada $\varepsilon > 0$, existe un polinomio $EP(x)$, con la propiedad de que

$$|f(x) - EP(x)| < \varepsilon, \text{para todas las } x \in [a,b]$$

Otra razón importante para considerar la clase de polinomios en la aproximación de funciones es que la derivada y la integral indefinida de un polinomio son fáciles de determinar y también son polinomios. Por esta razón, a menudo se usan polinomios para aproximar funciones continuas.
El uso de las funciones con los polinomios anterioremente mencionados se usan para aproximar dentro de un intervalo que pasa por puntos dados se conoce como interpolación.
Se opta por usar la interpolación por spline cúbico. Se modifica la definición de (Burden y cols., 2015) donde la función f son los valores de la serie de tiempo Y_t documentados. Se deben aproximar los Y_{t_N} la variable aleatoria de la serie de tiempo para la cual no se tienen datos. A continuación una definición que es de utilidad para la construcción de este spline.
Definición 2.(Burden y cols., 2015) Dada una función Y_t definida en $[a,b]$ donde a es el primer dato de la serie, b es el último dato de la serie de tiempo y un conjunto de nodos $a = t_0 < t_1 < \cdots < t_n = b$, un interpolante de spline cúbico S para Y_t es una función que satisace las siguientes condiciónes:

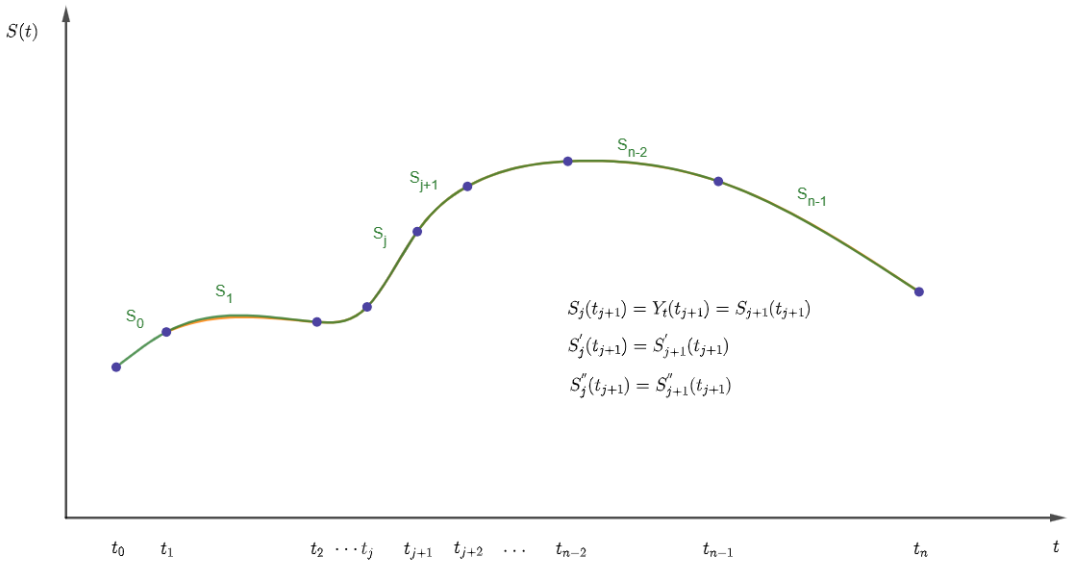
1. $S(t)$ es un polinomio cúbico, que se denota $S_j(t)$, en el subintervalo $[t_j, t_{j+1}]$ para cada

$j = 0, 1, \ldots, n-1$;

2. $S_j(t_j) = f(t_j)$ y $S_j(t_{j+1})$ para cada $j = 0, 1, \ldots, n-1$;
3. $S_{j+1}(t_{j+1)=S_j(t_{j+1})}$ para cada $j = 0, 1 \ldots, n-2$ (implícito en **b**).);
4. $S'_{j+1}(t_{j+1}) = S'^{(t_{j+1})}_j$ para cada $j = 0, 1, \ldots, n-2$;
5. $S''_{j+1}(t_{j+1}) = S''_j(t_{j+1})$ para cada $j = 0, 1, \ldots, n-2$;
6. Uno de los siguientes conjuntos de condiciones frontera se satisface:

   a) $S''(x_0) = S''_j(t_n) = 0$ (**frontera natural**);

   b) $S'(x_0) = f'(t_0)$ y $S'(t_n) = f'(t_n)$ (**frontera condicionada**).



Fuente: Elaboración propia

Dada esta definición se procede a la construcción de acuerdo a (Burden y cols., 2015) del spline cúbico: Se interpola para una función dada $f = Y_t$ (para este caso se emplea en la serie de tiempo), las condiciones en la definición se aplican a los polinomios cúbicos de la forma

$$S_j(t) = a_j + b_j(t - t_j) + c_j(t - t_j)^2 + d_j(t - t_j)^3,$$

para cada $j = 0, 1, \ldots, n-1$. Puesto que $S_j(t_j) = a_j = f(t_j)$, la condición **3**) de la definición anterior se puede aplicar para obtener

$$a_{j+1} = S_{j+1}(t_{j+1}) = S_j(t_{j+1}) = a_j + b_j(t_{j+1}) - t_j) + c_j(t_{j+1}) - t_j)^2 + d_j(t_{j+1}) - t_j)^3,$$

para cada $j = 0, 1, \ldots, n-2$.
Como los términos $t_{j+1} - t_j$ son usados repetidamente en este desarrollo, es conveniente introducir una notación más simple

$$h_j = t_{j+1} - t_j,$$

para cada $j = 0, 1, \ldots, n-1$. Si tambien defininmos $a_n = Y_t(t_n)$, entonces la ecuación

$$a_{j+1} = a_j + b_j h_j + c_j h_j^2 + d_j h_j^3 \tag{4.1}$$

se mantiene para cada $j = 0, 1, \ldots, n-1$.
De manera similar, defina $b_n = S^{'(t_n)}$ y observe que

$$S_j^{'}(t) = b_j + 2c_j(t - t_j) + 3d_j(t - t_j)^2$$

implica que $S_j^{'}(t_j) = b_j$, para cada $j = 0, 1, \ldots, n-1$. Al aplicar la condición (4.) obtenemos

$$b_{j+1} = b_j + 2c_j h_j + 3d_j h_j^2 \tag{4.2}$$

para cada $j = 0, 1, \ldots, n-1$.
Otra relación entre los coeficientes de S_j se obtiene al definir $c_n = S^{''}(t_n)/2$ y aplicar la condición (5.). Entonces, para cada $j = 0, 1, \ldots, n-1$,

$$c_{j+1} = c_j + 3d_j h_j. \tag{4.3}$$

Resolviendo para d_j en la ecuación (4.3) y sustituyendo este valor en las ecuaciones (4.1) y (4.2) obtenemos, para cada $j = 0, 1, \ldots, n-1$, las nuevas ecuaciones

$$a_{j+1} = a_j + b_j h_j + \frac{h_j^2}{3}(2c_j + c_{j+1}) \tag{4.4}$$

y

$$b_{j+1} = b_j + h_j(c_j + c_{j+1}). \tag{4.5}$$

La relación final que involucra los coeficientes se obtiene al resolver la ecuación adecuada en la forma de la ecuación (4.4), primero para b_j,

$$b_j = \frac{1}{h_j}(a_{j+1} - a_j) - \frac{h_j}{3}(2c_j + c_{j+1}), \tag{4.6}$$

y entonces, con una reducción del índice, para b_{j-1}. Esto nos da

$$b_{j-1} = \frac{1}{h_{j-1}}(a_j - a_{j-1}) - \frac{h_{j-1}}{3}(2c_{j-1} + c_j).$$

Al sustituir estos valores en la ecuación obtenida de la ecuación (4.5), con el índice reducido en uno, obtenemos el sistema lineal de ecuaciones

$$h_{j-1}c_{j-1} + 2(h_{j-1} + h_j)c_j + h_j c_{j+1} = \frac{3}{h_j}(a_{j-1} - a_j) - \frac{3}{h_{j-1}}(a_j - a_{j-1}), \tag{4.7}$$

Para cada $j = 1, 2, \ldots, n-1$. Este sistema sólo tiene los $\{c_j\}_{j=0}^{n}$ como incógnitas. Los valores de $\{h_j\}_{j=0}^{n-1}$ y $\{a_j\}_{j=0}^{n}$ están dados, respectivamente, por el espaciado e los nodos $\{t_j\}_{j=0}^{n}$ y los valores de f en los nodos. Por lo que, una vez que se determinan los valores de $\{c_j\}_{j=0}^{n}$, es sencillo encontrar el resto de las constantes $\{b_j\}_{j=0}^{n-1}$ a partir de la ecuación (4.6) y $\{d_j\}_{j=0}^{n}$, a partir de la ecuación (4.3). Entonces podemos construir los polinomios cúbicos $\{S_j(t)\}_{j=0}^{n-1}$.

4.2. Estimación

El paquete **tscount** ajusta modelos de la forma (3.1) mediante estimación máxima verosimilitud (ML) (función `tsglm`). Si se cumple el supuesto de Poisson, obtenemos un estimador ML ordinario.
Denotemos por $\theta = (\beta_0, \beta_1, \ldots, \beta_p, \alpha_1, \ldots, \alpha_q, \eta_1, \ldots, \eta_r)^\top$ el vector de parametros de regresión.

$$\Theta = \left\{ \theta \in \mathbb{R}^{p+q+r+1} : \beta_0 > 0, \beta_1, \ldots, \beta_p, \alpha_1, \ldots, \alpha_q, \eta_1, \ldots, \eta_r \geq 0, \sum_{k=1}^{p} \beta_k + \sum_{l=1}^{q} \alpha_l < 1 \right\}.$$

El intercepto β_0 debe ser positivo y todos los demás parámetros deben ser no negativos para garantizar la positividad de la media condicional λ_t. La otra condición garantiza que el modelo ajustado tiene una solucion estacionaria y ergódica con momentos de cualquier orden (Ferland, Latour, y Oraichi, 2006)(Doukhan, Fokianos, y Tjøstheim, 2012); véase también (Tjøstheim, 2015) para una revisión reciente. Para el modelo log-lineal 3.2 con covariables, el espacio de parámetros se toma como

$$\Theta = \left\{ \theta \in \mathbb{R}^{p+q+r+1} : |\beta_1|, \ldots, |\beta_p|, |\alpha_1|, \ldots, |\alpha_q| < 1, \left| \sum_{k=1}^{p} \beta_k + \sum_{l=1}^{q} \alpha_l \right| < 1 \right\}.$$

En (Liboschik, Fokianos, y Fried, 2017) esto se profundiza. En (Christou y Fokianos, 2014) señalan que con la parametrización (3.4) de la distribución Binomial Negativa la estimación de los parámetros de regresión θ no depende del parámetro de dispersión ϕ.Esto permite emplear un

enfoque de máxima verosimilitud basado en la verosimilitud de Poisson para estimar los parámetros de regresión θ, que se describe a continuación. El parámetro perturbador ϕ se estima por separado en un segundo paso. Este enfoque es diferente de una estimación máxima verosimilitud basada en la distribución Binomial Negativa, que, por ejemplo, se ha implementado en la función `glm.nb` del paquete `R` **MASS** (Venables y Ripley, 2013). En ese algoritmo, la maximización de la verosimilitud Binomial Negativa para un parámetro de dispersión estimado ϕ y estimación de ϕ dados los parámetros de regresión estimados θ se itera hasta la convergencia. El enfoque binomial negativo se ha elegido por su sencillez y su utilidad para obtener estimadores coherentes cuando el modelo para λ_t especficado correctamente (para más información (Ahmad y Francq, 2016)).

La log-verosimilitud, el vector de puntuación y la matriz de información se derivan de forma condicional de los valores pre-muestra de la serie temporal y del proceso de media condicional $\{\lambda_t\}$, precisamente de $\mathscr{F}_0$. Se necesita una inicialización apropiada para su evaluación. Para un vector de observaciones $\mathbf{y} = (y_{t_1}, \ldots, y_{t_n})^\top$, la función de log-similitud condicional viene dada por

$$\ell(\theta) = L(\theta | y_{t_n}, i = 1, ..., n) = \sum_{t=1}^{n} log p_t(y_t|\theta) = \sum_{t=1}^{n} \left(y_t ln(\lambda_t(\theta)) - \lambda_t(\theta) \right), \tag{4.8}$$

Donde $p_t(y;\theta) = P(Y_t = y|\mathscr{F}_{t-1})$ es la función de densidad de probabilidad de una distribución de Poisson definida como en (3.3). La media condicional es considerada como una función $\lambda_t : \Theta \to \mathbb{R}^+$ y estos son denotados por $\lambda_t(\theta)$ para todo t.

El estimador máxima verosimilitud (MLE) $\hat{\theta}_n$ de θ, suponiendo que exista, es la solución del problema de optimización no lineal con restricciones

$$\hat{\theta} := \hat{\theta}_n = \ \text{MAX}_{\theta \in \Theta} \ell(\theta). \tag{4.9}$$

Los cuales se maximizan computacionalmente.

4.3. Implementación

La configuración por defecto de estos argumentos se eligen basándose en muchos experimentos que deberían ser suficientes para la mayoría de las situaciones.

Las restricciones de parámetros que impone la condición $\theta \in \Theta$ puede ser formulada como d desigualdades lineales. Esto significa que existe una matriz $\mathbf{U}$ de dimensión $d \times (p+q+r+1)$ y un vector $\mathbf{c}$ de longitud d, tal que $\Theta = \{\theta | \mathbf{U}\theta \geq c\}$. Para modelos lineales (3.1) se necesita $d = p+q+r+2$ para garantizar la no negatividad de la media condicional λ_t y la estacionalidad del proceso resultante.

Para resolver numéricamente el problema de maximización (4.9) empleamos por defecto la fun-

ción `constrOptim`. Esta función aplica un algoritmo descrito por (Lange, Chambers, y Eddy, 2010) (Capítulo 14) que esencialmente impone las restricciones añadiendo un valor de barrera a la función objetivo y luego emplea un algoritmo para la optimización sin restricciones de esta nueva función objetivo, iterando estos dos pasos si es necesario.
Note que la log-verosimilitud (4.8) se da condicionada a valores pre-muestra no observados. Dependen del predictor lineal, que pueden calcularse recursivamente utilizando cualquier inicialización. Las recursiones y varias estrategias para su inicialización se pueden consultar en (Liboschik y cols., 2017) (argumentos `init.method` e `init.drop`). La solución del problema de optimización no lineal (4.9) requiere un valor inicial para el vector de parámetros θ. Este valor inicial puede obtenerse ajustando un modelo más sencillo para el que se disponga de un procedimiento de estimación. Consideramos la posibilidad de ajustar un MLG o de encajar un modelo ARMA. modelo. Una tercera posibilidad es ajustar un modelo ingenuo i.i.d. sin covariables. Además, el usuario puede asignar valores fijos. Todas estas posibilidades están disponibles mediante el argumento `start.control`.
Resulta que el algoritmo de optimización converge de forma muy fiable incluso si los valores iniciales no estén cerca del óptimo global de la verosimilitud. Un valor inicial más cercano al óptimo global suele requerir menos iteraciones hasta la convergencia. Sin embargo, hemos encontrado algunos ejemplos en los que los valores de partida cercanos a uno de los dos primeros métodos mencionados, no dan lugar al óptimo global. En consecuencia, es más recomendable ajustar el modelo ingenuo i.i.d. sin covariables para obtener valores de partida. Para más información algunos detalles se encuentran en (Liboschik y cols., 2017) la implementacion de la función `tsglm` y la explicacuón de sus argumentos técnicos.

4.4. Predicción

En términos del error cuadrático medio, el predictor óptimo 1-paso-por adelantado $\hat{Y}_{n+1}$ para Y_{n+1}, dado $\mathscr{F}_n$, por tanto, el pasado del proceso hasta el momento n y las covariables potenciales en el tiempo $n+1$, es la expectativa condicional λ_{n+1} dada en (3.1) (método `S3` de función `predict`). Por construcción del modelo condicional de $\hat{Y}_{n+1}$ una distribución Poisson (3.3) respectivamente distribución Binomial negativa (3.4) con media λ_{n+1}. Una predicción h-paso-adelante $\hat{Y}_{n+h}$ para Y_{n+h} es obtenido por predicciones 1-paso-adelante recursiva, donde los valores no observados $Y_{n+1},\ldots,Y_{n+h-1}$ son reemplazadas por su respectiva predicción 1-paso-adelante, $h \in \mathbb{N}$. La distribución de esta predicción h-paso-adelante $\hat{Y}_{n+h}$ no es conocida analíticamente pero puede ser aproximada numéricamente por un procedimiento bootstrap paramétrico, que es descrito abajo.
En aplicaciones, λ_{n+1} es sustituido por estimadores $\hat{\lambda}_{n+1} = \lambda_{n+1}(\hat{\theta})$, que depende de los parámetros de regresión estimados $\hat{\theta}$. El parámetro de dispersión *phi* de la distribución Binomial Nega-

tiva es reemplazada por su estimador $\hat{\phi}$. Note que introduciendo los parámetros estimados induce una incertidumbre adicional a la distribucion predictiva. Esta incertidumbre de estimación no se tiene en cuenta para la construcción de los intervalos de predicción descritos en los siguientes párrafos.

Los intervalos de predicción un-paso-adelante se puede obtener de forma directa desde la distribución condicional (argumento `method = "conddistr"`). Los intervalos de predicción obtenidos por un procedimiento bootstrap paramétrico (argumento `method = "bootstrap"`) son basados en B simulaciones de realizaciones $y_{n+1}^{(1)},\ldots,y_{n+h}^{(B)}$ (si `type = "quantiles"` o encontrar el intervalo más pequeño que contiene el menor $\lceil(1-\alpha)\cdot B\rceil$ de estas observaciones (si `type = "shortest"`). Este procedimineto de bootstrap puede ser acelerado distribuyendo a multiples núcleos simultáneamente (argumento `parallel = "TRUE"`), que requiere un cluster de computación registrado por el paquete `R parallel` (ver la página de ayuda de la función `setDefaultCluster`).

La regresión se usa a menudo para las predicciones, para realizar la predicción en R se emplea la función `predict`. La función `predict` ((Lander, 2014)): Se utiliza para generar predicciones o pronósticos basados en un modelo estadístico que se ha ajustado previamente a los datos. La función toma un objeto de modelo ajustado como argumento principal y produce valores o resultados previstos para puntos de datos nuevos o existentes.

El valor resultante de la función da como resultado un vector o matriz de predicciones, o una lista de conformada por las predicciones y los errores estándar. La función `predict` es comúnmente usada para extraer los valores ajustados de un objeto de ajuste, o en el caso de modelos generalizados, extraer el predictor lineal o aditivo. ((Chambers y Hastie, 1992)).

El principio del bootstrap que para el caso de las series de tiempo se usa de la siguiente manera: ((Dekking, Kraaikamp, Lopuhaä, y Meester, 2005)): Usar el conjunto de datos $x_1,x_2,\ldots,x_n$ para calcular una $\hat{F}$ para la distribución "verdadera" F. Reemplace la muestra aleatoria $X_1,X_2,\ldots,X_n$ de F por una muestra aleatoria $X_1^*,X_2^*,\ldots,X_n^*$ de $\hat{F}$, y aproximar la distribución de probabilidad de $h(X_1,X_2,\ldots,X_n)$ por la de $h(X_1^*,X_2^*,\ldots,X_n^*)$

$$\left(\text{Cuantil}_{\alpha/2}(\lambda_{t_N}),\text{Cuantil}_{1-\alpha/2}(\lambda_{t_N})\right),$$

donde λ_{t_N} corresponde al valor de $E(X_{t_N}|X_t$: t son valores para los cuales se tienen datos), para el tiempo t_N sin datos (tiempo a predecir). Por consiguiente la banda de confianza está dada por

$$\bigcup_{t_N \text{ valor futuro}} \left(\text{Cuantil}_{\alpha/2}(\lambda_{t_N}),\text{Cuantil}_{1-\alpha/2}(\lambda_{t_N})\right).$$

De esta manera, la banda de confianza es la unión de los intervalos de confianza para la predicción, es decir, para los t_N no observados para la serie de tiempo, Y_{t_N}.

4.5. Descripción base de datos

La base de datos ‘Control venta de gallinas.xlsx’ contiene los números de un negocio que distribuye y comercializa aves sacrificadas en unos municipios, los datos son recopilados de cuadernos donde estos llevan el control de los pedidos por cliente de cada municipio y las ventas del local de la plaza de mercado.
En contadas ocasiones estos datos no se encuentran anotados dentro de los cuadernos y dentro de la base de datos se encuentran como “NA” Estos datos por municipio y local estan dados en el total de aves requeridas según la ubicación, se recopilaron datos desde el 18 de octubre de 2019 hasta el 8 de octubre de 2022. Es decir, se tienen 1087 (días) datos, algo relevante en la venta de estas aves es que a diferencia del pollo estas solo se comercializan completas y no por presas. Por tanto, los valores que se encuentran en las variables son números enteros. Debido a la naturaleza de los datos se suponen los modelos con una distribución discreta como lo son la distribución binomial negativa y distribución de Poisson .
A continuación, la siguiente tabla presenta las 11 variables del documento. Estas son: ‘Fecha’, ‘Total de gallinas’, ‘Local’, ‘Restaurante’, ‘Municipio 1’, ‘Municipio 2’, ‘Municipio 3’, ‘Saldo de aves’, ‘AMR’, ‘TSGAMR’ y ‘Venta congelada’.

Variables de la base de datos		
Variable	Clase	Descripción
Fecha	Cualitativa	Registro de la fecha.
Total gallinas	Cuantitativa discreta	Cantidad de aves sacrificadas en total. Suma de las cantidades del local, los municipios, restaurante y AMR.
Local	Cuantitativa discreta	Cantidad de aves que se comercializan en un local de la plaza de mercado.
Restaurante	Cuantitativa discreta	Cantidad de aves sacrificadas en un restaurante popular del municipio.
Municipio 1	Cuantitativa discreta	Suma del número de aves sacrificadas para los comerciantes de la plaza de mercado del Municipio 1.

Municipio 2	Cuantitativa discreta	Suma del número de aves sacrificadas para los comerciantes de la plaza de mercado del Municipio 2.
Municipio 3	Cuantitativa discreta	Suma del número de aves sacrificadas para los comerciantes de la plaza de mercado del Municipio 3. (Actualmente no funciona)
Saldo de aves	Cuantitativa discreta	Cantidad de aves que no se lograron comercializar. Estas aves se congelan y posteriormente se venden a un menor precio o se regalan.
AMR	Cuantitativa discreta	AMR son la cantidad total de aves A= ahogadas, M= maltratadas y R= regaladas. Estas son aves que cuentan con defectos relacionados al peso y su apariencia.
TSGAMR	Cuantitativa discreta	Esta variable reune la cantidad Total Sin Gallinas Ahogadas, Maltratadas y Regaladas.
Venta congelada	Cuantitativa discreta	Cantidad de aves que se encuentran en el congelador y son vendidas a un menor precio que las otras.

En la sección 3.3 la serie de tiempo de conteo esta definida como $\{Y_t : t \in \mathbb{N}\}$. Para esta aplicación Y_t es la serie de tiempo “Total gallinas"donde cada dato representa las ventas en un día t. En la base de datos ‘control venta de gallinas’ los datos se extraen de cuadernos donde se lleva el registro de los pedidos de cada cliente, uno de los inconvenientes que se presentan es que alrededor de 22 días no tienen registro. Estos datos que no estan disponibles se encuentran de manera discontinua, por ejemplo se tiene un día martes sin registro pero los datos del día lunes y miércoles si se encuentran registrados.

En series de tiempo el orden juega un rol importante, es así que estos datos “NA” o no registrados no pueden ser descartados ya que afecta el orden de la serie. Suponga que es un dato del día domingo el que se descarta, esto provoca que el día que sigue (lunes) se tome como un día domingo. Para solucionar esto se emplea la interpolación definida en la sección 4.1. Los datos calculados con la imputación, son reemplazados en la cantidad total de aves “NA”, el respectivo ajuste con los nuevos datos imputados.

Los datos observados se encuentran separados por días desde 18 de octubre 2019 hasta el 8 de octubre de 2022. Es así que se asigna el dato Y_1 al total de aves del día 18 de octubre de 2019 y para el último dato Y_{1087} se asocia la cantidad del día 8 de octubre de 2022. A continuación la gráfica de la serie de tiempo con los valores interpolados.

Al graficar los 1087 datos de la variable ”total gallinas"se visualizan valores estan entre 0 y 2800,

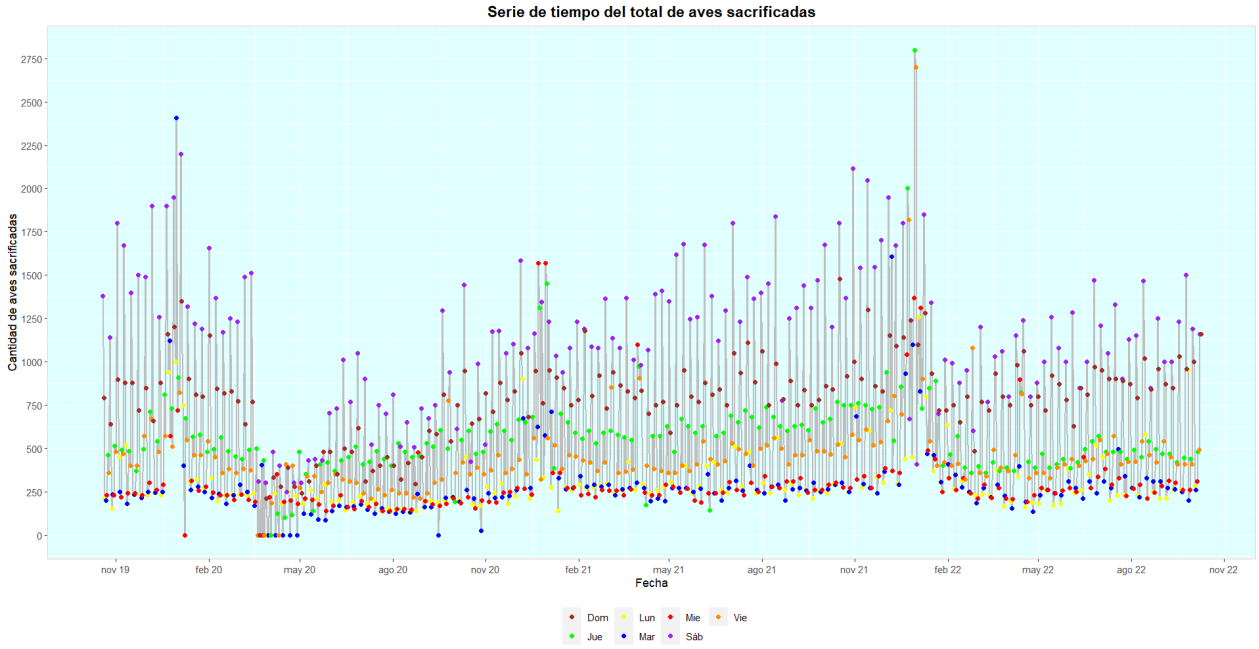


Figura 4.1

Gráfico de la serie de tiempo del total de aves sacrificadas durante 1087 días.

se ven valores cero seguidos entre marzo 2020 y mayo 2020 que se dieron por las restricciones de la pandemia de covid-19. Se observa que los datos del día sábado por lo general estan ubicados en lo más alto de la serie, posicionandose como el día en el que hay mejores ventas y que los valores más bajos se encuentran en los días lunes, martes y miércoles que tienen los colores, teniendo en cuenta esto se puede tener en cuenta el día de la semana como algo importante para las ventas en este negocio, así como los meses dado que se ve como varian especialmente los días sábados según el mes. Se observa en unos pocos días de la serie que valores de días martes, miércoles y jueves presentan valores similares muy por encima de la tendencia que estos tienen, siendo para fechas que tienen grandes ventas como lo son el 24 de diciembre, 31 de diciembre, etc.

5. Análisis y resultados

5.1. Implementación sin covariables

Inicialmente se prueba el comportamiento del modelo sin el uso de las covariables. Se presenta el que obtiene un mejor comportamiento de acuerdo con la obtención de valores plausibles para la predicción, es decir, el conjunto de valores del futuro que pueden verse como datos faltantes, estos son generados utilizando el modelo sin covariables que se adapta a los datos disponibles. Se debe tener en cuenta que las previsiones plausibles pueden cambiar con el tiempo, a medida que haya nueva información disponible, el pronóstico se puede modificar. En este caso el modelo que mejor se adapta toma 3 observaciones pasadas, 14 medias condicionales pasadas y como función de enlace la función identidad, se comparan las distribuciones binomial negativa y Poisson con el modelo dado por:

$$\lambda_t = \beta_0 + \sum_{k=1}^{3} \beta_k(Y_{t-i_k}) + \sum_{\ell=1}^{14} \alpha_\ell(\lambda_{t-j_\ell})$$

Los resultados de la gráfica 5.1 muestran de lado derecho el modelo empleando una distribución de Poisson y de lado izquierdo el modelo que emplea la distribución binomial negativa. Note que se tienen los mismos parámetros en ambos modelos y el criterio de Akaike del modelo con distribución de Poisson tiene un valor de $290976,5$ mientras que para el modelo con la distribución binomial negativa se tiene un valor diez o más veces menor siendo este $15711,75$ al comparar los criterios de información de Akaike el de menor valor es el que se considera mejor.

```
> coefficients(GeneralizadoSinCovariablesNB)
 (Intercept)       beta_3     alpha_14
5.532981e+02 6.934171e-12 2.069857e-06
> AIC(GeneralizadoSinCovariablesNB)
[1] 15711.75
> coefficients(GeneralizadoSinCovariablesP)
 (Intercept)       beta_3     alpha_14
5.532981e+02 6.934171e-12 2.069857e-06
> AIC(GeneralizadoSinCovariablesP)
[1] 290976.5
```

Figura 5.1

Parámetros y coeficientes del modelo sin covariables según su distribución.

Así, el modelo la distribución binomial negativa da un mejor ajuste con los datos e igualmente para verificar esto observe que en la figura 5.2 se muestra el diagrama de calibración marginal en donde entre más cercano este las lineas del modelo al cero es mejor. Es notorio que el modelo con distribución binomial negativa es mejor. A continuación, se muestra la serie de tiempo con la predicción que esta genera.

Como se puede notar en la figura 3.4 la serie de tiempo junto a las predicciones tiene una banda de confianza con unos límites inferiores y superiores (gris) demasiado grandes, esto tiene relación

a la variabilidad que hay entre días como lunes que se sacrifican números muy inferiores a por ejemplo el día sábado. También se puede notar además de predicciones tienen nula variación entre los días de la semana, haciendo que sean inferiores a las esperadas en un día sábado,domingo o sobrestimadas para los días lunes, martes y miércoles que tienden a ser casi la mitad de las generadas por el modelado. Para mejorar este modelo, se va a analizar la serie de tiempo con covariables para ver si existe una mejora considerable en el modelo.

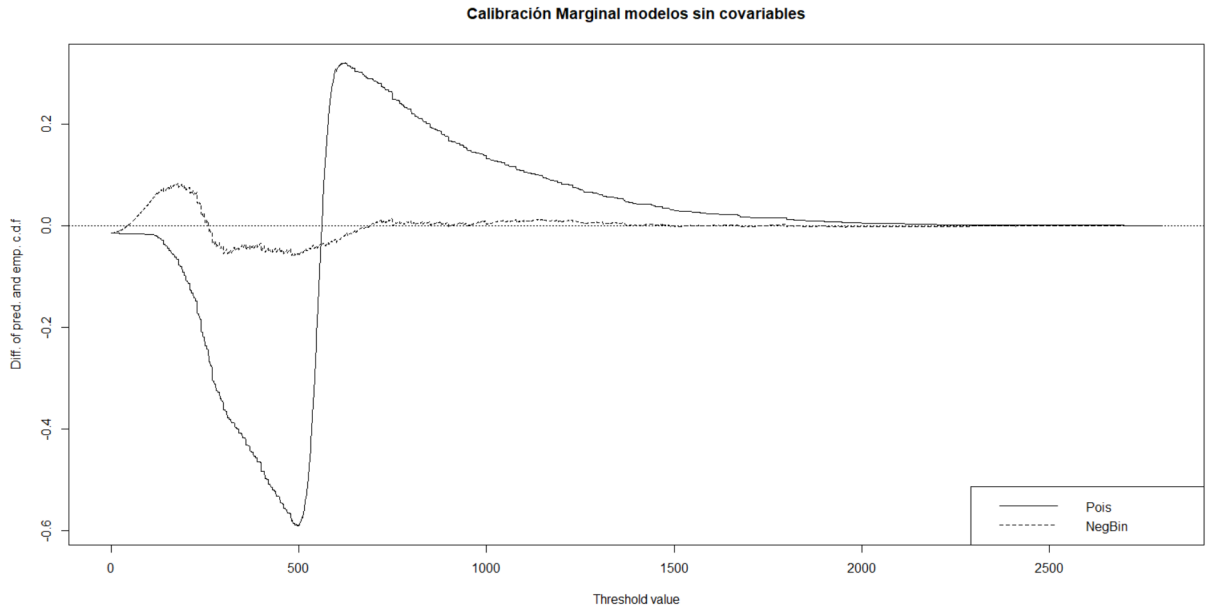


Figura 5.2

Diagrama de calibración marginal para los modelos sin covariables con distribución de Poisson y binomial negativa.

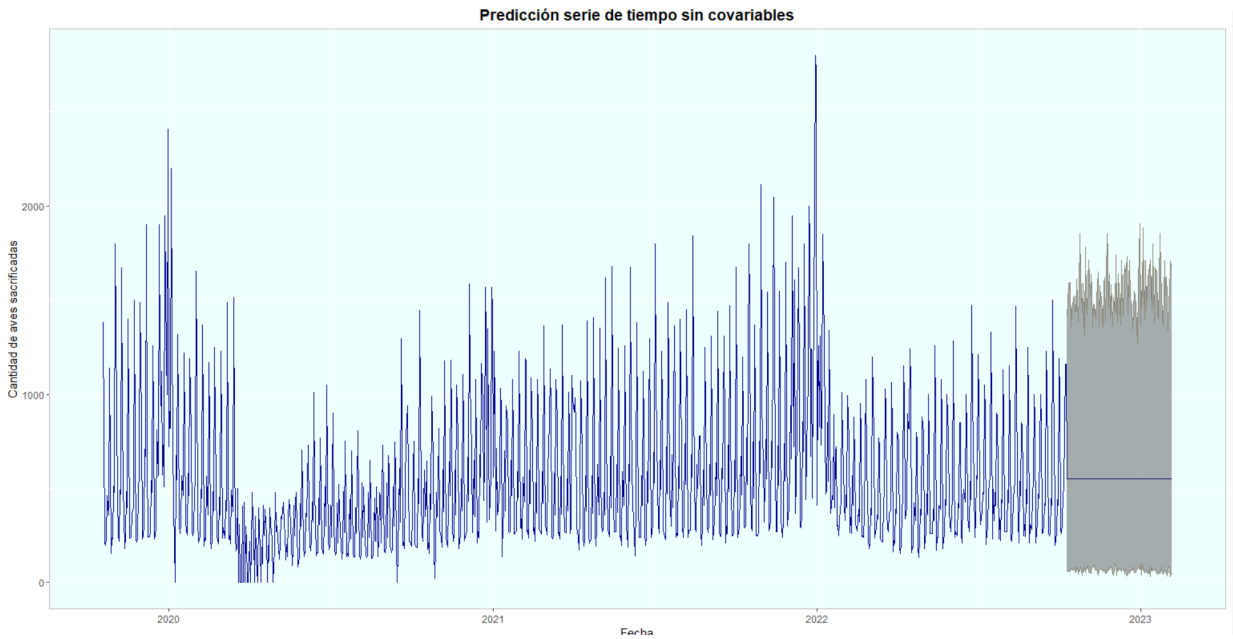


Figura 5.3

Serie de tiempo con su predicción para el modelo con distribución binomial negativa.

5.2. Covariables

Las covariables que se tendrán en cuenta para el modelo en la serie de tiempo Total aves sacrificadas serán los días de la semana y el mes. Primero se debe ver la dependencia de la serie de tiempo respecto al 'día de la semana' y 'mes'.

Se realiza una tabla de contingencia con 10 filas dadas por subintervalos de 280 unidades, en la variable 'día' 7 columnas y 12 columnas para 'mes'. Se analiza primero el día de la semana y la influencia que pueda tener en la cantidad de aves, para esto se realiza una tabla de contingencia entre las cantidades de aves y los días de la semana. Seguido a esto para visualizar de mejor manera los datos se realiza una gráfica con los datos de la tabla de contingencia en un diagrama de barras apiladas.

Note que en la figura 5.5 los días lunes, martes y miércoles son días con poca venta de aves; más de la mitad de los datos se encuentran acumulados en ventas de 0 a 280 gallinas seguido de 280 a 560 aves y se puede notar que los valores menos probables los lunes y martes son los pedidos mayores a 1120 aves. Por otro lado los días jueves y viernes tiene que almenos la mitad de los pedidos son entre 280 y 560 aves. De igual manera la posibilidad de un pedido mayor a 1120 aves es del 2.6% . Por otro lado, los días como el sábado y domingo son los que presentan la menor frecuencia de cantidades a sacrificar menores a 280 aves.

Ahora, se observa la relación entre covariable 'mes' y las ventas de las aves sacrificadas, se realiza una tabla de contingencia en las filas teniendo un conteo del total de aves dividido en 10 inter-

	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic	Sum
(0,280]	16	26	37	41	37	33	30	33	34	26	26	13	352
(280,560]	30	30	30	24	33	32	35	35	25	27	26	19	346
(560,840]	21	12	12	13	11	12	8	13	16	13	18	20	169
(840,1.120]	14	8	9	8	6	8	12	6	8	6	8	15	108
(1.120,1.400]	10	8	3	3	4	3	6	3	4	7	4	10	65
(1.400,1.680]	0	1	2	1	2	2	1	2	3	3	5	6	28
(1.680,1.960]	1	0	0	0	0	0	1	1	0	1	2	6	12
(1.960,2.240]	1	0	0	0	0	0	0	0	0	1	1	1	4
(2.240, 2520]	0	0	0	0	0	0	0	0	0	0	0	1	1
(2520,2800]	0	0	0	0	0	0	0	0	0	0	0	2	2
Sum	93	85	93	90	93	90	93	93	90	84	90	93	1087

Figura 5.4

Tabla de contingencia total de aves y mes

	Lun	Mar	Mie	Jue	Vie	Sáb	Dom	Sum
(0,280]	98	107	106	8	16	1	3	339
(280,560]	44	37	40	78	118	15	22	354
(560,840]	8	6	2	59	15	18	62	170
(840,1.120]	4	2	3	6	4	34	57	110
(1.120,1.400]	1	1	2	1	0	49	11	65
(1.400,1.680]	0	1	2	1	0	24	1	29
(1.680,1.960]	0	0	0	0	1	12	0	13
(1.960,2.240]	0	0	0	1	0	3	0	4
(2.240, 2520]	0	1	0	0	0	0	0	1
(2520,2800]	0	0	0	1	1	0	0	2
Sum	155	155	155	155	155	156	156	1087

Figura 5.5

Tabla de contingencia total de aves y día de la semana

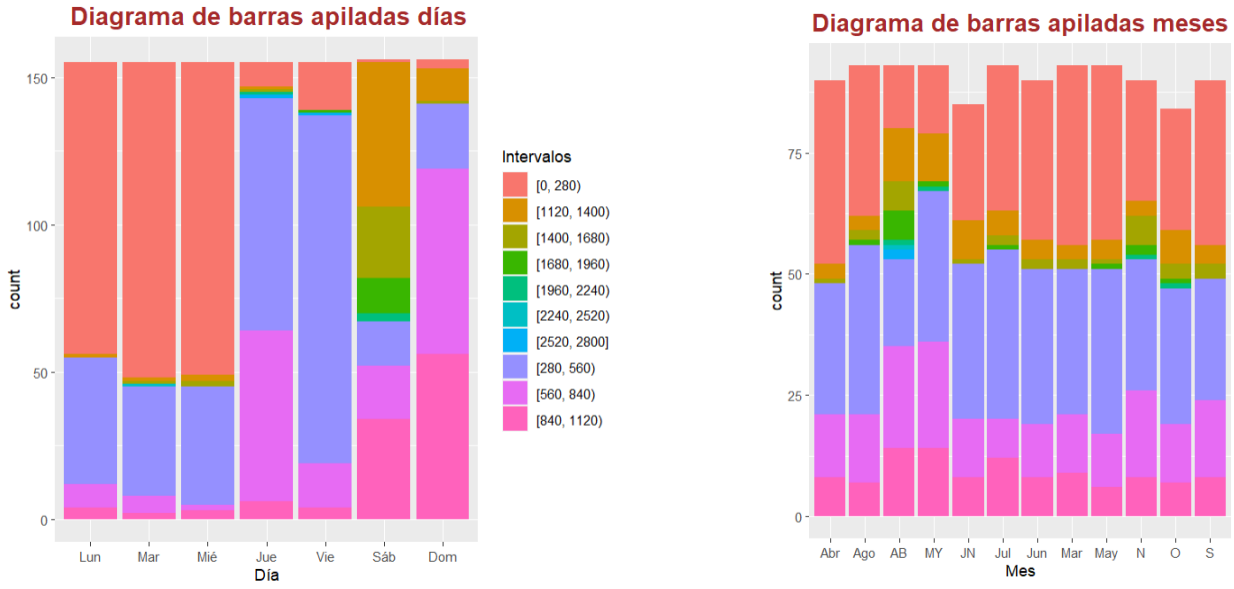


Figura 5.6

Diagrama de barras apiladas para los días de la semana y meses.

valos de 280 y en las columnas los 12 meses del año. Note que para los meses enero y diciembre

son poco frecuentes los pedidos de máximo 280 aves,los datos de enero a octubre recopilados durante estos tres años muestran que solo hay 2 o 3 ocasiones donde los pedidos son superiores a 1400 aves, el mes de enero y diciembre con la menor cantidad de sacrificios menores a 280 aves y diciembre como el mes que presenta los valores máximos en la serie de tiempo siendo el único con ventas superiores a 2240 aves.

Para verificar que existe independencia o relación de la categorización del total de aves por intervalos con la covariable días de la semana y mes se usa el estadístico de Chi-cuadrado χ^2. La figura 5.7 muestra los resultados de esta prueba. Se tiene que para los intervalos vs días se tiene que el $valor - p < 2{,}2e-16$ y los intervalos vs meses se tiene que $valor - p = 0{,}0002031$. Se puede rechazar con una confianza del 95 % que no existe una asociación estadísticamente significativa entre la cantidad de aves y el día de la semana e igualmente se rechaza que no existe alguna asociación entre la cantidad de aves y los meses del año.

```
        Pearson's Chi-squared test

data:  intervsdias
X-squared = 1227.9, df = 54, p-value < 2.2e-16

        Pearson's Chi-squared test

data:  intervsMA
X-squared = 156.55, df = 99, p-value = 0.0002031
```

Figura 5.7

Test χ^2 de Pearson entre intervalos y las covariables día, mes.

5.3. Implementación con covariables

Terminado el análisis de las covariables se procede a emplear el nuevo modelo con el uso de covariables. Anteriormente mediante la prueba de chi cuadrado se evidencio la dependencia que existe entre los días de la semana y la cantidad de aves sacrificadas, más aún la influencia que tienen los meses del año dentro de estas cantidades. Los modelos con covariables que se emplean son los siguientes:

$$\begin{aligned}\lambda_t = \beta_0 \quad &+ \quad \sum_{k=1}^{p} \beta_k \tilde{Y}_{t-i_k} + \sum_{\ell=1}^{q} \alpha_\ell (g(\lambda_{t-j_\ell}) \\ &+ \quad N_1^\top (\mathbf{X}_{t-j_\ell}^{(1)}) + N_2^\top (\mathbf{X}_{t-j_\ell}^{(2)})\end{aligned}$$

Donde $\mathbf{X}^{(1)}$ es la covariable 'Día' y $\mathbf{X}^{(2)}$ es la covariable 'Mes'. Para las variables cualitativas se emplea una matriz dummy que esta dada por vectores de variables (dummy). Esta es la denominación genérica para una variable que toma valores de 0 o de 1 y que se utiliza para re-expresar

variables cualitativas. En R se encuentra esta función en el paquete `'fastDummies'`. El comando arroja una matriz definida por la variable indicadora 'Día de la semana' y 'Mes' con tamaños de 1087 x 7 y 1087 x 12 respectivamente. Las matrices dummy para el 'día de la semana' y el 'mes' son respectivamente:

Teniendo expresados estas variables cualitativas se crea el vector de las covariables que está dado para los modelos de la siguiente manera:

regressors <- cbind(matrizm[,-(1:7)]) donde `'matrizm'` contiene los datos de la matriz dummy para el día y mes según la fecha. Los valores que quedan por definir son los efectos de las covariables y para esto cambiando valores entre las medias pasadas y observaciones manualmente, los mejores ajustes se obtienen con los siguientes valores:

>GeneralizadoP <- tsglm(serie, xreg= regressors, link=ïdentity", + distr = "poisson", model = list(past$_obs = 3, + past_mean = 14, + external = TRUE$)

>GeneralizadoNBin<- tsglm(serie, xreg= regressors, link=ïdentity", + distr = "nbinom", model = list(past$_obs = 3, + past_mean = 14, + external = TRUE$)

Donde GeneralizadoNBin, GeneralizadoP es el modelo con distribuciones binomial negativa y Poisson respectivamente, ambos usando las covariables día y mes. Se puede observar en la predicción del modelo con covariables de la figura 5.8 que la banda de confianza aún sigue presentandose muy grande y no aporta mucho al momento de tener una estimación de las aves a sacrificar porque las predicciones siguen arrojando datos cercanos a 550 aves para cualquier día de la semana. Se puede observar que el modelado con distribución binomial negativa tiene un mejor ajuste a los datos sin embargo aún es necesario tener una predicción más acertada para los días de la semana.

Modelado diario				
Modelo	Covariables	OP	MP	AIC
Poisson	Dia y mes	3	14	290805.5
Binomial negativo	Dia y mes	3	14	14344
Poisson	No	3	14	290976.5
Binomial negativo	No	3	14	15711.75

Discriminando con el valor del criterio de Akaike se puede decir: El modelo con distribución binomial negativa con uso de covariables es el mejor en comparación al modelado sin covariables, el modelo muestra que con distribución de Poisson este no mejora si se emplean estas covariables.

Para dar una mejor aproximación a la cantidad de aves necesaria para cada día de la semana, reducir las bandas de confianza tan grandes se filtran los datos de la serie de tiempo por día y se cambia a una medición semanal para cada día de la semana, que cuenta con 155 datos para los días del lunes a viernes y los fines de semana 156.

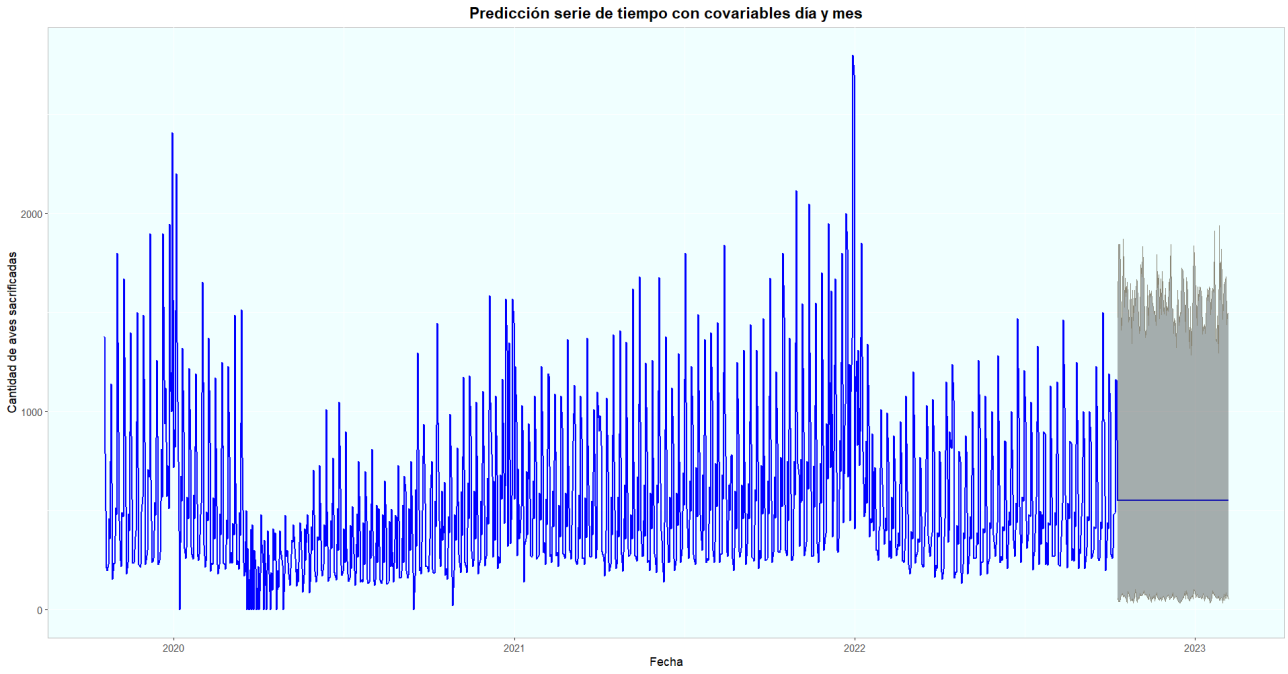


Figura 5.8

Serie de tiempo y predicción para el modelo binomial negativo con las covariables día y mes.

5.4. Modelo aplicado en la covariable día

Teniendo los datos filtrados por días se evalúa la relación que existe entre las cantidades de aves sacrificadas por día de la semana con la variable 'mes' que se extrae de la variable 'Fecha'. Para medir la correlación de estas variables cualitativas se hace mediante el coeficiente V de Cramer cuya definición se da a continuación.

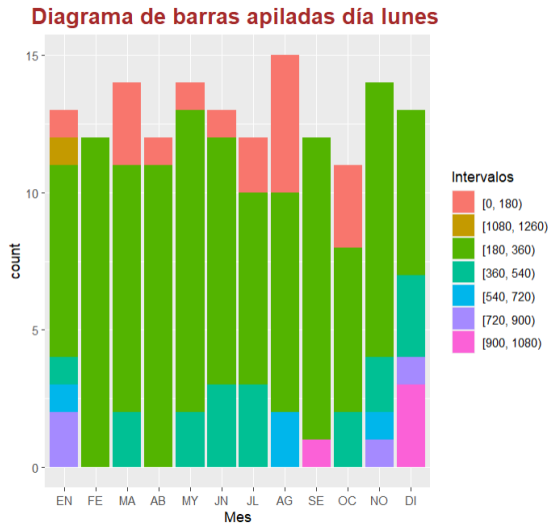
Definición: (Gravetter y cols., 2020) La V de Cramer es la más popular de las medidas de asociación nominal basadas en chi-cuadrado porque proporciona una buena norma de 0 a 1 independientemente del tamaño de la tabla, cuando los marginales de las filas son iguales a los marginales de las columnas. V es igual a la raíz cuadrada de chi-cuadrado dividida por el tamaño de la muestra, n, multiplicado por m, que es el menor de (filas - 1) o (columnas - 1):

$$V_{cramer} = \sqrt{\frac{\chi^2}{n \cdot m}}$$

Interpretación del tamaño del efecto			
Covariable 1	Covariable 2	V_{Cramer}	Interpretación
Cantidad aves por intervalos lunes	Meses	0.3243492	Fuerte
Cantidad aves por intervalos martes	Meses	0.3178504	Fuerte
Cantidad aves por intervalos miércoles	Meses	0.2800577	Moderadamente fuerte
Cantidad aves por intervalos jueves	Meses	0.3165296	Fuerte
Cantidad aves por intervalos viernes	Meses	0.3288715	Fuerte
Cantidad aves por intervalos sábado	Meses	0.3262007	Fuerte
Cantidad aves por intervalos domingo	Meses	0.2964604	Moderadamente fuerte

Para el día lunes se tiene que el test V de Cramer da un resultado de 0.3243492 es decir se tiene una correlación fuerte para estos datos se observa gran concentración de datos entre 180 y 360 aves, estos intervalos se ven en menor medida para meses como enero, noviembre y diciembre que presentan muy pocos datos hasta 180 aves, el promedio de estos datos esta en 305 aves.
Para el día martes se tiene que el test V de Cramer 0.3178504, una correlación fuerte entre los meses y la cantidad de aves para este día. Se puede observar que para el mes de enero y diciembre las ventas de aves van desde 241 aves en adelante, en los meses de febrero a noviembre hay una tendencia a tener solo pedidos de hasta 482 aves, el mes de abril muestra una disminución en las ventas en comparación a los demás meses dado que es el que frecuenta los pedidos de máximo 241 gallinas.
Los resultados del día miércoles arrojan una correlación moderadamente fuerte entre meses y las cantidades por intervalos de aves. Se observa para todos los meses que los pedidos frecuentan la cantidad de aves de 200 a 400 aves, los meses de febrero y diciembre solo cuentan con pedidos superiores a 200 aves, el mes de diciembre presenta los días donde se sacrificaron más aves, el promedio de estos datos se encuentra en alrededor de 300 aves sacrificadas.
Para el día jueves se tiene que el test de chi-cuadrado tiene un valor p de 0.01011, lo que significa que hay menos de un 0.5 % de probabilidad de obtener un resultado tan extremo si no hubiera relación entre las variables. Por lo tanto, se rechaza la hipótesis nula de independencia y para el coeficiente de correlación V de Cramer se obtiene un valor de 0.31653, lo que indica una asociación moderada entre el mes y las aves sacrificadas el día jueves.

Para el día viernes se tiene que el test de chi-cuadrado para independencia entre la cantidad de aves sacrificadas los viernes y el mes, se tiene que el valor de p es 0.003919, lo que significa que hay menos de un 0.5 % de probabilidad de obtener un resultado tan extremo si no hubiera relación entre las variables. Por lo tanto, se rechaza la hipótesis nula de independencia y el coefi-



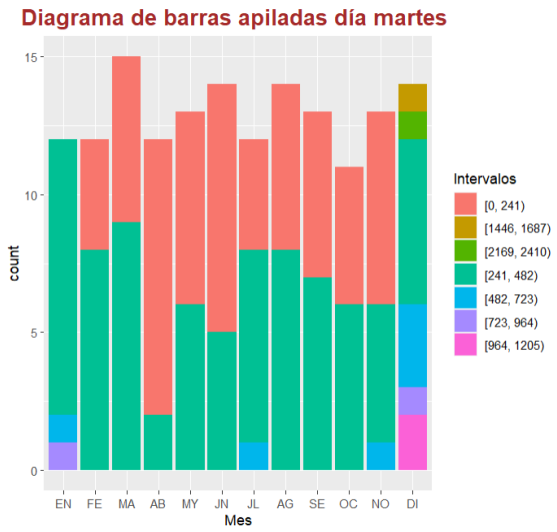
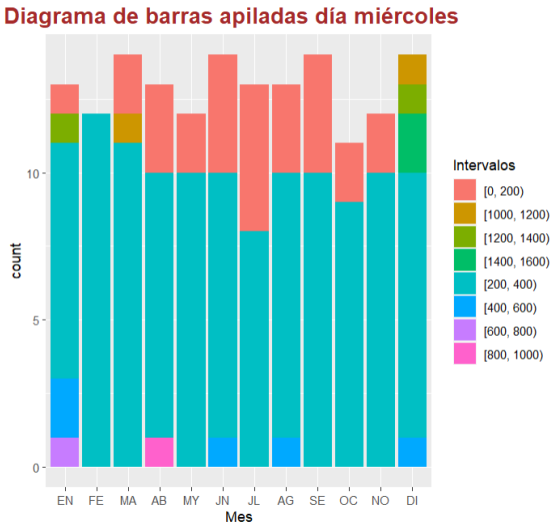


Figura 5.9

Diagramas de barras apiladas de la cantidad de aves por intervalos de los días lunes y martes contra la variable mes.



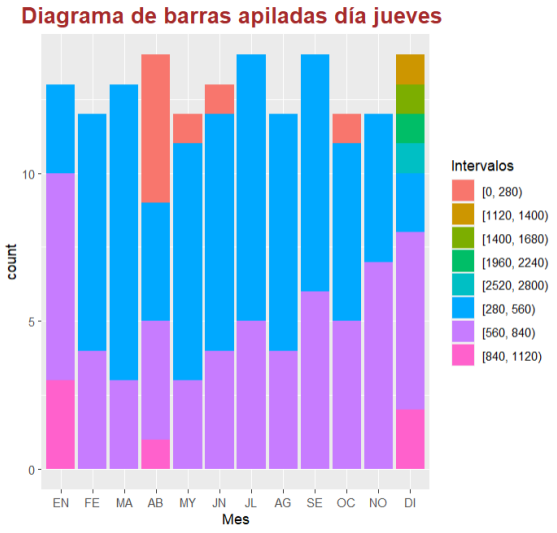


Figura 5.10

Diagramas de barras apiladas de la cantidad de aves por intervalos de los días miércoles y jueves contra la variable mes.

ciente de correlación V de Cramer es 0.32887, lo que indica una asociación moderada entre las dos variables.

Para el día sábado se tiene que el coeficiente V de Cramer arroja un valor que se interpreta como fuerte, para este mes se observa que hay variedad de pedidos y no se puede decir como en los

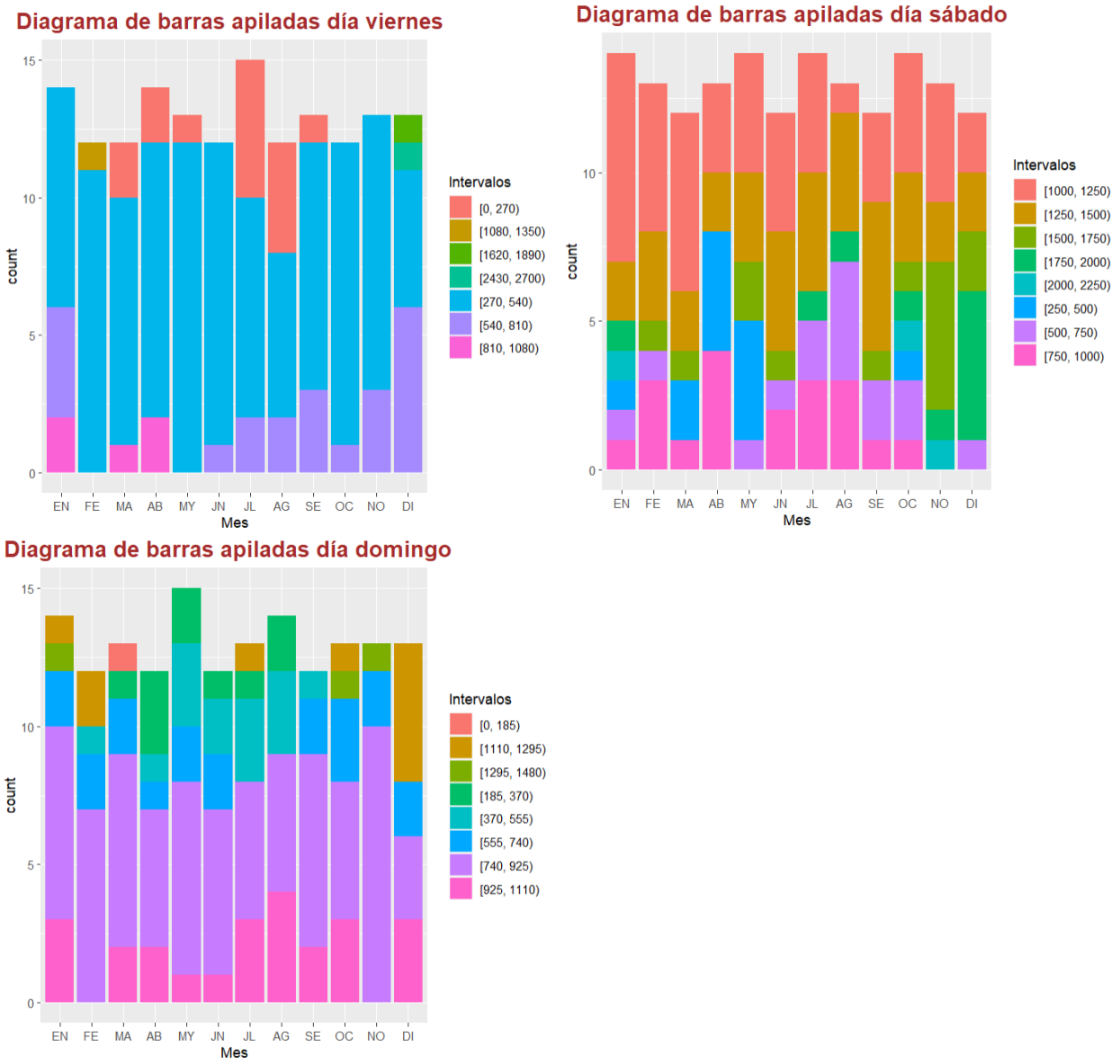


Figura 5.11

Histogramas cantidad de aves por intervalos vs variable mes

días previamente analizados que existe una cantidad que se presente de manera frecuente cada mes, solo se puede decir que para este día de la semana los pedidos son los más altos en comparación al resto de días de la semana independiente del mes.

Para el día domingo el resultado muestra que existe una correlación moderadamente fuerte entre los meses y la cantidad de aves por intervalos. Se observa que en este día de la semana es muy poco probable tener pedidos inferiores a 185, hay una mayor tendencia a que se sacrifiquen en estos días cantidades superiores a 740 aves.

Teniendo en cuenta los resultados mostrados anteriormente para todos los modelos se emplean

estas covariables en los modelos que se definen acontinuación.

5.5. Implementación del modelado semanal

Anteriormente, se observo la dependencia que existe entre las cantidad de aves a sacrificar según el día de la semana y la covariable mes. Para los datos tomados semanalmente se tiene un efecto moderado con esta covariable. El modelo para estos días de la semana variando los p y q en la ecuación son:

$$\begin{aligned}\lambda_t = \beta_0 &+ \sum_{k=1}^{p} \beta_k \tilde{Y}_{t-i_k} + \sum_{\ell=1}^{q} \alpha_\ell (g(\lambda_{t-j_\ell}) \\ &+ N_1^\top (\mathbf{X}^{(1)}_{t-j_\ell})\end{aligned}$$

Donde $\mathbf{X}^{(1)}$ es la covariable 'Mes'
El tratamiento de la matriz dummy para los meses es igual que la definida en **??**. Para todos los días de la semana se presentaran los resultados de forma tabular, calibración marginal y predicción que esta genera.

Predicción lunes				
Modelo	Covariable	MP	OP	AIC
Poisson	Mes	1	6	11756.14
Binomial negativa	Mes	1	6	2000.796
Poisson	No	1	6	160950.1
Binomial negativa	No	1	6	1988.844

Los resultados del modelado para el día lunes muestra como la distribución de Poisson con un uso de covariables mejora su ajuste en comparación al que no la emplea. Sin embargo para los modelos con distribución binomial negativa se observa muy poca variación en el ajuste, mostrando como mejor modelo el que no emplea covariables por una diferencia de 12. En la figura 5.12 se puede notar como el comportamiento de los modelos con distribución binomial negativa es muy similar para ambos, el modelo con distribución Poisson con covariables siendo mejor exceptuando los valores cercanos a 200.

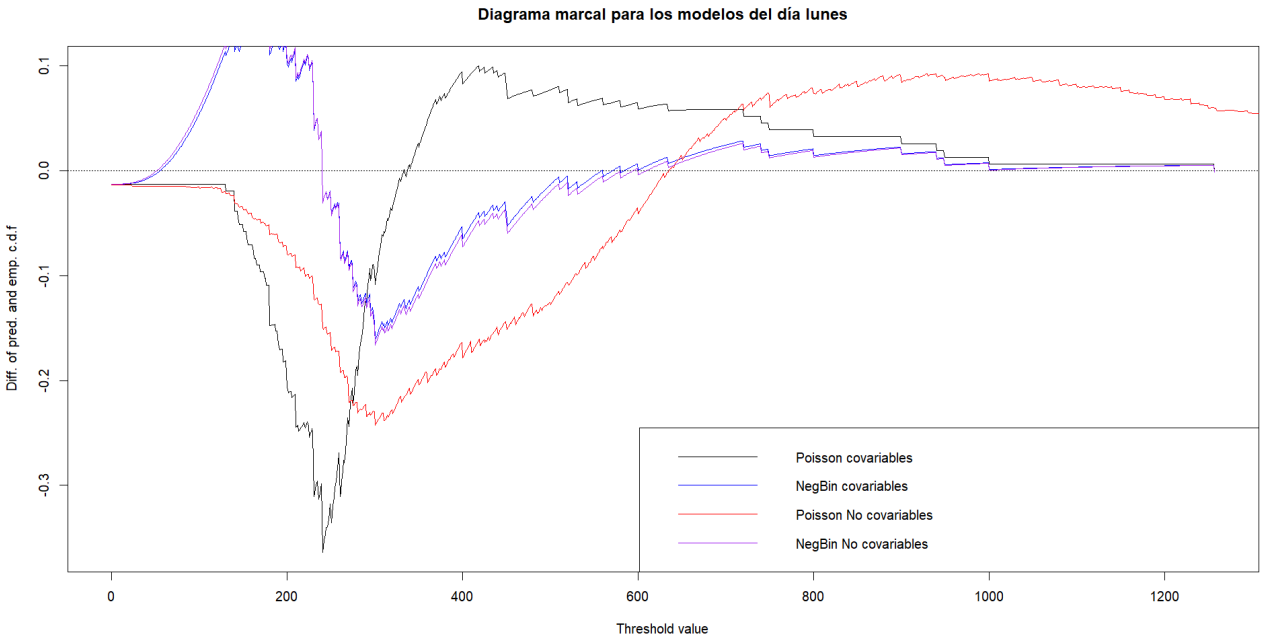


Figura 5.12

Diagramas de calibración marginal de los modelos del día lunes.

Modelos martes				
Modelo	Covariable	MP	OP	AIC
Poisson	Mes	2	4	13429.26
Binomial negativa	Mes	2	8	2043.177
Poisson	No	4	4	22351.02
Binomial negativa	No	4	8	2063.01

Para el modelado del día martes se tienen mejores ajustes cuando se emplean las covariables, siendo más notorio con la distribución de Poisson dado que al usar la covariable mes el AIC se reduce en gran medida, para el modelo binomial negativo no tiene tanta mejora al respecto entre el uso o no de las covariables. En la figura 5.13 se puede notar como el modelo con distribución binomial negativo con covariables es mejor al estar más cercano a la recta discontinua, seguido por el modelo de Poisson con covariables.

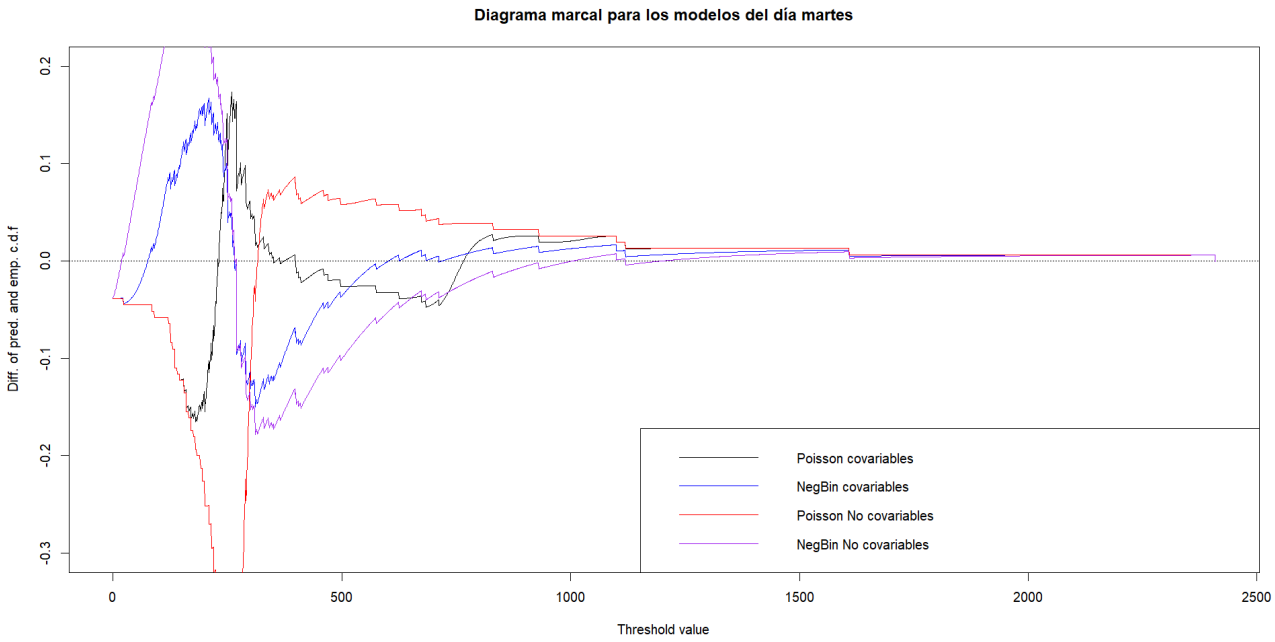


Figura 5.13

Diagramas de calibración marginal para los modelos del día martes con distribución de Poisson y binomial negativa.

Modelos miércoles				
Modelo	Covariable	MP	OP	AIC
Poisson	Mes	2	5	17708.15
Binomial negativa	Mes	2	5	2028.631
Poisson	No	1	6	13802.61
Binomial negativa	No	1	6	1983.716

Para el modelado del día miércoles se tiene que el mejor modelo es el que emplea la distribución binomial negativa, 1 media pasada y 6 observaciones pasadas. Teniendo un AIC apenas 40 unidades menor que el que emplea el uso de covariables, como en los casos anteriores se nota una leve mejora o un leve deterioro en el ajuste para los que emplean binomial negativo y en el caso del modelado de Poisson se observa en que en este caso no mejora al emplear las covariables. Los resultados para la calibración marginal muestran como un mejor modelo el que emplea el modelado de Poisson sin el uso de covariables y de igual manera el modelado sin covariables con distribución binomial negativa, mostrando de esta manera que el uso de covariables no vuelve significativamente mejor el modelo para este día de la semana.

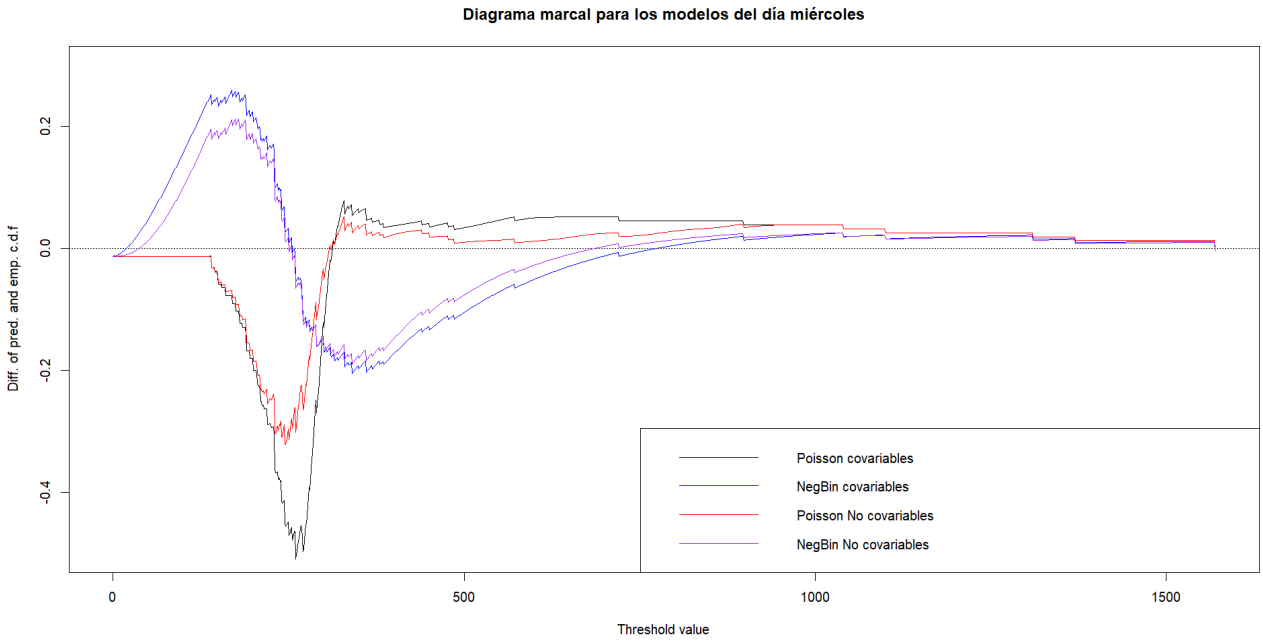


Figura 5.14

Diagramas de calibración marginal de los modelos del día miércoles.

Modelos jueves				
Modelo	Covariable	MP	OP	AIC
Poisson	Mes	1	5	17131.9
Binomial negativa	Mes	1	5	2171.369
Poisson	No	2	6	16167.41
Binomial negativa	No	2	6	2135.14

El día jueves presenta como mejores modelos los que aplican distribución binomial negativa con poca diferencia y con un deterioro en el ajuste al emplear el uso de la covariable mes para el modelado con distribución de Poisson, en caso de emplearse el modelado de Poisson para las predicciones sería mejor emplear el modelo sin covariables. Por otro lado, en la figura 5.15 el peor modelo es el que usa la distribución de Poisson con covariables, mientras que los modelos binomial negativo tienen el mejor ajuste y no hay una notoría diferencia en cual es mejor dado que en ocasiones es mejor el que emplea covariables y en otras ocasiones no tanto.

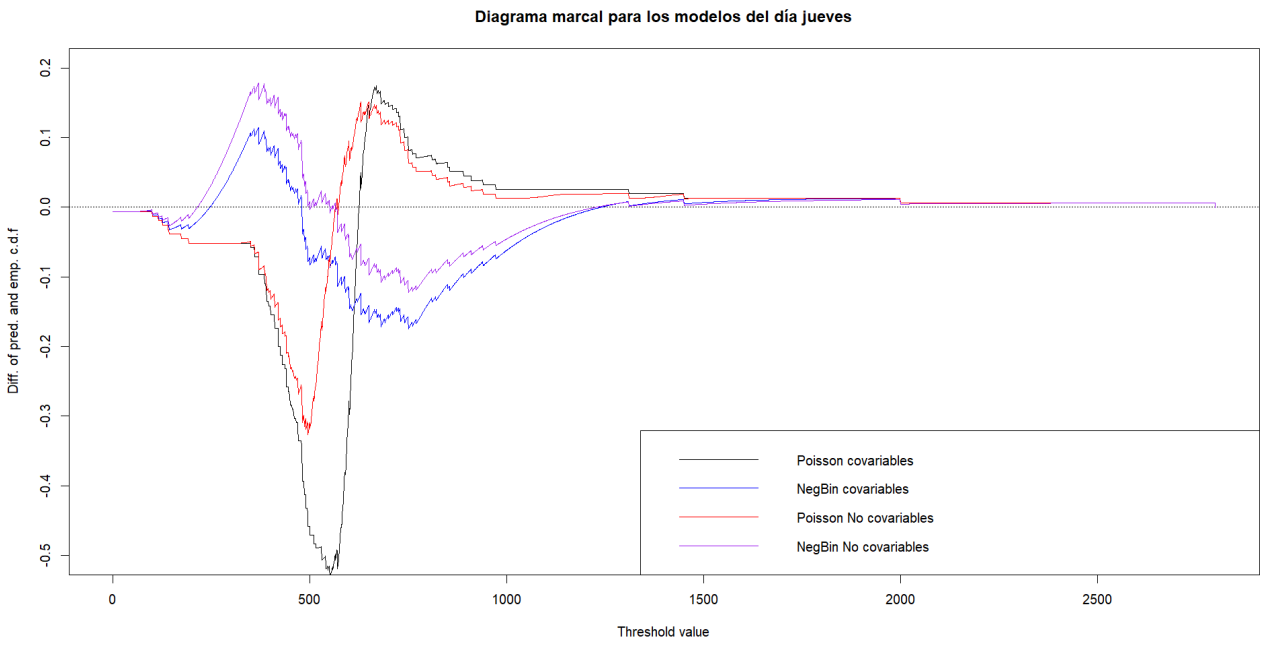


Figura 5.15

Diagramas de calibración marginal de los modelos del día jueves.

Modelos viernes				
Modelo	Covariable	MP	OP	AIC
Poisson	Mes	3	12	11074.71
Binomial negativa	Mes	3	12	2115.235
Poisson	No	3	12	15451.48
Binomial negativa	No	3	12	2119.458

Los resultados del día viernes muestran que el uso de covariables en el modelo de Poisson mejora significativamente el ajuste del modelo y el modelado con distribución binomial negativa tiene una nula mejora empleando las covariables, siendo por muy poco mejor ajuste el modelado binomial negativo sin covariables. Además en la figura 5.16 se evidencia como la curva asociada al modelo de Poisson con covariables tiene un empeoramiento que el modelo sin el uso de covariables no presenta, el modelado con distribución binomial negativa es mejor alternativa a estos modelos y no se observa que el uso o no uso de covariables mejore notoriamente el ajuste.

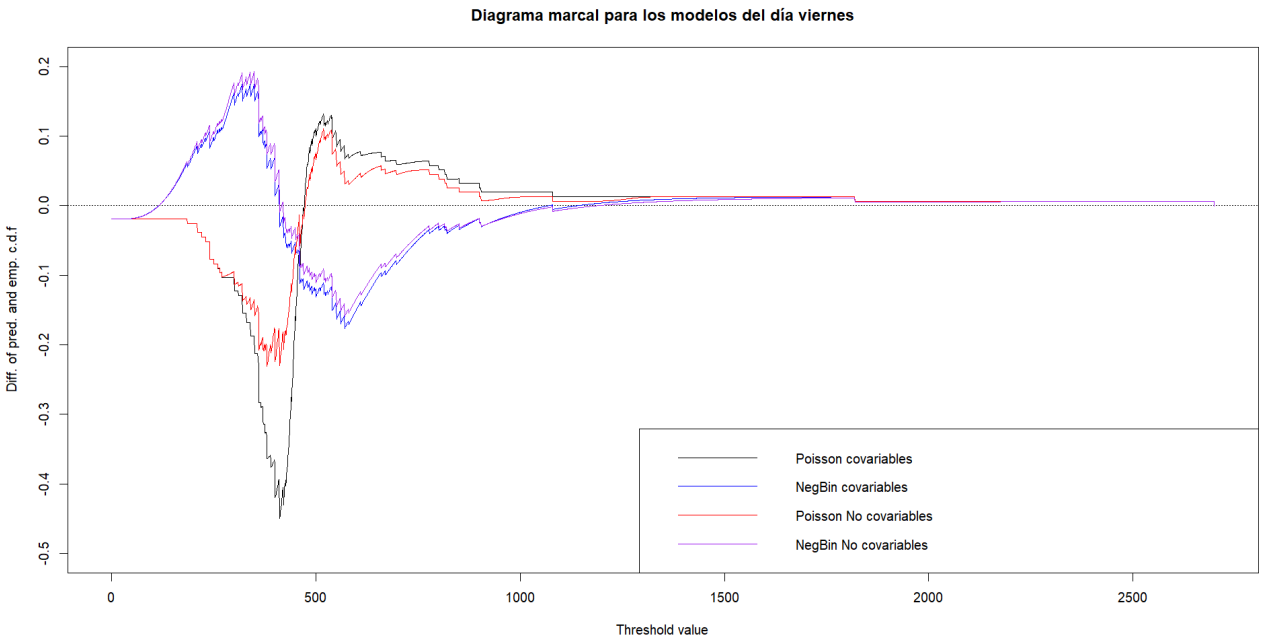


Figura 5.16

Diagramas de calibración marginal de los modelos del día viernes.

Modelos sábado				
Modelo	Covariable	MP	OP	AIC
Poisson	Mes	3	12	15842.87
Binomial negativa	Mes	3	12	2288.487
Poisson	No	4	8	17925.14
Binomial negativa	No	4	8	2288.014

En el día sábado se observa como el modelo con covariables tiene un mejor ajuste a los datos frente modelo que no las emplea, por otro lado se tiene que el ajuste de los modelos con distribución binomial negativa tienen una nula mejora. En la figura 5.17 muestra que es mejor emplear el modelo sin covariables porque el que emplea los meses tiene una notorio empeoramiento y en los modelos binomial negativo se observa como es mejor emplear las covariables para un mejor ajuste.

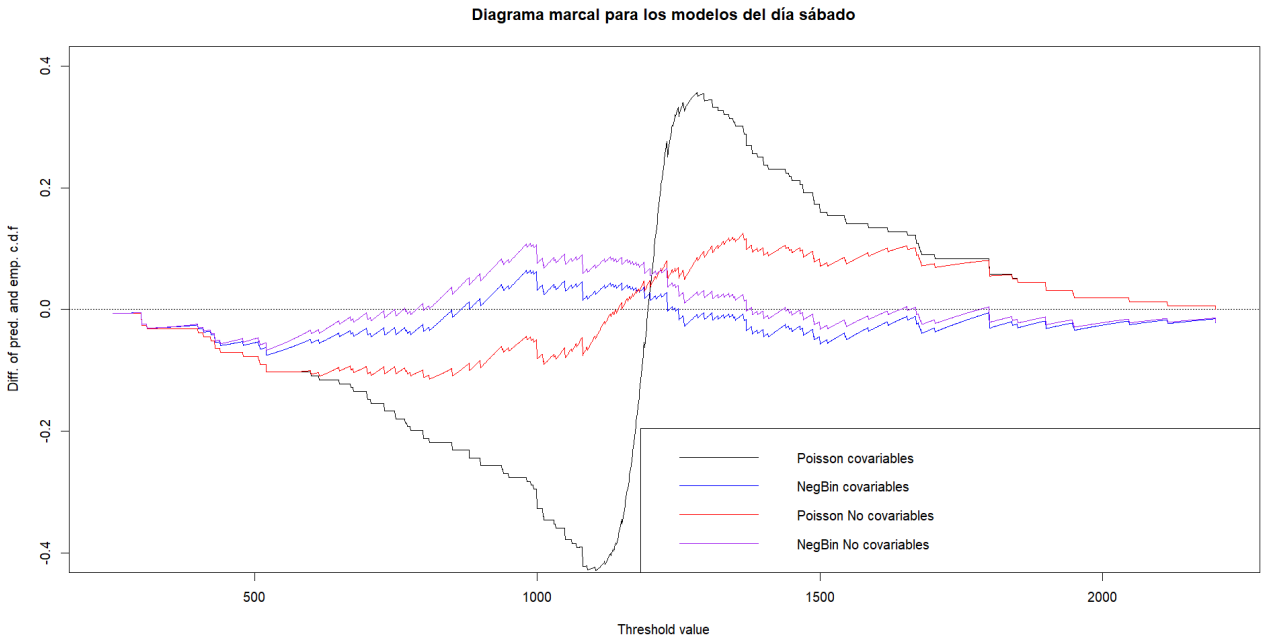


Figura 5.17

Diagramas de calibración marginal de los modelos del día sábado.

Modelos domingo				
Modelo	Covariable	MP	OP	AIC
Poisson	Mes	1	8	7968.084
Binomial negativa	Mes	1	8	2169.605
Poisson	No	2	8	12968.6
Binomial negativa	No	2	8	2247.819

En la tabla asociada a los resultados del día domingo muestra como hay una significativa mejora en el ajuste de los datos según el AIC que obtiene el modelado con distribución de Poisson tener en cuenta la covariable mes, de igual manera por muy pocas unidades el modelo con covariables y distribución de Poisson es mejor. En la figura 5.18 se evidencia como el modelado con distribución Poisson y uso de covariables es notoriamente mejor que el que no las emplea, mientras que en el caso del modelado binomial negativa hay poca diferencia entre un modelo u otro porque no se observa que sea mejor alguno en todo momento del modelo.

5.6. Predicciones para los modelos semanales

Ahora, se hace una comparación entre las predicciones que se obtienen con los diferentes modelos según el día de la semana con la distribución binomial negativa. Donde se resalta la me-

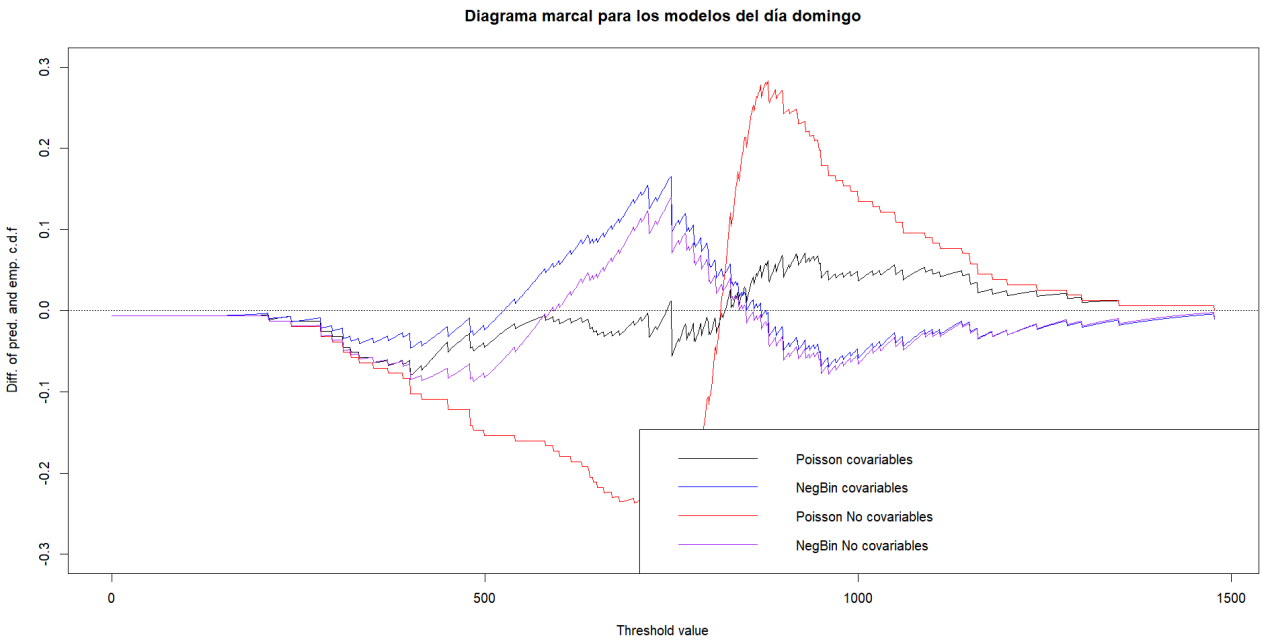


Figura 5.18

Diagramas de calibración marginal de los modelos del día domingo.

jora de los intervalos de predicción que ya no tienen tanta variabilidad como sucede en el primer modelado donde se toma $t = 1$ día, dado que el rendimiento de días con poca venta como los lunes, martes y miércoles aumenta la variación al tener valores más altos como lo son días sábado y domingo. Se solucionan problemas como las predicciones más realistas del negocio, donde estas se acercan mucho más a lo esperado cada día de la semana a excepción de algunos días como se ve a continuación.

En las predicciones que se observan en la figura 5.19 se obtiene para el modelo del día lunes con distribución binomial negativa mejora mucho respecto al modelo inicial de días de la semana dado que gran parte del comportamiento de este día se ve en predicciones de alrededor de 220 aves con mínimos y máximos de 340 que son los intervalos donde hay más frecuencia según lo mostrado en el diagrama de barras apiladas de la figura 5.9.

La predicción del modelo para el día martes de la figura 5.20 no tuvo el resultado esperado a pesar de que se implementaron varios valores para las medias condicionales y observaciones pasadas con sacrificios de 70 a 230 aves, teniendo en cuenta que los meses que esta prediciendo son noviembre, diciembre y enero que son meses donde se evidencia una mayor presencia de pedidos superiores a 241 aves.

La predicción del modelo dada en la figura 5.21 para el día miércoles dado en la figura 5.10 muestra una gran variabilidad dentro de su banda de confianza, sin embargo da un número promedio de aves a sacrificar que se ve similar al comportamiento que tiene durante los registros de

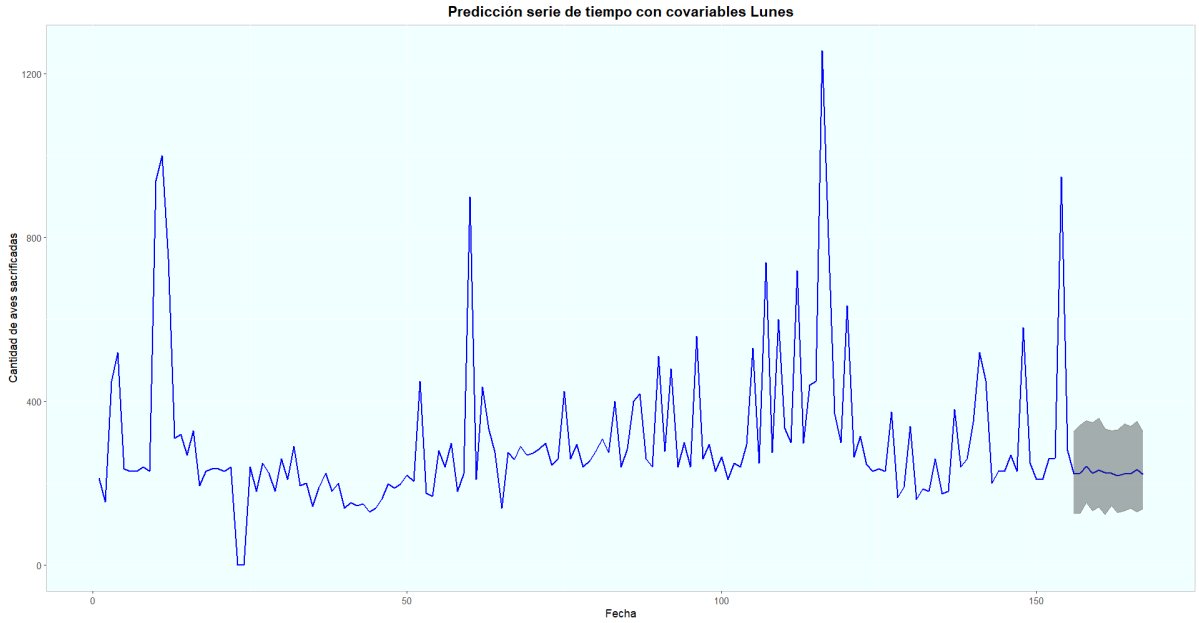


Figura 5.19

Serie de tiempo día lunes con predicciones modelo BN.

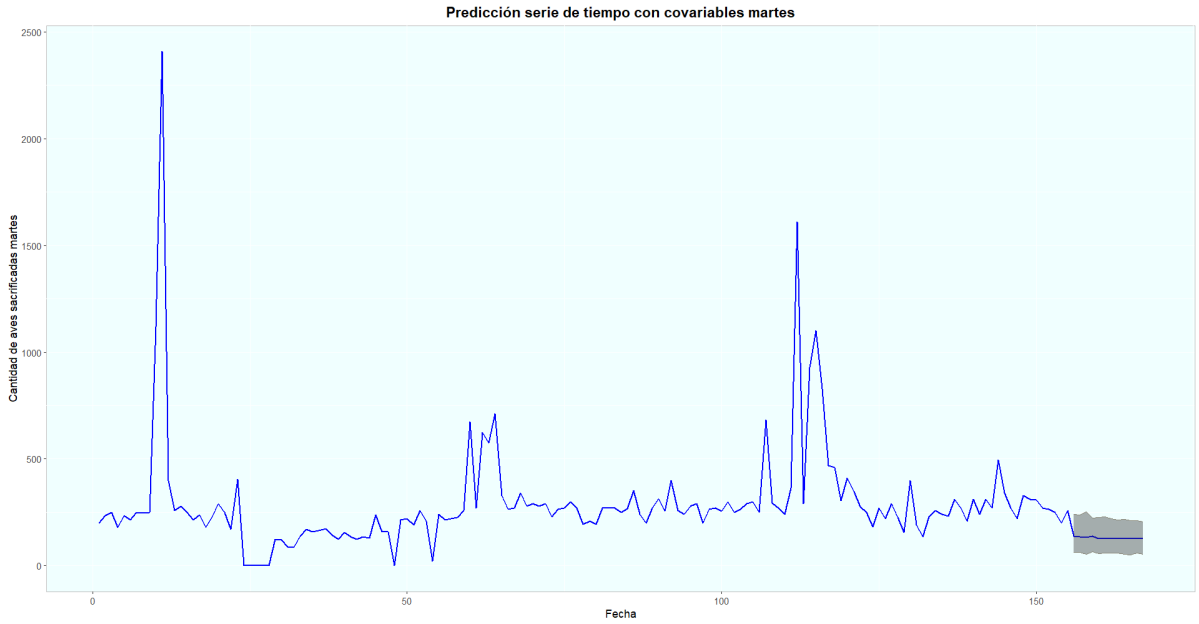


Figura 5.20

Serie de tiempo día martes con predicciones modelo BN.

la serie de tiempo, puesto que en la figura 5.10 se observa que esta dentro de los valores que tienen mayor frecuencia.

El día jueves muestra la figura 5.22 unas predicciones donde la cantidad de aves irá en aumento

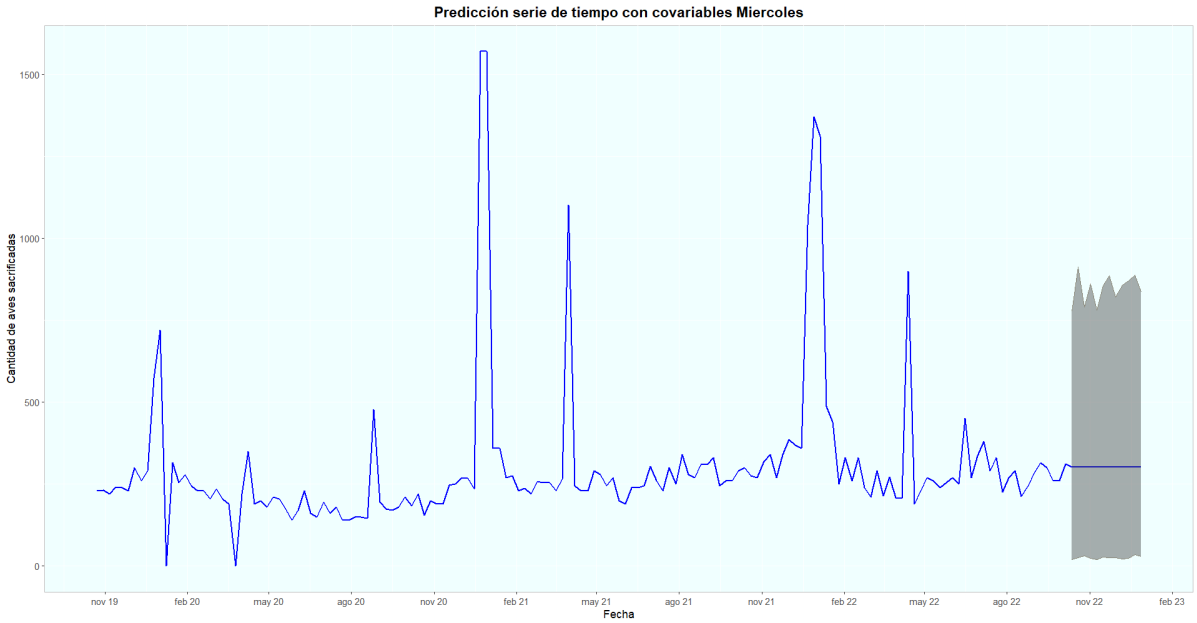


Figura 5.21

Serie de tiempo día miércoles con predicciones modelo BN.

en comparación a los mostrados en la serie de tiempo, dado que durante los tiempos donde la venta decreció fue por la suspensión de la ruta del Municipio 3 dado que las ganancias que dejaba esta ruta eran muy pocas para el proceso que se debía realizar.

La predicción que realiza el modelo del día viernes que muestra la figura 5.23 se acerca más a los valores que se muestran en el histograma de la figura 5.11 con predicciones de alrededor de 350 aves y límites superiores 560 aves e inferiores de 200 aves.

El día sábado presenta una predicción con límites inferiores de 720 y límites superiores de alrededor 1360 aves con predicciones de casi 1000 aves, un comportamiento similar al que se tiene en el diagrama de barras apiladas 5.11, sin embargo no se logra en los modelos tener con certeza algunos días que parecen tener los valores más altos en cada día de la semana.

Por último las predicciones del día domingo 5.25 tienen una gran variabilidad dentro de la banda de confianza pero tambien tienen unos valores adecuados para lo mostrado en el diagrama de barras apiladas 5.11, en estas semanas muestra como si existieran algunas disminuciones en las ventas.

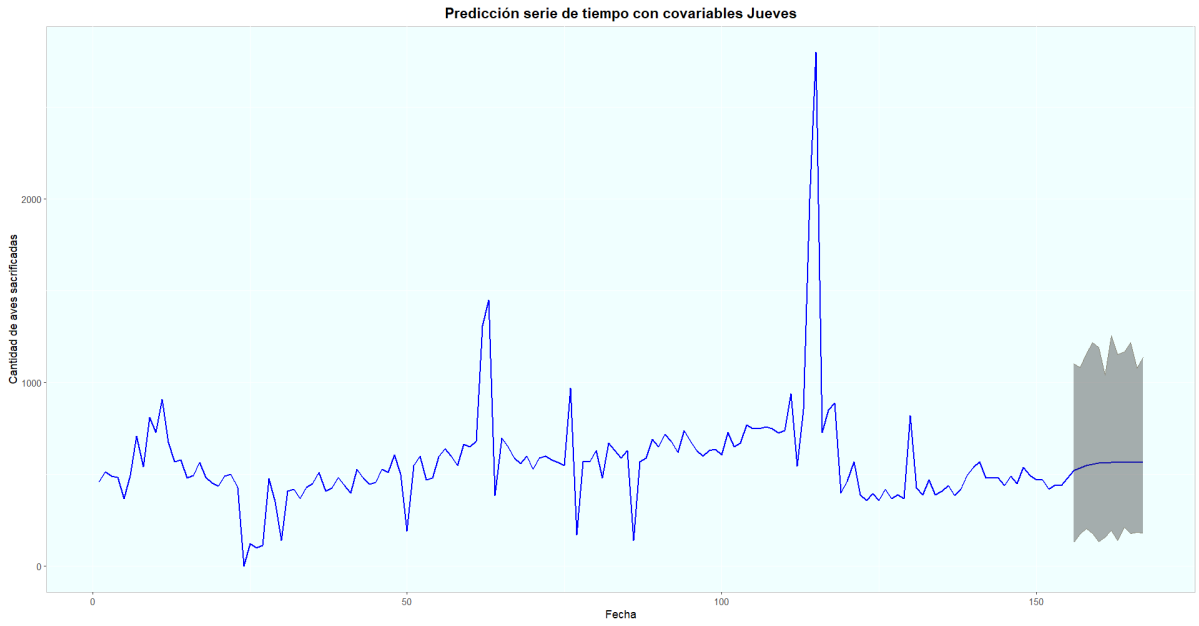


Figura 5.22

Serie de tiempo día jueves con predicciones modelo BN.

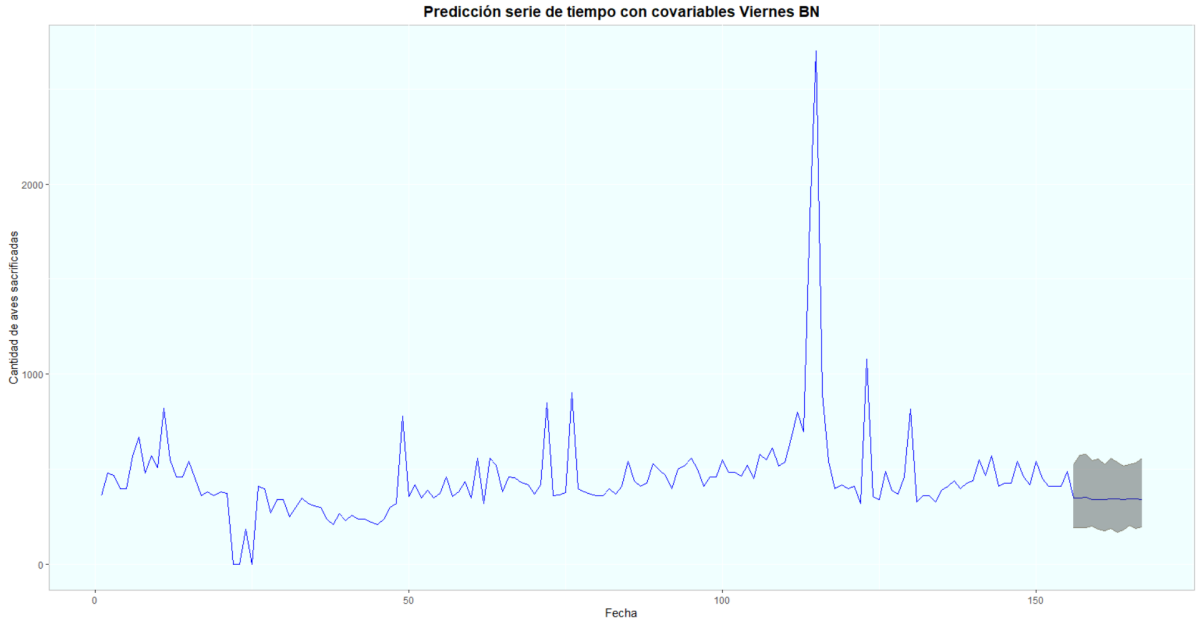


Figura 5.23

Serie de tiempo día viernes con predicciones modelo BN.

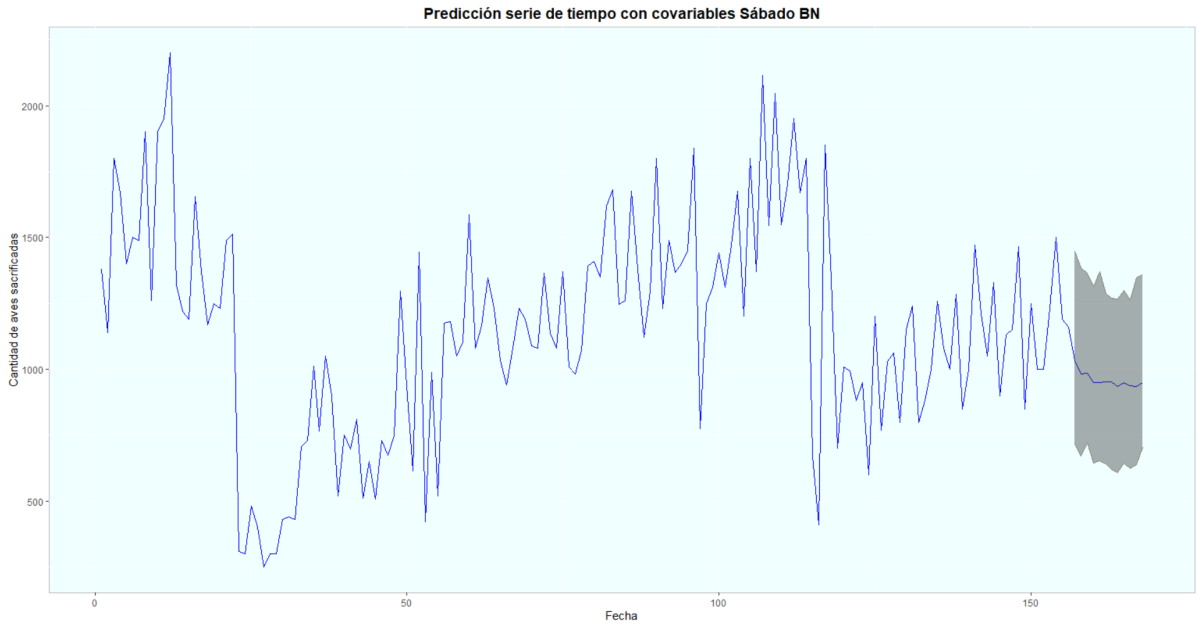


Figura 5.24

Serie de tiempo día sábado con predicciones modelo BN.

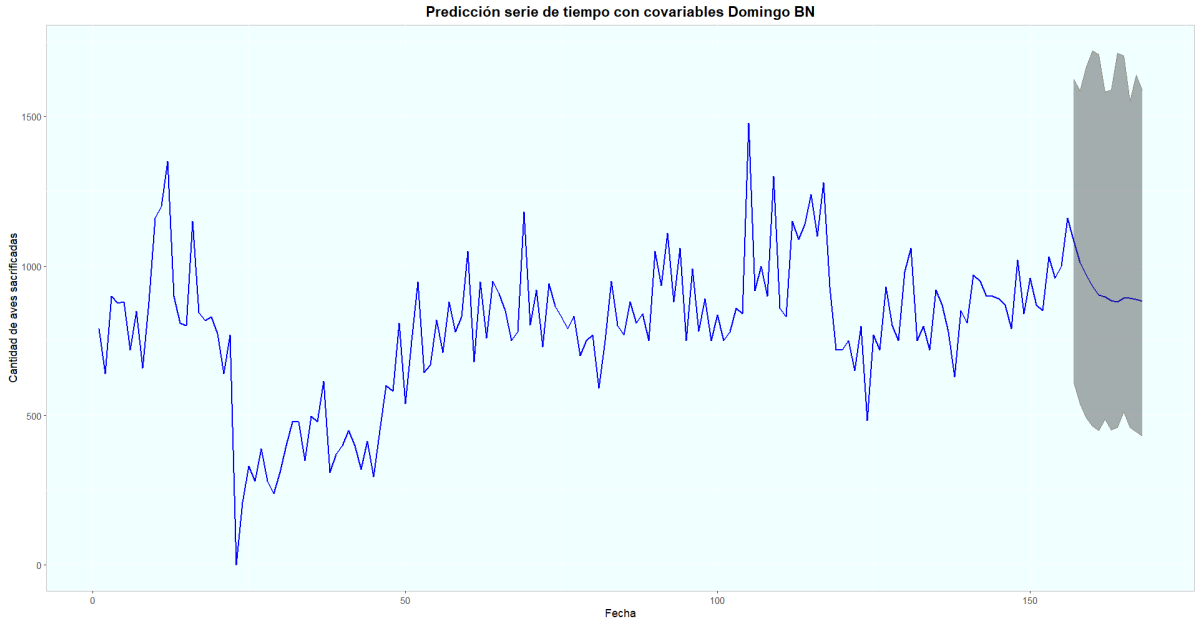


Figura 5.25

Serie de tiempo día domingo con predicciones modelo BN.

6. Conclusiones

• En la serie de tiempo y las predicciones, existe una gran variabilidad de los datos día a día, la solución de modelar de manera semanal resulta mejor al momento de tener una estimación algo más realista y con menos variabilidad al primer modelado que se presenta.

• Se observo que existe una tendencia a disminuir la cantidad de aves. El modelo no cuenta con la covariable del costo del ave sacrificado que últimamente ha venido en aumento dada la inflación el precio de mantenimiento de estas aves para una granja avícola, teniendo en cuenta que los bultos de purina, el huevo, así como el transporte de los mismos ha ido en aumento. Esto perjudica el precio final del ave y hace que la gente opte por comprar otro producto o directamente pasar de comprar la gallina.

• Este modelado plantea una posible solución frente al problema que presenta el negocio relacionado a las ocasiones donde la estimación por experiencia del administrador puede terminar con aves congeladas. Al tener en cuenta el modelo semanal se puede minimizar pérdidas en las ventas.

• El modelo con distribución binomial negativa es una mejor opción al momento de modelar estos datos debido a los supuestos que tiene la misma y cambiando muy poco realmente cuando se emplean las covariables, este modelo se resalta frente al modelado de Poisson almenos en esta aplicación en particular.

• El modelado con distribución de Poisson presenta un peor ajuste sin embargo este puede mejorar notablemente si se emplean las covariables en algunos casos.

• El uso de otras covariables como pueden ser el peso del ave y su precio pueden mejorar el modelo, así como fechas exactas en las que se observa un aumento inesperado en las ventas sin importar que día de la semana sea dado que son conocidos solo el 24 y 31 de diciembre sin embargo parece existir más días donde tienden a presentarse cantidades de aves mayores que por limitantes en los datos no se pudieron contemplar.

Referencias

Ahmad, A., y Francq, C. (2016). Poisson qmle of count time series models. *Journal of Time Series Analysis*, *37*(3), 291–314.

Anderson, D., y Burnham, K. (2004). Model selection and multi-model inference. *Second. NY: Springer-Verlag*, *63*(2020), 10.

Blanco Castañeda, L. (2013). Probabilidad. *Editorial UN*.

Box, G. E., Jenkins, G. M., y Bacon, D. W. (1967). *Models for forecasting seasonal and non-seasonal time series*. University of Wisconsin–Madison, Department of Statistics.

Brockwell, P. J., y Davis, R. A. (2009). *Time series: theory and methods*. Springer science & business media.

Burden, R. L., Faires, J. D., y Burden, A. M. (2015). *Numerical analysis*. Cengage learning.

Casella, G., y Berger, R. L. (2021). *Statistical inference*. Cengage Learning.

Chambers, J. M., y Hastie, T. J. (1992). Statistical models in s., wadsworth & brooks/cole. *Pacific Grove, CA*.

Christou, V., y Fokianos, K. (2014). Quasi-likelihood inference for negative binomial time series models. *Journal of Time Series Analysis*, *35*(1), 55–78.

Dekking, F. M., Kraaikamp, C., Lopuhaä, H. P., y Meester, L. E. (2005). *A modern introduction to probability and statistics: Understanding why and how* (Vol. 488). Springer.

Doukhan, P., Fokianos, K., y Tjøstheim, D. (2012). On weak dependence conditions for poisson autoregressions. *Statistics & Probability Letters*, *82*(5), 942–948.

Ferland, R., Latour, A., y Oraichi, D. (2006). Integer-valued garch process. *Journal of time series analysis*, *27*(6), 923–942.

Gravetter, F. J., Wallnau, L. B., Forzano, L.-A. B., y Witnauer, J. E. (2020). *Essentials of statistics for the behavioral sciences*. Cengage Learning.

Jacod, J., y Protter, P. (2004). *Probability essentials*. Springer Science & Business Media.

Lander, J. P. (2014). *R for everyone: Advanced analytics and graphics*. Pearson Education.

Lange, K., Chambers, J., y Eddy, W. (2010). *Numerical analysis for statisticians* (Vol. 1). Springer.

Liboschik, T., Fokianos, K., y Fried, R. (2017). tscount: An R package for analysis of count time series following generalized linear models. *Journal of Statistical Software*, *82*, 1–51.

Liboschik, T., Kerschke, P., Fokianos, K., y Fried, R. (2016). Modelling interventions in ingarch processes. *International Journal of Computer Mathematics*, *93*(4), 640–657.

Moritz, S., Gatscha, S., Wang, E., Hause, R., Moritz, M. S., y ByteCompile, T. (2019). Package 'imputets'. *cran. r-project. org*.

Ríos, A. (2019). *Modelos epidemiológicos estocásticos y su inferencia: casos SIS y SEIR* (Tesis de Master no publicada). Universidad Nacional de Colombia, Bogotá D.C..

Sutherland, I. (1963). John graunt: A tercentenary tribute. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *126*(4), 537–556.

Tjøstheim, D. (2015). Count time series with observation-driven autoregressive parameter dynamics. *Handbook of Discrete-Valued Time Series, Handbooks of Modern Statistical Methods*, 77–100.

Venables, W. N., y Ripley, B. D. (2013). *Modern applied statistics with s-plus*. Springer Science & Business Media.