Compañía de Cervecerías Unidas CCU
Gerencia Inteligencia de Mercado

Auditoría Analítica de Modelos Econométricos para la Proyección de Venta Volumen de la Industria de Cervezas y Gaseosas

M. Ignacia Vicuña - Cristian Vásquez

Resumen

En este informe se entregan los principales resultados de la auditoría analítica que se hizo sobre los dos modelos econométricos desarrollados por CCU para la proyección de venta volumen de la industria de cervezas y gaseosas en Chile. Mediante análisis estadísticos que son presentados en todo el documento se identifica el no cumplimiento de suspuestos econométricos para los dos modelos auditados, específicamente problemas de colinealidad entre las variables predictoras. Al finalizar el informe se entrega un detalle con las propuestas metodológicas para el tratamiento de los problemas identificados. [1].

Introducción

La Gerencia de Inteligencia de Mercado de CCU desarrolló dos modelos econométricos para proyectar la venta volumen de la industria de cervezas y gaseosas en Chile. Ambos modelos están construídos en base a registros mensuales y consideran como variable objetivo la venta volumen de la industria, que se calcula como la división de Market Share CCU *(Fuente: Nielsen)* y los Volúmenes CCU *(Fuente: Sell In CCU)*. Las dimensiones utilizadas como variables independientes para explicar la venta volumen son:

- Temperatura Máxima: Corresponde al promedio mensual de las temperaturas máximas (en °C) de la Región Metropolitana. *(Fuente: Accuweather)*.
- Precio Categoría: Precio por litro ponderado por categoría. *(Fuente: Nielsen Canales Supermercado y Compra Nielsen)*.
- Tasa de Desempleo: Tasa de desocupación nacional. *(Fuente: INE)*.
- Retiros AFP: Variable de creación interna en CCU para representar el impacto de los tres retiros de las AFP. *(Fuente: Interna)*
  - Toma el valor 1 en Agosto 2020, Septiembre 2020, Octubre 2020, Diciembre 2020, Enero 2021, Mayo 2021, Junio 2021 y Julio 2021.
  - Toma el valor 0.5 en Noviembre 2020, Febrero 2020 y Agosto 2021.
  - Toma el valor 0 en todos los otros meses.
- Índice de Movilidad: Corresponde al promedio mensual del % de movilidad de las 6 categorías regristradas: retail and recreation, grocery and pharmacy, parks, transit stations,workplaces, residential . *(Fuente: Google)*

[1] Última edición: 26 de Noviembre, 2021

Los dos modelos desarrollados utilizan la técnica estadística de "Modelos de regresión lineal con errores SARIMA" donde las dimensiones mencionadas anterioremente explican la venta volumen de la industria de manera lineal para cada categoría (cervezas y gaseosas), y los errores del modelo de regresión lineal son ajustados mediante un modelo SARIMA$(p,d,q)(P,D,Q)_s$:

$$Y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \ldots \beta_p x_{tp} + \varepsilon_t$$

$$\Phi_P(B^s)\phi_p(B)(1-B^s)^D(1-B)^d\varepsilon_t = \Theta_Q(B^s)\theta_q(B)u_t.$$

La variable objetivo Y representa la venta volumen mensual de la industria de la categoría, el subíndice t es la secuencia correlativa con los meses, los parámetros β son desconocidos y las variables $x's$ son las dimensiones consideradas para proyectar la venta volumen de la industria. El error aleatorio del modelo de regresión lineal, es representado por ε_t el cual es modelado por un ARIMA estacional. Los parámetros de la componente no estacional, dados por (p,d,q), representan: p el grado autoregresivo, q el grado de medias móviles y d el grado de diferenciación. Por otra parte, los parámetros de la componente estacional, dados por $(P,D,Q)_s$, representan: P el grado autoregresivo estacional, Q el grado de medias móviles estacional, D el grado de diferenciación estacional y s el período estacional. Las innovaciones u_t son incorrelacionadas entre sí con distribución normal de media cero y varianza constante. Para mayor detalle sobre los modelos Arima estacional ver página 203 [1].

El objetivo es estimar los parámetros desconocidos β del modelo de regresión lineal y los parámetros asociados a la parte SARIMA para poder realizar las proyecciones de las ventas volumen de la industria.

Para la estimación del modelo, la Gerencia de Inteligencia de Mercado utilizó la función `auto.arima( )` de la librería `forescast` del software R. Esta función hace una estimación bietápica. Primero estima los coeficientes de regresión mediante "mínimos cuadrados ordinarios" . Posteriormente a los errores del modelo ajusta modelos SARIMA y escoge aquel que tenga menor AIC. Para consultar detalles de la función ver [2].

El documento se encuentra elaborado de la siguiente manera: En la primera sección, se encuentra una descripción general de las principales variables utilizadas en cada uno de los modelos econométricos. En la segunda sección, se realiza un análisis bivariado de cada una de las variables independientes con la variable objetivo (cervezas o gaseosas), se presentan gráficos de dispersión y el grado de correlación de cada dimensión con la venta volumen de la industria. En la tercera sección, se presenta un análisis multivariado de las dimensiones de cada modelo, con el objetivo de identificar posibles problemas de información redundante. En la cuarta sección, se revisan los supuestos econométricos que deben cumplir estos modelos. En la sección 5 se presenta un resumen y conclusiones de la auditoría y en la última sección se presentan las propuestas de mejoras. Todos los análisis presentados en el documentos se hicieron con el software R, principalmente utilizando el IDE (entorno de desarrollo integrado) RStudio.

1. Análisis Descriptivo de la Información

En esta sección del documento se revisa la información entregada por CCU para la construcción de los dos modelos analíticos.

La variable objetivo en cada uno de los modelos es la venta volumen de la industria, de cervezas para el primero y gaseosas para el segundo. Para el modelo de cervezas, esta información se encuentra disponible mensualmente desde Enero 2014 hasta Agosto 2021 y para el modelo de gaseosas desde Enero 2015 hasta Agosto 2021. Las

Figuras 1 y 2 presentan los gráficos de las ventas mensuales de volumen de la industria de cervezas y gaseosas respectivamente, junto con los histogramas de distribuciones.

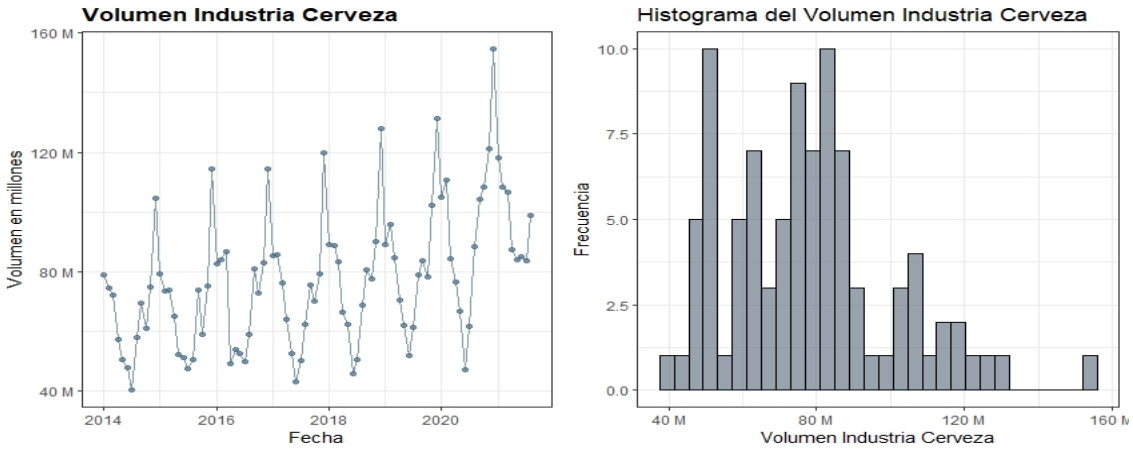

Figura 1: Venta volumen de la industria cervezas

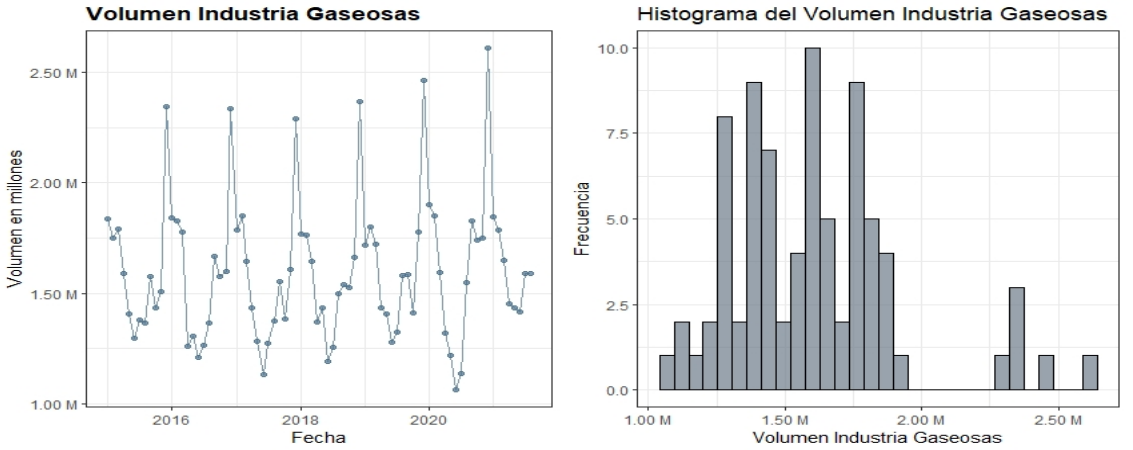

Figura 2: Venta volumen de la industria gaseosas

Todos los datos se encuentran completos en las fechas indicadas. A partir de los gráficos se aprecia un comportamiento estacional (con un patrón anual) en las industrias de cervezas y gaseosas. El volumen de venta en la industria de cerveza presenta un tendencia positiva, algo que no ocurre de forma clara con el volumen de venta de la industria gaseosas. Los valores de las distribuciones se encuentran dentro del rango esperado según las tendencias, y los valores extremos observados en las distribuiciones corresponden a los máximos anuales.

A continuación se presenta un reporte con las principales estadísticas descriptivas de las variables objetivos:

```
#====== 1.1 Estadísticas Descriptivas Industria Cervezas =========#

skimr::skim(select(datoscer,CERVEZAS))

-- Variable type: numeric -------------------------------------------------------
# A tibble: 1 x 10
  skim_variable n_missing complete_rate      mean        sd        p0       p25
* <chr>             <int>         <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
1 CERVEZAS              0             1 77794404. 22694205. 40283324. 61139913.
        p50       p75       p100
*     <dbl>     <dbl>      <dbl>
1 76378925. 87667648. 154766087.

#====== 1.1 Estadísticas Descriptivas Industria Gaseosas =========#

skimr::skim(select(datosgas,GASEOSAS))

-- Variable type: numeric -------------------------------------------------------
# A tibble: 1 x 10
  skim_variable n_missing complete_rate     mean      sd       p0      p25
* <chr>             <int>         <dbl>    <dbl>   <dbl>    <dbl>    <dbl>
1 GASEOSAS              0             1 1596755. 310277. 1062442. 1378451.
       p50      p75     p100
*    <dbl>    <dbl>    <dbl>
1 1578016. 1769254. 2610418.
```

Como se puede ver del cuadro, toda la información se encuentra completa. A diferencia de los gráficos, en las estadísticas descriptivas, las escalas de las variables son las informadas por CCU para tener una mayor sensibilidad de los números. Se observa que el volumen de venta (por litro) promedio es de 77.794.404 para la industria de cervezas y de 1.596.755 para gaseosas.

Para la dimensión Precio Categoría se realiza un reporte descriptivo similar. Las Figuras 3 y 4 contienen los gráficos de tendencia mensual del precio por litro de la industria de cervezas y gaseosas, respectivamente, junto a sus histogramas.

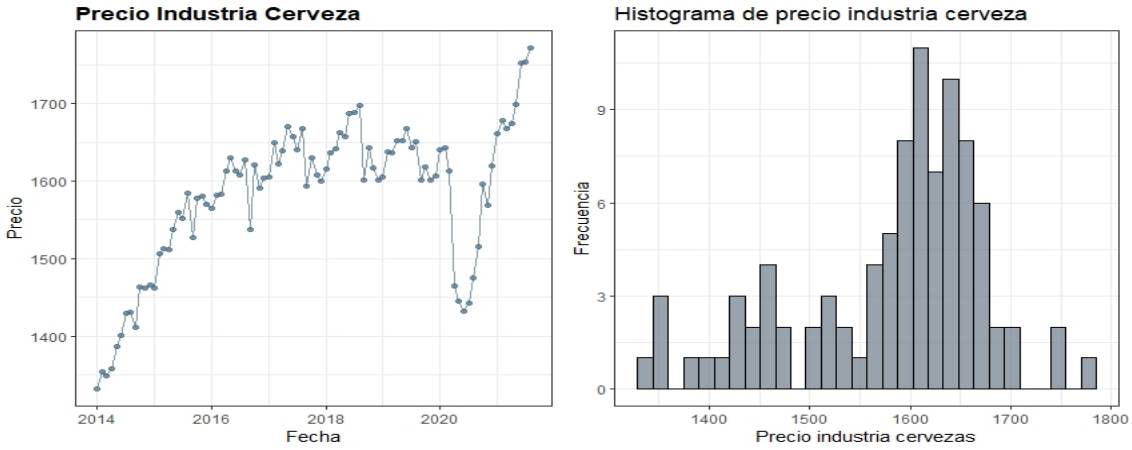

Figura 3: Precio Industria Cervezas

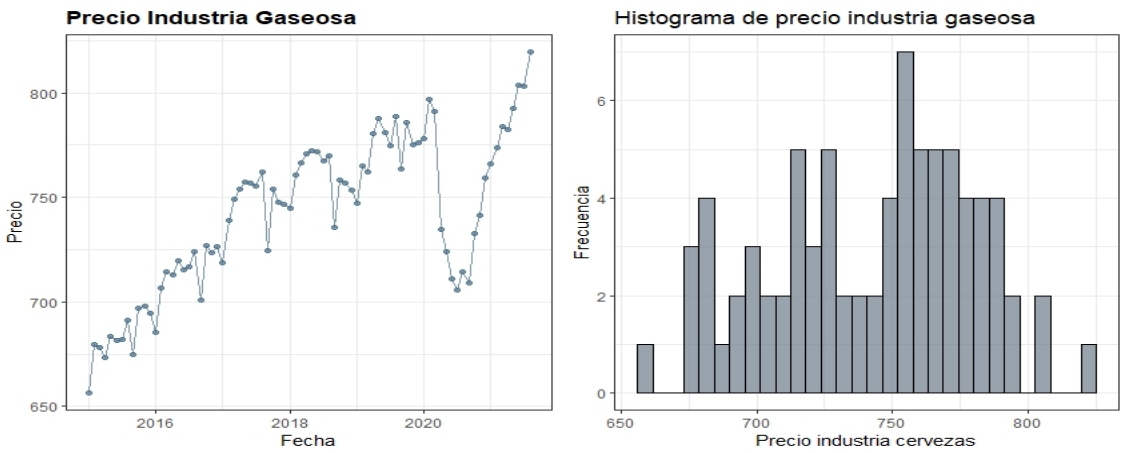

Figura 4: Precio Industria Gaseosas

Ambas variables se encuentran disponibles con datos completos, se tiene más historia de información para la industria de la cerveza, y las distribuciones se encuentran dentro de los valores admisibles según la escala de medición. Se observa un cambio estructural en la tendencia de ambas series en los meses que se comenzaron a aplicar medidas sanitarias por parte de las autoridades locales para reducir la movilidad de las personas y frenar el avance la pandemia.

Las estadísticas descriptivas son:

```
#====== 1.2 Estadísticas Descriptivas Precio Industria Cervezas =========#

skimr::skim(select(datoscer,PCERVEZAS_T))

-- Variable type: numeric -------------------------------------------------------------------
# A tibble: 1 x 10
  skim_variable n_missing complete_rate  mean    sd    p0   p25   p50   p75  p100
* <chr>              <int>         <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 PCERVEZAS_T            0             1 1582.  95.4 1331. 1535. 1606. 1642.  1772

#====== 1.2 Estadísticas Descriptivas Precio Industria Gaseosas =========#

skimr::skim(select(datosgas,PGASEOSAS_T))

-- Variable type: numeric -------------------------------------------------------------------
# A tibble: 1 x 10
  skim_variable n_missing complete_rate  mean    sd    p0   p25   p50   p75  p100
* <chr>              <int>         <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 PGASEOSAS_T            0             1  742.  37.2  656.  714.  749.  771.  820.
```

El precio promedio por litro de cervezas redondeado a su valor entero es de $1582, el precio más bajo alcanzado es de $1331, el precio máximo es de $1772 y la desviación estándar es de $95.4. Para la categoría de gaseosas, se observa que el precio promedio por litro es de $742, el precio menor es de $656, el precio máximo es de $820 y la desviación estándar es de $37.2 que es mucho menor a la obtenida para la categoría de cervezas.

Note que las dimensiones:

- Temperatura Máxima
- Tasa de Desempleo
- Índice de Movilidad
- Retiros AFP

Se utilizan en ambos modelos econométricos, por lo tanto se decide entregar un reporte descriptivo de estas variables considerando la información disponible en los datos de desarrollo de la industría cerveza (tiene mayor profundidad histórica).

La Figura 5 contiene los gráficos de tendencias de cada unas las cuatro dimensiones mencionadas:

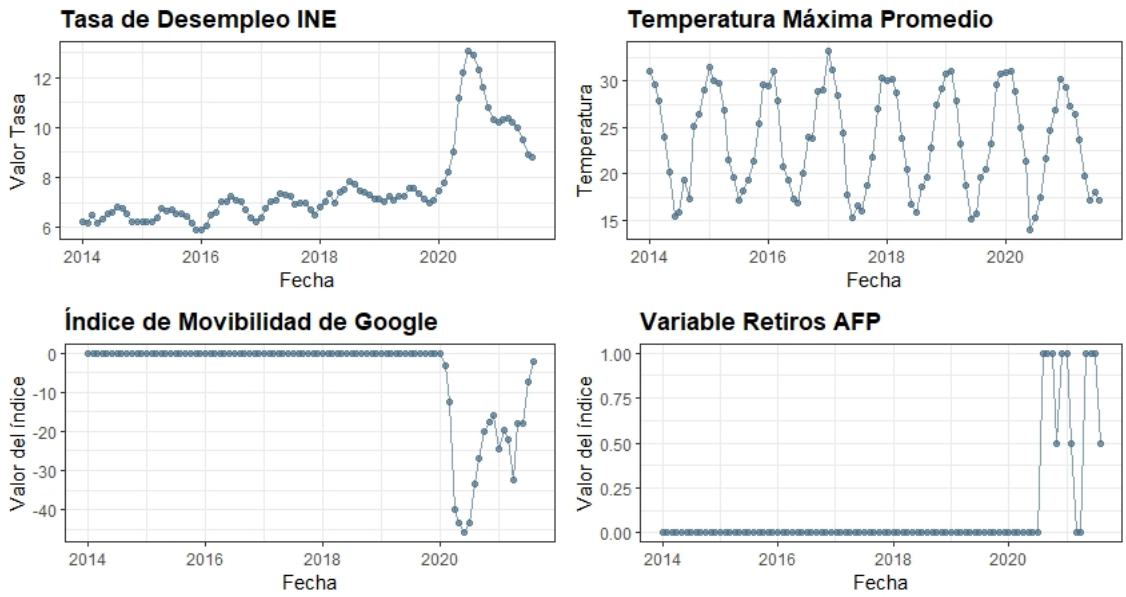

Figura 5: Gráficos de tendencia mensual de las dimensiones consideradas en los modelos

A partir de los gráficos de tendencia se pueden observar algunas características:

- La tasa de desempleo presenta una tendencia positiva con una fuerte alza en el año 2020, seguramente en parte explicado por pandemia en Chile.
- La variable temperatura máxima promedio tiene un comportamiento estacional anual (propio de las temperaturas), similar al presentado en las ventas volumen por industria y por las características naturales de esta variable, debe ser uno de los ejes más relevantes para proyectar la venta volumen por industria.
- El índice de movilidad calcula cómo cambia la cantidad de visitantes en los lugares categorizados (retail and recreation, grocery and pharmacy, parks, transit stations,workplaces, residential) en comparación con los días de referencia. El día de referencia es la mediana del período de 5 semanas comprendido entre el 3 de enero y el 6 de febrero del 2020. Se dispone del índice a partir de Febrero 2020. Además, se debe tener en cuenta que en el cálculo de los valores de referencia, no se ha tenido en cuenta la estacionalidad. Cabe mencionar

que se amputó con valor cero los registros donde no se disponía información sobre el índice de movilidad, lo cual en estricto rigor debiera ser un valor "NA".

- Los modelos consideraron una variable de Retiros, que fue creada por el equipo de CCU. Esta variable intenta explicar el impacto de los retiros de AFP. Al analizar los valores que toma la variable, se observan tres niveles: 1, 0.5 y 0. El equipo CCU nos menciona que la variable tomará valor 1 el mes que se anuncia el retiro y los meses siguientes rezagados tomará valor 0.5. En los otros meses la variable tomará el valor cero.

La Figura 6 contiene los gráficos de distribuciones de las cuatro dimensiones:

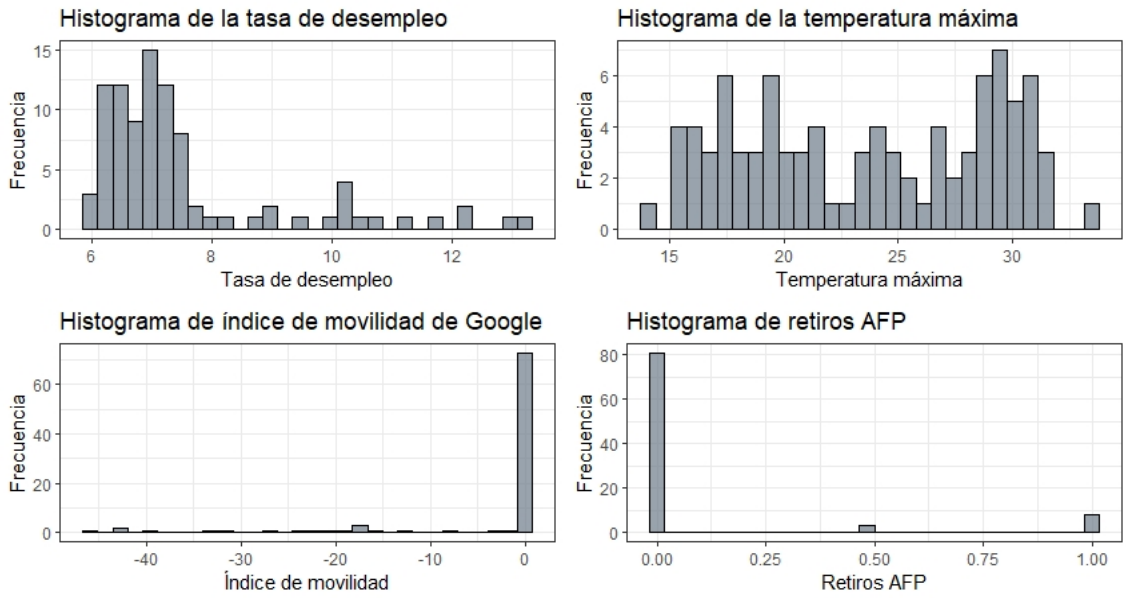

Figura 6: Gráficos de distribuciones de las dimensiones consideradas en los modelos

La siguiente tabla contiene las estadísticas descriptivas para las dimensiones cuantitativas:

```
#====== 1.2 Estadísticas Descriptivas   =========#

skimr::skim(select(datoscer,TEMP_MAX))
skimr::skim(select(datoscer,DESEMPLEO))
skimr::skim(select(datoscer,MOVILIDAD))


-- Variable type: numeric -------------------------------------------------------------------
# A tibble: 1 x 10
  skim_variable n_missing complete_rate   mean     sd     p0    p25    p50    p75   p100
* <chr>             <int>         <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
1 TEMP_MAX              0             1   23.7   5.47   13.9   18.8   23.9   29.1   33.3
2 DESEMPLEO             0             1    7.57  1.66    5.87   6.55   7.04   7.52   13.1
3 MOVILIDAD             0             1   -4.85  11.2   -45.9    0      0      0      0
```

Al analizar las medidas de resumen, se observa que la temperatura máxima promedio están dentro de un rango

estándar. El promedio es 23.7 ºC está dentro de un rango aceptable, las desviaciones no exceden los 5.4ºC grados, la temperatura máxima promedio más baja es de 13.9 ºC y la temperatura máxima promedio más alta es de 33.3 ºC. En cuanto a la tasa de desempleo, se observa que la tasa promedio es de 7.5 %, la desviación estándar es de 1.66, la tasa mínima de desempleo es de 5.8 % y la tasa máxima de desempleo es de 13.1 %. El índice de movilidad, presenta una media negativa de -4.85, una alta desviación estándar de 11.2, su valor mínimo es de -45.9 y su valor máximo es de 0. Estos valores no son representativos debido a la imputación de ceros a los valores faltantes. Cabe mencionar que la variable retiros es una variable categórica, por lo tanto no tiene sentido calcular medidas descriptivas.

Observaciones Importantes:

- La variable "Movilidad" debe ser corregida, los registros sin información deben tratarse como valores "NA".
- La variable "Retiros" tiene una definición subjetiva. Las fechas de los 3 retiros anunciados por el gobierno, son 30 Julio 2020 para el primer retiro, 10 Diciembre 2020 para el segundo y 28 Abril 2021 para el tercero. Para la fecha del primer retiro, consideran que la variable "Retiros" tomará el valor uno por tres meses consecutivos (Agosto, Septiembre y Octubre) luego decae a 0.5 en el mes de Noviembre. Sin embargo no consideran que el día 30 y 31 de Julio ya se podían realizar retiros y considera ese mes como valor nulo. Por otro lado, para el segundo retiro, consideran que la variable toma el valor 1 por dos meses consecutivos, desde el mes de anuncio (Diciembre). Luego en Febrero 2021 cae al valor 0.5 y en Marzo y Abril 2021 vuelve a caer al valor 0. Por último para el tercer retiro la variable toma el valor 1 por tres meses consecutivos: Mayo, Junio y Julio para caer a 0.5 en Agosto. Sin embargo, no considera que desde el 28 Abril ya se podia realizar el tercer retiro. Nos parece un poco subjetivo el período que permanece el valor uno la variable. En el primer retiro y tercero permanece por 3 meses y en el segundo retiro por 2 meses.

  Por otro lado, la variable "Retiros"tiene un tratamiento no adecuado en la regresión, esta información cualitativa no debe ser considerada cuantitativa en los ajustes.
- En los dos modelos econométricos utilizan la variable promedio de temperatura máxima mensual para predecir las ventas volumen por industria. Hay que tener en cuenta que esto generará un conflicto al momento de realizar proyecciones para el mes siguiente, dado que no se conoce el valor de las variable de temperatura. Por su parte, CCU informa que ellos cuentan con la proyección de esta variable y que los valores proyectados de estas temperaturas difieren muy poco de los valores observados.
- Ambos modelos utilizan la variable índice de Movilidad de google, y al momento de hacer predicciones asumen que este índice toma el valor cero, lo cual es un supuesto erróneo.
- Ambos modelos utilizan la variables creada de Retiros, y al momento de hacer predicciones asumen que esta variable toma el valor cero, lo cual podría no ser cierto si se aprueban más retiros.
- Ambos modelos utilizan el precio ponderado y al hacer predicciones el equipo de CCU informa que lo estiman según inflación o según la data disponible de las unidades de Marketing.

2. Análisis Bivariado

El análisis bivariado consiste en estudiar el comportamiento que tiene la variable objetivo de las industrias de cervezas y gaseosas con las dimensiones utilizadas para la proyección (comportamiento de a par de variables). Se realizan análisis gráficos de dispersión, estadísticos de correlación y sentido lógico (signo) de los coeficientes estimados en los modelos.

A través de la información entregada por CCU, se logró replicar los dos modelos econométricos (uno por industria), donde no se observaron diferencias entre lo reportado por CCU y los ajustados en la auditoría. Esto es de gran utilidad para comparar los resultados del análisis bivariado con los modelos ajustados y detectar incoherencias. Notar que los modelos econométricos ajustados por CCU, no dividen la muestra en datos de entrenamiento y validación, por lo cual en está sección se trabaja con toda la información reportada.

Las Figuras 7 y 8 grafican la dispersión presentada por las variables independientes y la venta volumen de la industria de cervezas y gaseosas, respectivamente. Para todos los gráficos se considera lo siguiente: En el eje de ordenadas (eje Y) se encuentra la venta volumen de la industria y en el eje de las abscisa (eje X) las variables independientes utilizadas en el modelo, con rojo se gráfica la recta estimada que pasa por los puntos.

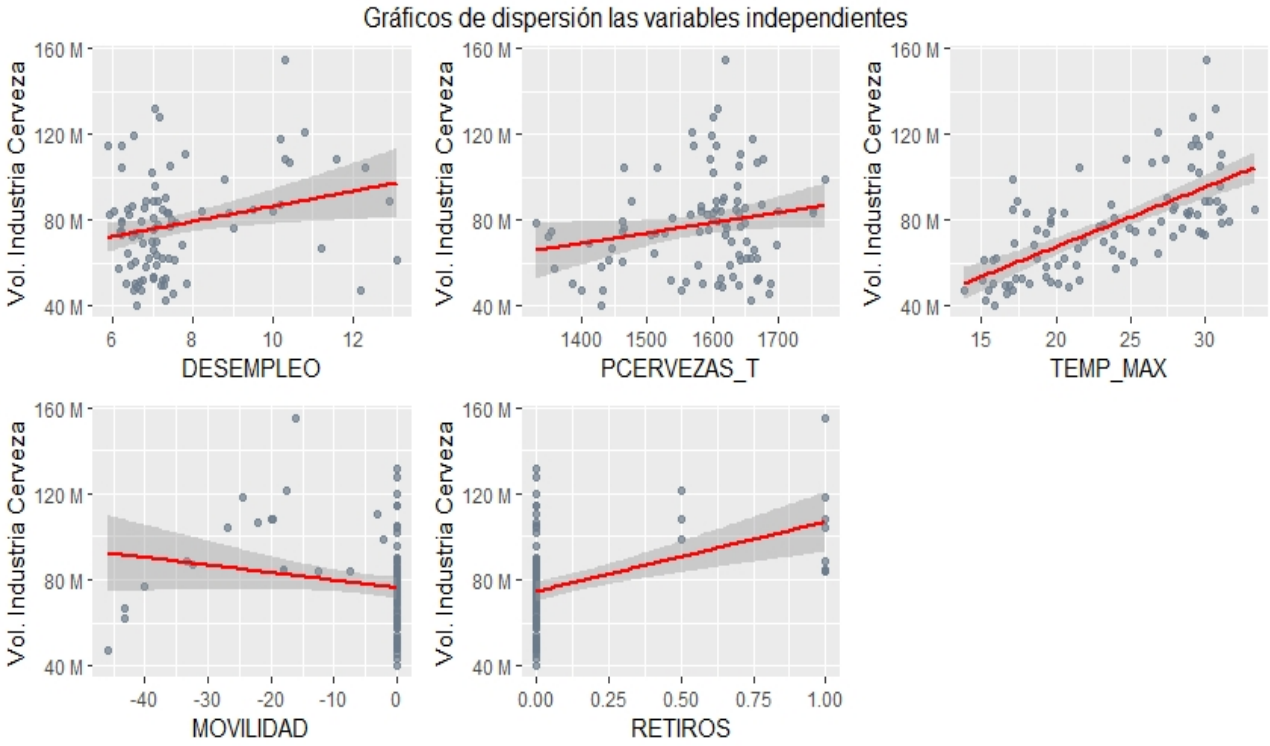

Figura 7: Gráficos de dispersión de las variables independientes con el Volumen de la industria Cerveza

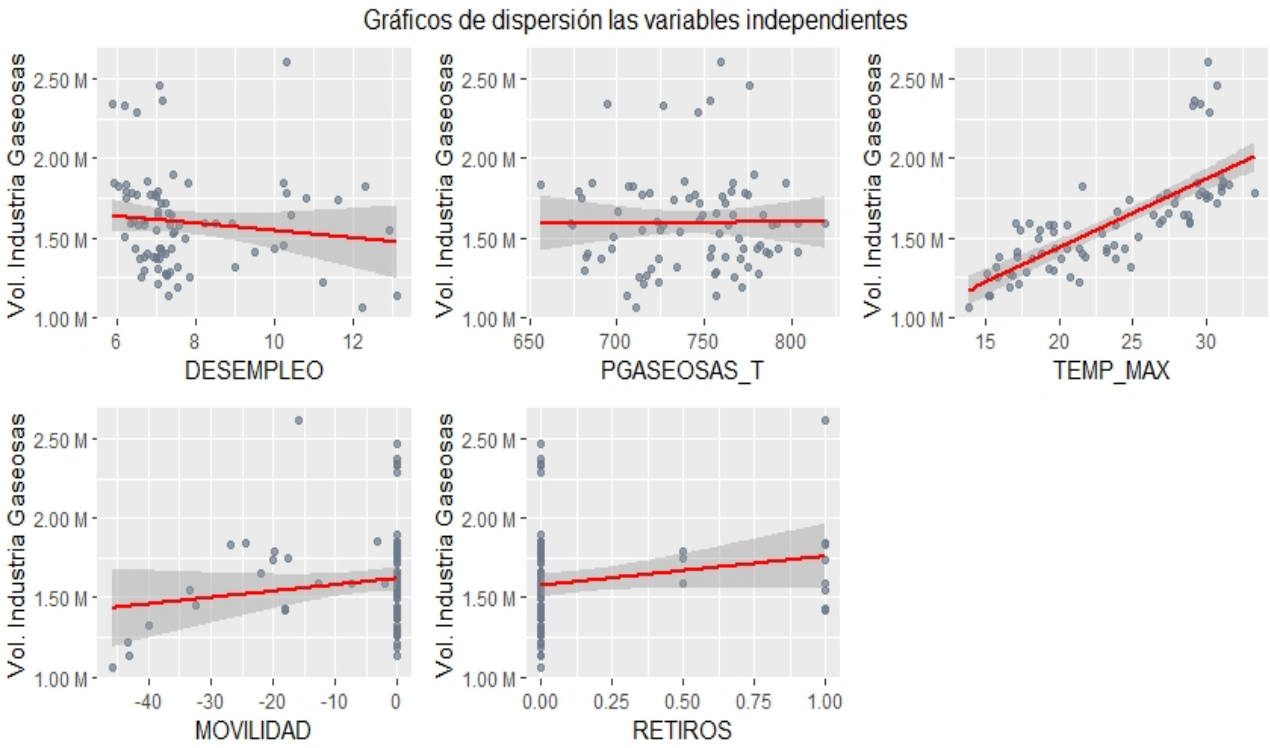

Figura 8: Gráficos de dispersión de las variables independientes con el Volumen de la industria Gaseosa

- Industria Cerveza: Se observa que, salvo con la variable movilidad, todas las relaciones tienen un sentido positivo, es decir, a medida que aumentan las variables explicativas se observa que aumenta la venta volumen de cervezas. Por lo tanto, se espera que los coeficientes β que acompañan a estas variables en los modelos econométricos sean positivos. Las rectas graficadas (estimadas) sobre los pares de puntos tienen mayor pendientes en la variable temperatura máxima promedio. Al analizar el gráfico de dispersión de la movilidad con la venta volumen de cervezas, se aprecia leve sentido negativo, esto se debe a los ceros imputados en la data no disponible. Si se corrigieran por "NA", se apreciaría asociación positiva con la venta volumen de cervezas. Por otro lado, se observa que el precio por categoría es un fenómeno inelástico en el modelo, una alternativa será analizar esta dimesión en valor UF y observar su comportamiento.
- Industria Gaseosa: A partir de los graficos note que la variable temperatura máxima promedio, presenta una relación positiva con la venta volumen de gaseosas, es decir, en la medida que aumenta la temperatura máxima promedio, aumenta también la venta volumen de gaseosas. De esta manera, se espera que el coeficiente β que acompañan a la variable temperatura máxima en el modelo econométrico sea positivo. No se observa relación entre la variable precio gaseosa y la variable venta volumen. Además las variables movilidad, tasa desempleo y retiros presentan muy baja relación con la variable objetivo.

A continuación se presentan todas las correlaciones de Person de cada una de las variables independientes con las variables objetivos (venta volumen cervezas y gaseosas) ordenado por magnitud de la correlación (por valor absoluto):

```
#========== 2.1 Correlación variables con venta volumen cervezas ==========#
# A tibble: 5 x 2
  Variable    Correlacion
  <chr>             <dbl>
1 TEMP_MAX          0.666
2 RETIROS           0.416
3 DESEMPLEO         0.258
4 PCERVEZAS_T       0.201
5 MOVILIDAD        -0.175

#========== 2.1 Correlación variables  venta volumen gaseosas ==========#
# A tibble: 5 x 2
  Variable    Correlacion
  <chr>             <dbl>
1 TEMP_MAX         0.761
2 RETIROS          0.185
3 MOVILIDAD        0.153
4 DESEMPLEO       -0.125
5 PGASEOSAS_T      0.003
```

La variable que presenta mayor magnitud de correlación es la temperatura máxima promedio en ambos modelos. La variable que presentan menor grado de correlación con la venta volumen de cervezas es el índice de movilidad. Se aprecia que las dimensiones retiros, movilidad, tasa de desempleo presentan muy baja correlación con la variable venta volumen de gasesosas, y el precio de gaseosas tiene casi nula correlación con el volumen de vente de la industria gaseosa.

Se replicaron los modelos ajustados por CCU, el cual corresponde a un modelo SARIMA$(3,0,0)(0,1,1)[12]$ para la venta volumen de cervezas y un modelo SARIMA$(0,0,0)(1,1,0)[12]$ para la venta volumen de gaseosas. Note que la función `auto.arima()` arroja en ambos modelos un coeficiente de `drift` lo que quiere decir que agrega una tendencia determinística al modelo, de la forma bt donde t es el tiempo (en meses) y la pendiente b es el `drift`.

A partir de los modelos ajustados por CCU se identifican algunas inconsistencias a lo obtenido en el análisis bivariado, dado que algunos de los coeficientes estimados no presentan el signo esperado. En las tablas siguientes, bajo el nombre de cada variable se encuentran los parámetros estimados para cada uno de los modelos, y destacados en color gris aquellos coeficientes que no tienen sentido lógico según el análisis bivariado. Estos coeficientes que se encuentran relacionados al precio, y movilidad para el modelo de cervezas y tasa de desempleo y precio para el modelo de gaseosas, tienen una incorrecta interpretación y se deben investigar la causa (usualmente se debe a problemas de colinealidad).

Por otro lado, al analizar la significancia de los coeficientes del modelo de cervezas, se observa que a un nivel de significancia de 5 %, el coeficiente ar1 del modelo SARIMA no es significativo, es decir, estadísticamente ese coeficiente es cero. En cuanto a la significancia de los coeficientes de las variables predictoras, se observa que la variable tasa de desempleo y precio de cervezas no son significativos.
En el modelo de gaseosas, se observa que todos los parámetros del modelo son significativos a un nivel de 5 %, a excepción de la variable tasa de desempleo.

```
#========== Modelo Ajustado para industria Cervezas ==========#

Series: Cervezas
Regression with ARIMA(3,0,0)(0,1,1)[12] errors

Coefficients:
         ar1     ar2     ar3    sma1     drift   TEMP_MAX  DESEMPLEO
      0.0931  0.3481  0.4317  -0.661  530818.5  1709976.9    2038296
s.e.  0.1021  0.1004  0.1081   0.149  170210.5   361085.9    1045297

      PCERVEZAS_T   MOVILIDAD   RETIROS
       -24201.25    588753.2   5918466
s.e.     21529.65   154884.9    2335186

sigma^2 estimated as 2.593e+13:  log likelihood=-1347.42
AIC=2716.84   AICc=2720.73   BIC=2743.05
```

```
#========== Modelo Ajustado para industria Gaseosas ==========#

Series: Gaseosas
Regression with ARIMA(0,0,0)(1,1,0)[12] errors

Coefficients:
         sar1     drift    DESEMPLEO  PGASEOSAS_T   TEMP_MAX  MOVILIDAD
      -0.5437  4745.195     2784.153   -2019.8108  28623.463   7993.726
s.e.   0.1089  1330.043    15172.209     583.0434   4535.574   1925.482

        RETIROS
      179702.85
s.e.   28439.48

sigma^2 estimated as 4.23e+09:  log likelihood=-848.52
AIC=1713.04   AICc=1715.48   BIC=1730.79
```

3. Análisis Multivariado

El método de mínimos cuadrados para estimar los paramétros desconocidos β de una regresión lineal supone que las dimensiones consideradas en el modelo no presentan un alto nivel de correlación. En esta sección se analiza la correlación de todas las variables utilizadas en los modelos.

Por medio de gráficos de heatmaps, las Figuras 9 y 10 presentan las correlaciones de las dimensiones consideradas en cada uno de los econométricos (cervezas y gaseosas respectivamente). En ambos modelos las variables tasa de desempleo e índice de movilidad presentan altas correlaciones (en torno a -0.9), y además las variables desempleo y retiros también presentan alta correlación (en torno a 0.65) .

Estas altas correlaciones (en magnitud) pueden ser la fuente de poblemas de signo en las estimaciones de los parámetros de los modelos, esto será revisado en la siguiente sección por medio los estadísticos de inflación de la varianza (VIF).

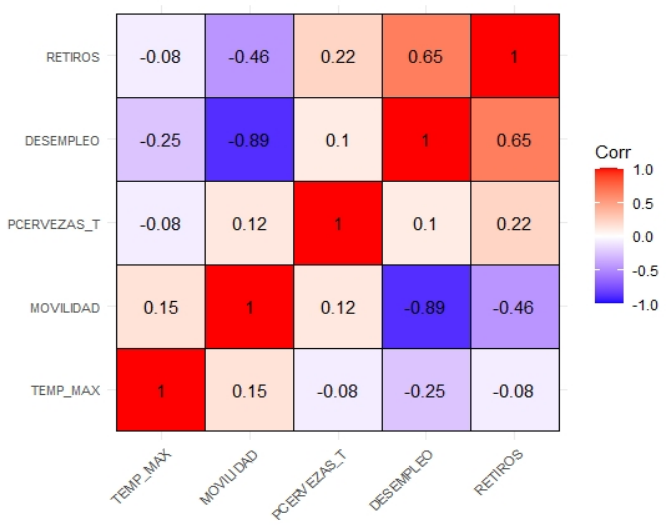

Figura 9: Heatmap con la matriz de correlación de las variables independientes del modelo de proyección de venta volumen de cervezas

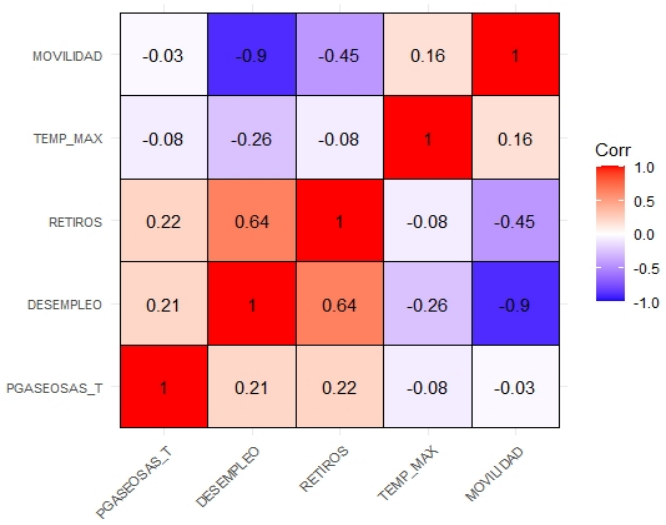

Figura 10: Heatmap con la matriz de correlación de la variables independientes del modelo de proyección de venta volumen de gaseosas

4. Verificación de Supuestos Econométricos

Los modelos de proyección de venta volumen desarrollados por CCU deben cumplir con supuestos econométricos para su correcto uso. En esta sección se revisan los supuestos:

- Normalidad: Los errores u_t del modelo tiene una distribución normal.
- Independencia: Los errores u_t no presentan correlación serial.
- No Colinealidad: Las variables independientes no presentan alta correlación.
- Homocedasticidad: Los errores u_t tienen varianza constante.

El supuesto de linealidad de las variables independientes con la venta volumen de la industria ya fue analizado en la sección 2 y para el revisar el supuesto de Independencia lineal, se utilizará de apoyo lo revisado en la sección 3.

Normalidad

Dado que se desconocen los errores teóricos del modelos, estos se estiman por medio de los residuos que se definen de la siguiente manera;

$$\widehat{u_t} = y_t - \widehat{y_t} - \widehat{\epsilon_t},$$

donde y_t corresponde a la venta volumen mensual de cervezas (o gaseosas), $\widehat{y_t}$ corresponde venta volumen mensual de cervezas (o gaseosas) proyectado por medio de la regresión lineal y $\widehat{\epsilon_t}$ corresponde al ajuste del modelo SARIMA a los errores del modelo lineal. Para cada industria se determinan los residuos y, por medio de análisis gráficos y test de hipótesis estadísticos se evalúa si existe evidencia empírica para descartar la normalidad.

Primero se realizan los gráficos QQ-plot, donde se comparan los cuantiles empíricos de los residuos con los cuantiles teóricos de la distribución normal. Junto con el QQ-plot se realiza un gráfico de la distribución de los residuos con una curva normal sobrepuesta. Las Figuras 11 y 12 contienen los qqplot y sus histogramas para la industria de cervezas y gaseosas, respectivamente.

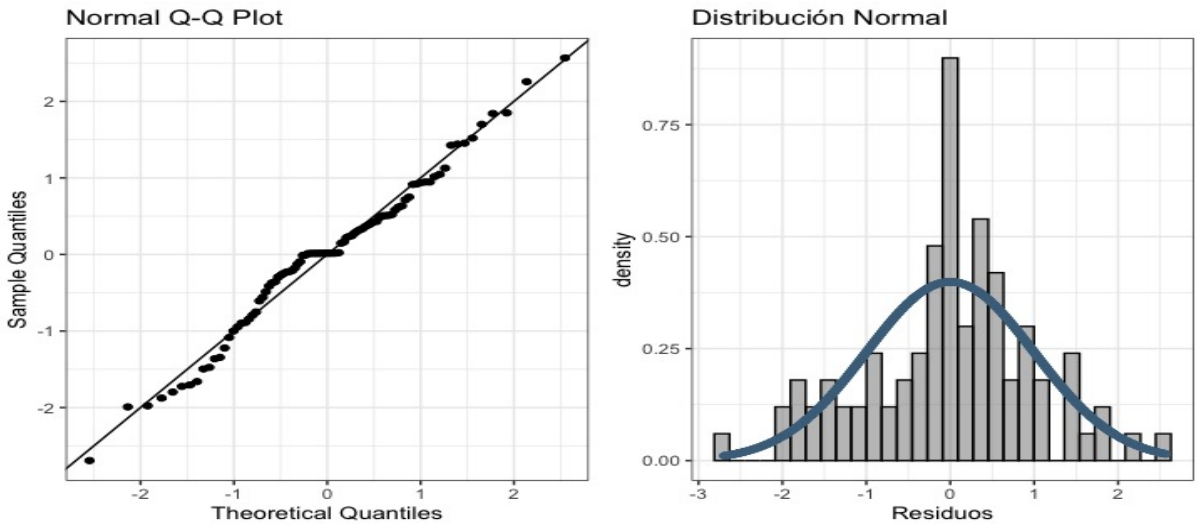

Figura 11: QQ-plot y el histograma de los residuos del modelo de proyección para la industria de cervezas

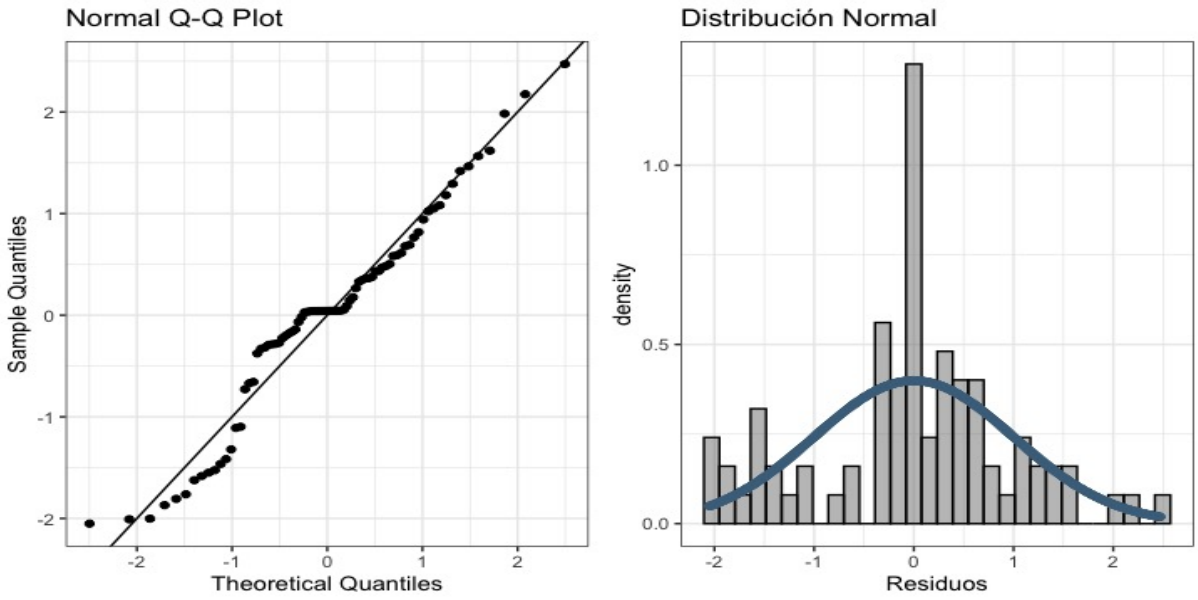

Figura 12: QQ-plot y el histograma de los residuos del modelo de proyección para la industria de gaseosas

Para los dos modelos, observamos que los cuantiles empíricos de los residuos estandarizados se asemejan a los cuantiles teóricos de la distribución normal estándar. Se aprecia que en el centro de la distribución la distribución empírica es más densa densidad que la teórica. Además, los histogramas de los residuos presentan un comportamiento similar a la distribución normal. Para evaluar de manera estadística esta hipótesis, se realiza el test de bonda de ajuste de Kolmogorov-Smirnov (bajo normalidad). Para ver detalles del test consultar en [3] página 428.

```
#======== 3.1 Test de Normalidad Residuos del modelo de cervezas ========#

        One-sample Kolmogorov-Smirnov test

data:  datoscer$Residuos
D = 0.10408, p-value = 0.2537
alternative hypothesis: two-sided

#======== 3.1 Test de Normalidad Residuos del modelo de gaseosas =====#

        One-sample Kolmogorov-Smirnov test

data:  datosgas$Residuos
D = 0.13337, p-value = 0.1059
alternative hypothesis: two-sided
```

Valores pequeños de `p-value` (menores a 0.05) indican que existe fuerte evidencia empírica para rechazar la hipótesis de normalidad de los residuos. Valores grandes de `p-value` (mayores que 0.05) indican que no existe evidencia estadística en contra de la hipótesis de normalidad de los residuos. Dado que los `p-value` de los dos modelos son grandes, 0.2537 para el modelo de cervezas y 0.1059 para el de gaseosas, no existe evidencia estadística para descartar la normalidad de los residuos.

Independencia

El modelo SARIMA supone que los errores u_t son independientes y, dado que el modelo se aplica a una variable medida mensualmente, es importante verificar el cumplimiento de este supuesto. Para este análisis se debe trabajar con los residuos del modelo SARIMA $\hat{u}_t$ que son una estimación del error, y por medio de los residuos se realiza un gráfico de la función de autocorrelación (ACF) y el test de Box-Ljung.

El gráfico ACF (función de autocorrelación) muestra la correlación que presentan los residuos de un modelo dado un intervalo fijo de tiempo. Valores grandes en la función de autocorrelación son una alerta para indicar que no se está cumpliendo el supuesto de independencia. Complementariamente, el test de Box-Ljung evalúa simultaneamente si todas las correlaciones a través del tiempo son estadísticamente significativas. Para ver detalles del test de Box-Ljung consultar [1] página 36.

La Figura 13 contiene los gráficos de ACF de los residuos de los modelos. Se observa que ambos modelos no presentan altos valores de autocorrelación para diferentes lag (longitud de tiempo), todas estas autocorrelaciones se encuentran bajo los intervalos de confianza (línea punteada azul en el gráfico), eso garantiza que los residuos de los modelos no presentan correlación serial.

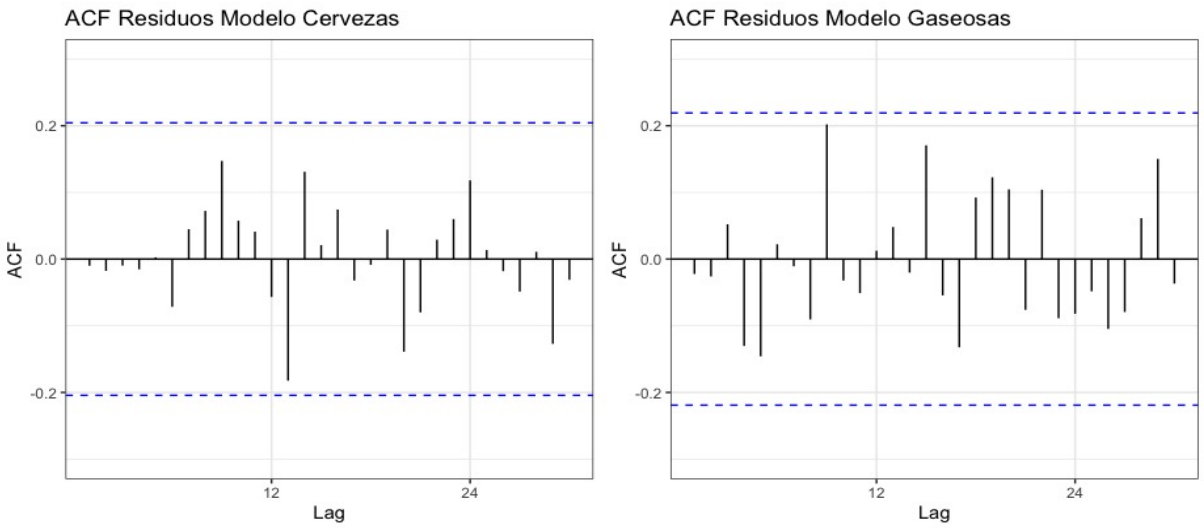

Figura 13: Gráfico de autocorrelación de los residuos de los modelos

A continuación se realiza el test de Ljung–Box para los residuos de ambos modelos considerando diferentes longitudes de dependencia.

```
#======== 4.1 Test de Box-Ljung de los Residuos del modelo cervezas ========#

# A tibble: 12 x 3
     Lag Estadistico valor_p
   <int>       <dbl>   <dbl>
 1     1     0.00951   0.922
 2     2     0.0397    0.980
 3     3     0.0490    0.997
 4     4     0.0721    0.999
 5     5     0.0728    1.00
 6     6     0.588     0.997
 7     7     0.793     0.998
 8     8     1.33      0.995
 9     9     3.59      0.936
10    10     3.94      0.950
11    11     4.12      0.966
12    12     4.47      0.973
```

```
#======== 4.1 Test de Box-Ljung de los Residuos del modelo gaseosas ========#

# A tibble: 12 x 3
     Lag Estadistico valor_p
   <int>       <dbl>   <dbl>
 1     1      0.0414   0.839
 2     2      0.0985   0.952
 3     3      0.331    0.954
 4     4      1.80     0.773
 5     5      3.66     0.600
 6     6      3.70     0.717
 7     7      3.71     0.812
 8     8      4.46     0.814
 9     9      8.23     0.511
10    10      8.33     0.597
11    11      8.57     0.661
12    12      8.59     0.738
```

Valores pequeños del `p-value` (menos que 0.05) indican que existe evidencia estadística para rechazar la hipótesis "nula" de que existe autocorrelación entre los residuos del modelo a diferentes longitudes de tiempo (autocorrelación significativamente distinta de cero). Para cada modelo se aplicó el test de Box-Ljung hasta una longitud de 12 meses, de los cuales, en los dos modelos no se encontraron autocorrelaciones significativamente distintas de cero. La Figura 14 muestran los valores-p del estadístico de Box-Ljung para distintos valores de lag (hasta 30) para el modelo cervezas (panel izquierdo) y el modelo gaseosas (panel derecho). Se aprecia que todos los valores-p están sobre la línea roja punteada (correspondiente al valor 0.05). Por lo tanto, se concluye que los residuos de ambos modelos no presentan correlación serial, por lo cual se cumpliría el supuesto de independencia.

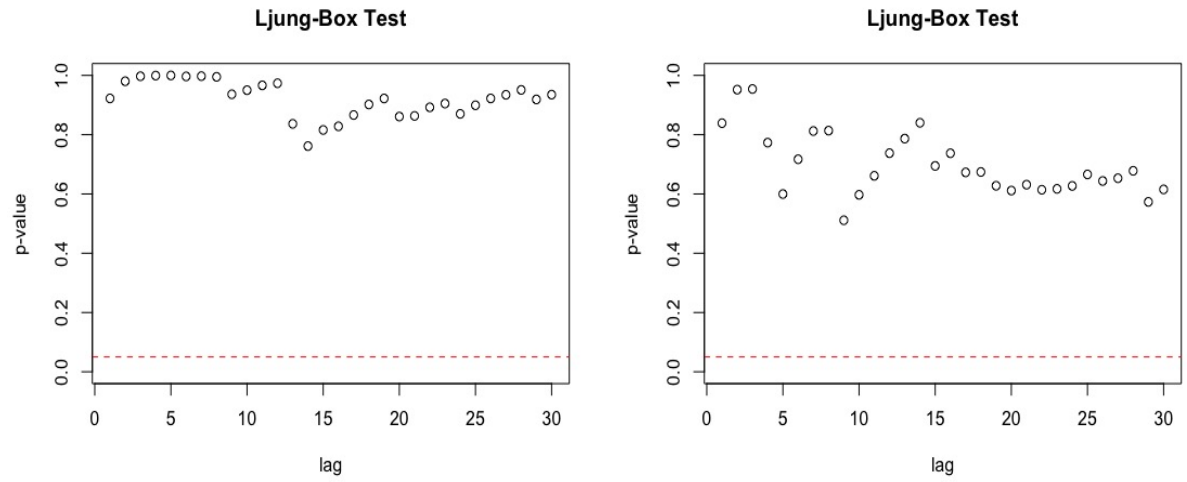

Figura 14: Gráfico de valores-p del test Box-Ljung. Panel izquierdo modelo cervezas, panel derecho modelo gaseosas

4.1. No Colinealidad

Cuando se estiman los parámetros de un modelo de regresión lineal múltiple con el método de mínimos cuadrados, se necesita un supuesto de no colinealidad entre las dimensiones consideradas. El término de colinealidad hace referencia cuando dos variables o más son proporcionales entre sí. La colinealidad genera problemas de estimación (sesgo) en los parámetros del modelos y aumenta la varianzas de los estimadores. En la sección 3 ya se analizó la colinealidad de las variables independientes mediante las correlaciones (gráficos de la matriz de correlación con heatmaps) y se identificaron problemas en los dos modelos. En esta sección se apoya ese análisis con el cálculo de los índices de inflación de la varianza (VIF), este estadígrafo identifica que parámetros (y variables) presentan problemas de estimación debido a su fuerte colinealidad. Si quiere ver detalles puede consultar en [4] página 254.

```
#======== 5.1 VIF parámetros estimados modelo cervezas ========#

  Variable       VIF
  <chr>        <dbl>
1 DESEMPLEO     193.789
2 PCERVEZAS_T   189.882
3 TEMP_MAX       18.857
4 MOVILIDAD       8.578
5 RETIROS         1.893

#======== 5.1 VIF parámetros estimados modelo gaseosas ========#

  Variable       VIF
  <chr>        <dbl>
1 DESEMPLEO     209.937
2 PGASEOSAS_T   210.772
3 TEMP_MAX       20.191
4 MOVILIDAD       8.554
5 RETIROS         2.109
```

Todos los VIF señalados con gris indican alta colinealidad (mayores al punto de corte 5), por lo tanto, la varianza de esos estimadores está inflada por la correlación y las estimaciones de los parámetros tienen sesgo. Estos resultados concuerdan con los análisis presentados en la sección 2 y 3, donde se identifico problemas en los signos de los parámetros estimados y alta correlación entre las variables utilizadas para proyectar las ventas volumen de cervezas y gaseosas.

Homocedasticidad

La forma más directa de evaluar el supuesto de homocedasticidad es realizar un gráfico de los valores proyectados con los residuos, cualquier patrón que genere un aumento o disminución en la variabilidad de los datos será considerado como presencia de heterocedasticidad. La Figura 15 contiene los gráficos de los residuos de cada modelo con las ventas volumen proyectadas. En ambos modelos se aprecia que no hay mayor dispersión de los residuos para todos los valores proyectados de la venta volumen de las industrias. Por lo tanto, se concluye que no existe algún patrón que genere un aumento en la variabilidad de los residuos.

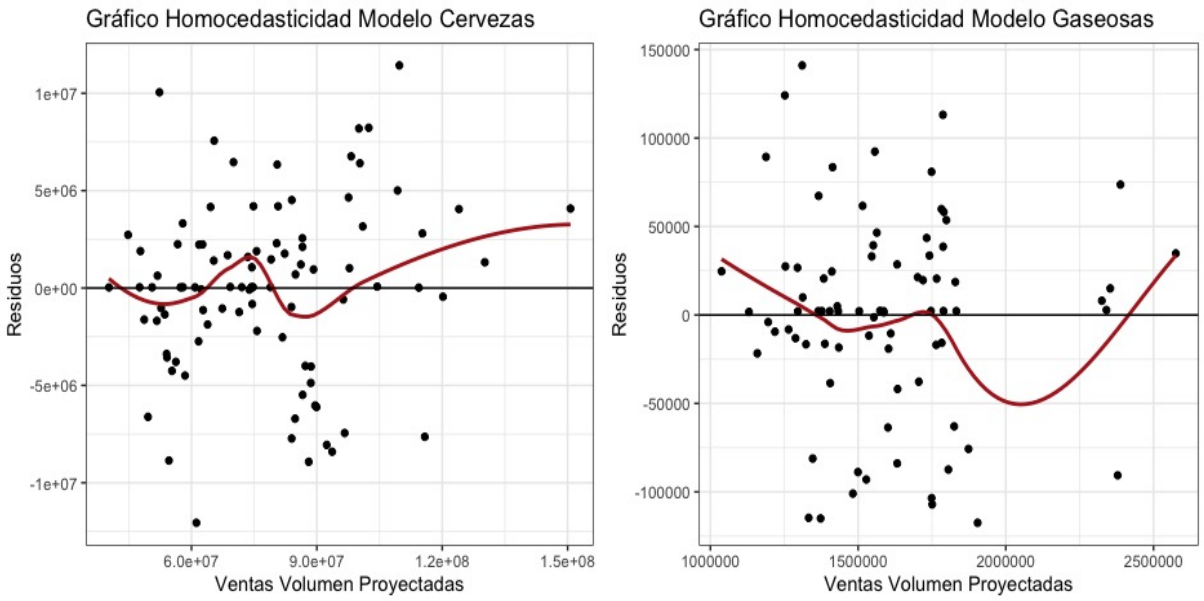

Figura 15: Gráfico de dispersión entre la venta volumen de la industria proyectadas y los residuos del modelo. Panel izquierdo modelo de cervezas, panel derecho modelo de gaseosas

5. Conclusiones

A continuación un detalle con los resultados de la auditoría presentados en todas las secciones del documento:

- Gerencia Inteligencia de Mercado de CCU puso a disposición para una auditoría analítica dos modelos econométricos para proyectar la venta volumen de la industria (cerveza y gaseosa). Junto con los modelos, se entregó toda la información para poder replicar los modelos y realizar los análisis pertinentes de la auditoría. Todos los modelos fueron replicados a la perfección en la auditoría y toda la información se encontraba completa, lo que permitió realizar un diagnóstico completo.

- A partir del primer análisis descriptivo de las variables se identificaron tres problemas.
    - La variable `Retiros AFP`, no es adecuada la forma de construir esa variable, también hay un error en la utilización de esta para la estimación de los parámetros en los modelos econométricos. La variable `Retiros AFP` se debe generar por medio de variables dummy's y los ponderadores de las categorías deben ser estimados en el modelo estadístico, no fijados subjetivamente. Por otro lado, para hacer predicciones asumen que la variable tomará el valor cero en los próximos meses, lo cual podría no ser real si se aprueban mas retiros.
    - La variable `índice de movilidad`, tiene una imputación de todos los valores previos al año 2020. No es correcto asumir que los datos faltantes son cero. Además para hacer predicciones asumen que este índice será cero, lo cual es erróneo.
    - La variable `Temperatura Máxima` se utiliza de manera simultánea en los modelos, por lo tanto, se debe contar con la predicción a futuro de la misma. Para el correcto diagnóstico predictivo se debe utilizar la predicción de la temperatura máxima mensual y no el valor observado.
- Problemas con los parámetros estimados: Ambos modelos econométricos ajustados tienen problema de signo con los parámetros estimados. En la sección 2 se realiza un análisis bivariado de todas las dimensiones consideradas en cada uno de los modelos con la variable objetivo, y las relaciones observadas y el signo de las pendientes no siempre son consistentes con los valores estimados por mínimos cuadrado para cada regresión lineal.
- Se observa que ambos modelos ajustados incorporan un efecto determinístico lineal en la tendencia (parámetro `drift`). Se necesita aclarar si el fenómeno tiene tendencia determinística,y además algunos coeficientes no son significativos a un nivel de 5 %.
- Las variables independientes `desempleo`, `movilidad` y `retiros` presentan altos niveles de correlación. Esto se identificó en la sección 3 y en la sección 4 fue confirmado mediante los estadístico VIF. Esto tiene como consecuencia un alto sesgo en la estimación de los parámetros de los modelos econométricos.
- Para un correcto análisis y uso de estos modelos, siempre se debe separar la información disponible en muestra de entrenamiento y muestra de validación. La muestra de entrenamiento (por lo general un 80 % de los datos) será utilizada para estimar el modelo. Luego, con la muestra de validación (el 20 % restante) se utiliza para evaluar el nivel predictivo del modelo, utilizando indicadores como por ejemplo el MAE, MAPE y RMSE.

6. Propuestas de Mejora

A continuación se presentan las alternativas metodológicas usuales de la literatura para solucionar los problemas mencionados en la sección 5 de conclusiones. Además se mencionan otras posibles mejoras considerando la estructura de la información:

- La variable `Retiros AFP` se debe construir mediante variables dummy's. También se puede evaluar considerar otra dimensión macro que esté relacionada con el aumento de dinero circulante en tiempos de pandemia.
- Se debe evaluar el uso de las proyecciones de la temperaturas para realizar los ajustes de los modelos econométricos.
- Evaluar corregir el precio de la categoría de cervezas y gaseosas por deflacción del IPC o UF.
- Si se desea estimar los parámetros de los modelos con mínimos cuadrados, se debe utilizar un proceso de selección de variables que considere aquellas dimensiones que tengan mayor poder predictivo con la variable objetivo y baja correlación entre ellas.

- Construir una variable cualitativa para los meses del año y hacer con tratamiento de dummy con ella. Evaluar si este efecto es significativo en el modelo.
- Si se desea utilizar todas las variables independientes a pesar de la alta correlación que presentan, se necesita cambiar el método de estimación, dado que mínimos cuadrados produce sesgo con inflación de las varianzas de los estimadores. Nuestra recomendación sería evaluar un método de regularización, Lasso es una buena opción.
- Separar la información disponible en muestra de entrenamiento y muestra de validación. Luego calcular los índices de desempeño MAPE y MAE.
- Evaluar otras dimensiones relacionadas a la venta volumen:
  - Índice mensual de confianza empresarial (IMCE)
  - Índice de precios de alimentos
  - Índice de percepción de la economía
  - Índice Mensual de Actividad Económica (IMACEC)

Referencias

[1] P.J. Brockwell and R.A. Davis. *Introduction to Time Series and Forecasting*. Springer Texts in Statistics. Springer New York, 2006.

[2] R Core Team. R: A language and environment for statistical computing. 2021. https://cran.r-project.org/web/packages/forecast/forecast.pdf.

[3] William Jay Conover. *Practical nonparametric statistics*. Wiley series in probability and statistics. Wiley, New York, NY [u.a.], 3. ed edition, 1999.

[4] G.A.F. Seber and A.J. Lee. *Linear Regression Analysis*. Wiley Series in Probability and Statistics. Wiley, 2003.