Modelos GAMLSS
Consumo de cerveza en Brasil

Daniela Gómez, Mateo Tuberquia, Brayan Pérez
Universidad Nacional de Colombia, Sede Medellín
dgomezbe@unal.edu.co,mtuberquiag@unal.edu.co,beperezm@unal.edu.co

Resumen

La cerveza es una de las bebidas más antiguas de la humanidad, ha ido evolucionando y poco a poco se ha convertido en unas de las bebidas alcohólicas más consumidas a nivel mundial, es elaborada a partir de azúcares obtenidas de cereales y otros granos, saborizada y aromatizada con lúpulo, que luego son fermentados en agua con levaduras; es consumida por todo tipo de personas mayores de edad en diferentes horas del día.

La cerveza es la bebida alcohólica más consumida en Brasil, sobre todo los días en los que la temperatura sube, siendo este el tercer país con mayor producción de tal bebida en el mundo. A continuación presentamos el uso de modelos GAMLSS (Generalized Additive Model for Location Scale and Shape) para determinar la influencia de algunas medidas meteorológicas sobre el consumo promedio de cerveza en nuestra población objetivo; estos modelos nos permiten modelar los parámetros de la distribución y ajustar un modelo no lineal haciendo uso del paquete "gamlss" de R.

Introducción

En este trabajo se analizan los datos sobre la medición en litros del consumo de cerveza en un área universitaria donde hay algunas fiestas con grupos de estudiantes mayores de edad, con el objetivo de estudiar qué tan significativas pueden ser las medidas de la temperatura del ambiente, la precipitación meteorológica y si es fin de semana o no.
La base de datos fue obtenida del sitio web "kaggle", contiene 365 observaciones que corresponden a un periodo de un año, seis covariables, una de ellas cualitativa, y una variable respuesta. Los datos se recopilaron en Sao Paulo-Brasil, en una población de 18 a 28 años de edad (en promedio) en el año 2015.

Análisis descriptivo

Las variables que están contenidas en nuestra base de datos son las siguientes:

- **ConsumoB:** Consumo en litros de cerveza por persona en un año (Variable respuesta).
- **TmpMed:** Temperatura promedio por día (12.9 C- 28.86 C)
- **TmpMin:** Temperatura mínima por día (10.6 C- 24.5C)
- **TmpMax:** Temperatura máxima por día (14.5 C - 36.5 C)
- **Precipitación:** Precipitación pluvial (0 mm – 94.8 mm)
- **Data:** Fecha en la que fue tomada la observación (2015/01/01 – 2015/12/31)
- **Finde:** Fecha correspondiente a un fin de semana (variable cualitativa)

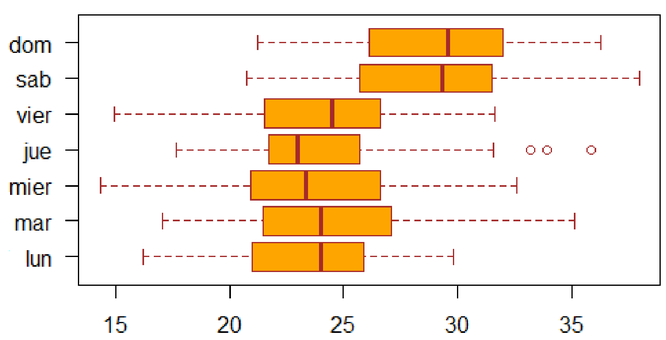


Figura 1

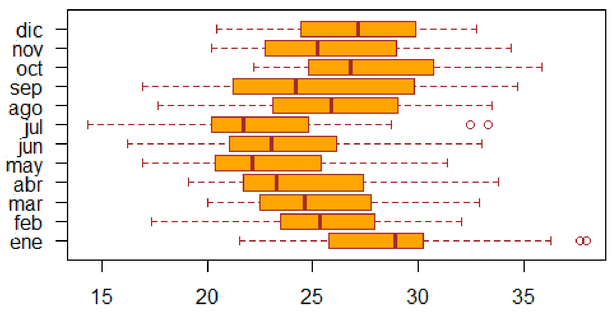


Figura 2

En la figura 1 se muestra un diagrama de caja del consumo de cerveza por día, donde observamos que el mayor consumo se da los días sábado y domingo.
En la figura 2 se muestra un diagrama de caja del consumo de cerveza dependiendo del mes. Se observa que el mes en el que hay mayor consumo es en enero y en donde hay menor es en julio, con una media de 29 litros aproximadamente y 23 litros aproximadamente, respectivamente

Ajustando un modelo lineal

Ajustamos dos modelos lineales, uno para fin de semana y otro para semana, relacionando temperatura máxima con el consumo de cerveza, donde 1 corresponde al fin de semana y 0 a día de la semana.

x_6:Temperatura máxima
$\hat{y}$:Consumo de cerveza

$\hat{y} = 11{,}656 + 0{,}655 * X_6$ para fin semana
$\hat{y} = 5{,}6823 + 0{,}685 * X_6$ para semana

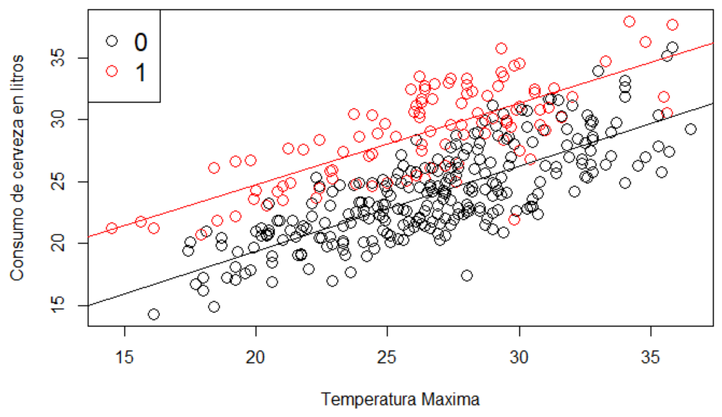


Figura 3

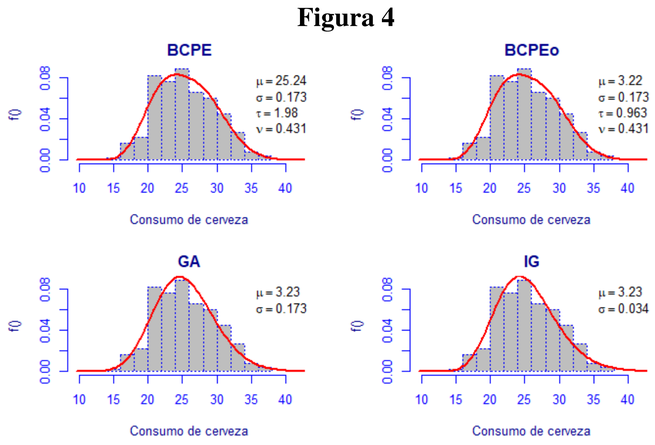
En la figura 3 se muestra un gráfico de dispersión con los modelos lineales ajustados. Se observa una tendencia lineal en el consumo de cerveza con respecto a la temperatura máxima, es decir, a medida que la temperatura máxima aumenta, el consumo de cerveza aumenta.

Ajustando el modelo Gamlss

	Correlacion	MSE	BIC	R2 adj	AIC
BCPE	0.8494	**5.3875**	1617.08	0.7765	**1578.09**
BCPEo	0.8455	5.5136	1651.61	0.7583	1608.71
GA	0.8456	5.4910	1676.75	0.7283	1649.45
IG	0.8459	5.5135	1674.25	0.7247	1650.84
LOGNO2	0.8459	5.5010	1674.15	0.7260	1650.75

Mejores Ajustes

Figura 4



En la figura 4 tenemos cuatro histrogramas con los ajustes de cada una de las mejores distribuciones seleccionadas según el menor BIC y menor AIC

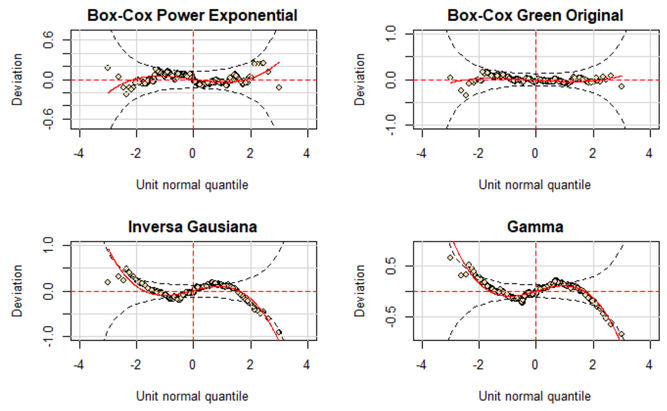


Figura 5

Este gráfico es una herramienta de diagnóstico para verificar los residuos dentro de los diferentes rangos, donde se evidencia fallas de normalidad en los errores ajustados para las distribuciones Inversa Gaussiana(IG) y Gamma(GA), por lo cual se descartan estas distribuciones para el modelo final.

Ajuste del modelo final

$\hat{\mu} = 6{,}018746 + 0{,}683263 * (X1) + 4{,}841118 * (X2) - 0{,}057524 * (X3)$

- X1: TmpMax.
- X2: Finde
- X3: Precipitacion

Resultados

Empleando el método de eliminación hacia adelante obtuvimos que las covariables que mejor explican la variable respuesta (consumo de cerveza) son: La temperatura máxima (TmpMax), la precipitación (Precipitación) y el fin de semana (Finde). Es razonable que las covariables temperatura mínima (TmpMin), temperatura media (TmpMed) no sean significativas a la hora de predecir el consumo de cerveza.
De todos los anteriores modelos se seleccionó el que asume la distribución BCPE (Box Cox Power Exponential), los criterios para elegir el modelo fueron la correlación, el error cuadrático medio, el AIC, el BIC y el coeficiente de correlación.

Conclusiones

La cerveza es una bebida alcohólica con un alto consumo en el mercado, es perfecto para cualquier situación independiente de la hora o el día; sin embargo, se logró ver que el consumo aumenta a media que la temperatura aumenta, pues las personas la toman como una bebida refrescante, los días que son fin de semana también hay mayor consumo ya que estos días se suele salir de festividad.
Después de hacer un análisis de los datos concluimos que la mejor distribución de probabilidad que se le ajusta es la Box Cox Power Exponential (BCPE), por este hecho el modelo GAMLSS de esta familia es el que más nos facilita la predicción del consumo de cerveza.

Referencias

Alexandre George Lustosa (2018). Beer Consumption – Sao Paulo Dataset. Kaggle. Estraído de : https://www.kaggle.com

Semana (2017). "Siete cosas que debe saber sobre la cerveza". Extraído de: https://www.semana.com

Viajando en Brasil(2018). Estraído de: https://viajandoenbrasil.com

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Ian Howson(2019), "wp: Worm plot". Estraído de: https://rdrr.io/cran/gamlss/man/wp.html