Análisis de series de tiempo no estacionarias

Samuel Méndez Villegas - A01652277[1]

[1]Módulo 5: Estadística avanzada para la ciencia de datos

[2]Inteligencia artificial avanzada para la ciencia de datos II. Grupo 502

2 de diciembre del 2022

Resumen: *Las series temporales se refieren a aquellas secuencias de datos que se registran en intervalos de tiempos regulares. Al trabajar con el tiempo como variable independiente, se tienen muchas aplicaciones, como la meteorología, telecomunicaciones, economía, entre otros. En este reporte, se analiza un caso de ventas de televisores, en el que se utilizan series de tiempo no estacionarias y otras herramientas como suavizamientos y modelos de regresión lineal con el objetivo de predecir las ventas a futuro. Como resultados del análisis, se obtuvo un modelo de regresión lineal el cual cumple con los requerimientos necesarios de normalidad, significancia de parámetros, entre otros para catalogarse como un buen modelo.*

Introducción

Las series de tiempo se refieren a aquellas secuencias de datos que se recopilan, observan o registran en intervalos de tiempo regulares. Estos intervalos pueden representar días, semanas, semestres, años, etc. En otras palabras, es el conjunto de observaciones sobre los valores que toma una variable cuantitativa a través del tiempo. Visualmente, las series temporales se visualizan como un curva que evoluciona en el tiempo [1].

Las series temporales están conformadas por 4 elementos básicos de variación, los cuales son los siguientes:

- **Tendencia:** se refiere al patrón gradual y consistente de las variaciones a largo plazo.
- **Variación estacional:** se refiere a las variaciones que ocurren año tras año en los mismos meses más o menos con la misma intensidad debido a la influencia de las estaciones.
- **Variación cíclica:** se refiere a las variaciones que presentan secuencias alternas de puntos abajo y puntos arriba de la linea de tendencia que dura más de un año.
- **Variación irregular:** se refiere a la variación imprevisible y no recurrente en el tiempo. Se debe a factores a corto plazo y como se explica la variabilidad aleatoria, no se puede esperar predecir su impacto sobre la serie de tiempo.

Es importante mencionar que las series temporales se pueden clasificar en dos ramas, series estacionarias, y series no estacionarias. Las primeras se refieren a aquellas series que no presentan alguna tendencia, es decir que se tiene una media y una variabilidad constante. Por otro lado, las series no estacionarias presentan un valor medio ascendiente o descendiente a lo largo del tiempo [2].

Debido a la fácil implementación y entendimiento de las series, éstas son muy utilizadas para predecir eventos futuros tomando como base, los eventos históricos. Es por dicha razón, que en este trabajo se centrará en analizar una serie temporal con el objetivo de predecir las ventas a futuro de una compañía de televisores. De igual forma, algunas preguntas que pueden derivarse de la problemática son: ¿En qué trimestre del año se tiene un mayor número de ventas? ¿Cuál es el número de televisores que se esperan vender durante el siguiente año?

Por lo tanto, utilizando técnicas de suavizamiento, y modelos de regresión lineal en la herramienta *RStudio*, se buscará crear un modelo matemático el cual tome en cuenta la estacionalidad de las ventas para de esta manera predecir el número de televisores vendidos en los siguientes trimestres.

Análisis de los resultados

Datos a utilizar

Como primer paso para cualquier análisis estadístico, es importante indagar en la información que se tiene disponible. En este caso, se tienen datos históricos de los últimos 4 años sobre las ventas de número de televisores. Cada año está dividido en sus respectivos trimestres, teniendo así un total de 16 registros de ventas. En la figura 1 se muestra esta información.

Año	1				2				3				4			
Trimestre	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Ventas (miles)	4.8	4.1	6.0	6.5	5.8	5.2	6.8	7.4	6.0	5.6	7.5	7.8	6.3	5.9	8.0	8.4

Figura 1: Ventas de televisores a lo largo de 4 años.

Ya conociendo los datos que se tienen al alcance, ahora si se pasará a realizar el análisis correspondiente de la serie.

Análisis de tendencia y estacionalidad

Como primer paso del análisis, se realizó un diagrama de dispersión para observar la tendencia y los ciclos de los datos. En la figura 2 se muestra el gráfico correspondiente.

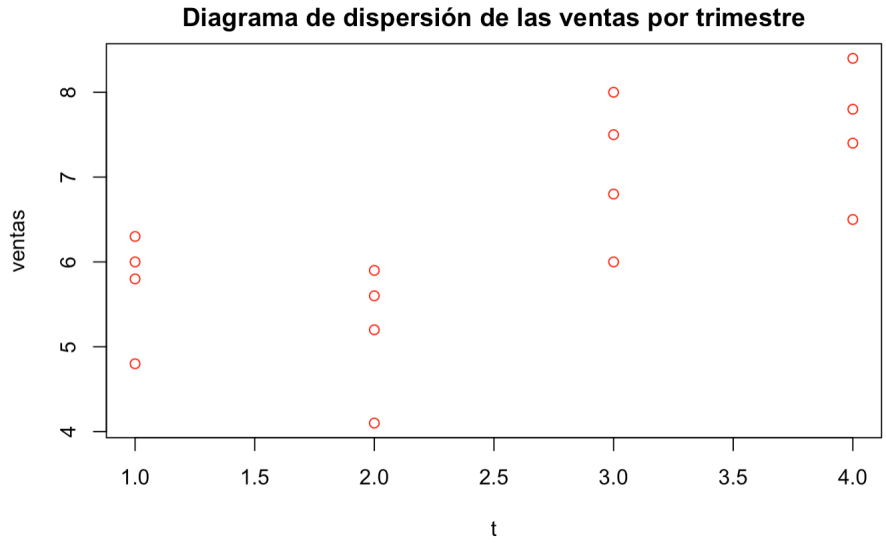

Figura 2: Diagrama de dispersión de los datos.

Se observa que dependiendo del trimestre del año, las ventas tienden a comportarse de manera distinta. En el cuarto trimestre se observa que las ventas son mayores, mientras que en el segundo trimestre, disminuyen.

Posteriormente, se graficó la serie temporal, ya tomando en cuenta los trimestres y los años. En la figura 3 se puede visualizar su comportamiento, en el cual se puede observar una tendencia creciente, y por lo tanto, puede ser un indicador de que la serie es no estacionaria.

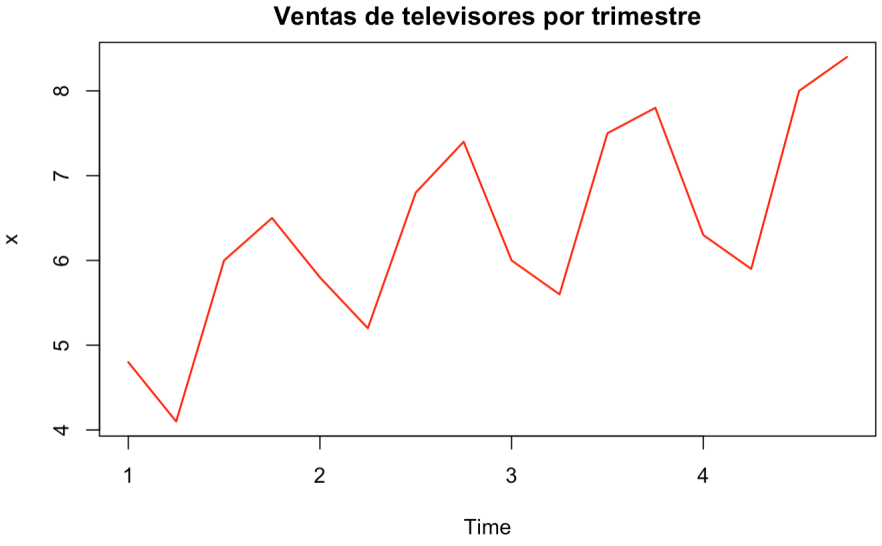

Figura 3: Serie de tiempo de la venta de televisores.

De igual forma, la serie se descompuso en 3 de sus diferentes componentes, es decir en la tendencia, en su variación estacional y en su variación irregular. Esta descomposición se observa en la figura 4.

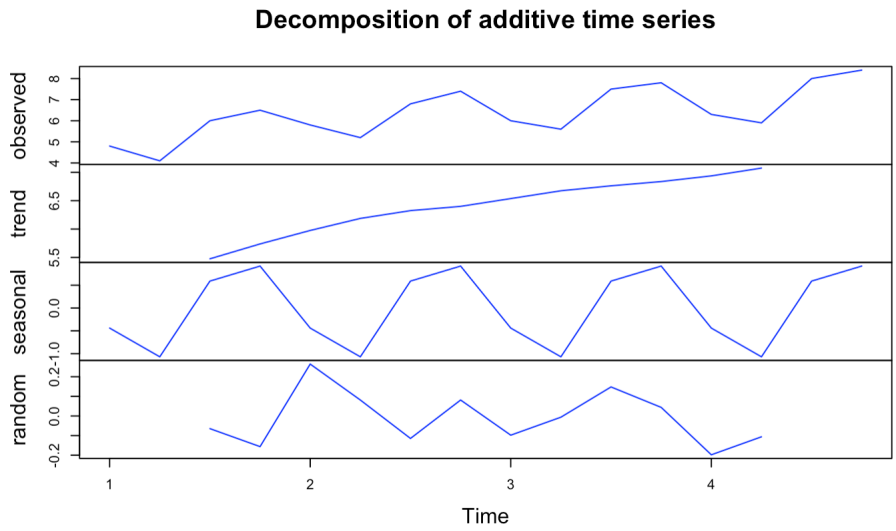

Figura 4: Descomposición de la serie en sus diferentes componentes.

En la serie de tiempo original, es decir la primer gráfica (figura 3), se observa claramente que no es estacionaria, ya que presenta una tendencia a aumentar las ventas con el paso del tiempo. Sin embargo, se puede ver que existe estacionalidad, es decir que se sigue un ciclo que se repite cada año.

Lo mencionado anteriormente se puede observar de mejor forma en la descomposición de la serie de tiempo (figura 4). La primer gráfica muestra la serie original. La segunda, gráfica muestra la tendencia, que como se mencionó, es una tendencia positiva. La tercer gráfica muestra la estacionalidad, en donde se observa que los años tienen un comportamiento bastante similar, es decir que se sigue un patrón de disminución de ventas al inicio, pero hay un aumento al final. El último componente, es el perteneciente a la variación irregular, la cual es la variación imprevisible que no se puede esperar predecir su impacto sobre la serie.

Análisis del modelo lineal de la tendencia

Ahora, se pasará a realizar un modelo de regresión lineal con el objetivo de predecir las ventas a futuro de televisores. Sin embargo, hay que tener en cuenta que si el modelo se construyera sobre los datos tal y como aparecen en la figura 3, no se obtendría un buen modelo, pues vemos que aunque la tendencia es lineal, existe mucha variación entre los trimestres, dando puerta abierta a tener error de pronóstico muy alto.

Por lo tanto, antes de aplicar el modelo, se aplicarán métodos de suvizamiento para tener un comportamiento más lineal y de igual forma calcular los índices estacionarios, los cuales permiten tomar en cuenta la estacionalidad de la serie. Para llevar a cabo dicha acción, en primer lugar se calculan los promedios móviles centrados, los cuales se obtienen a partir de los primeros promedios móviles. En este caso se estará utilizando el promedio móvil de cuatro trimestres. En la figura 5 se muestra los promedios móviles correspondientes, en los cuales los primeros 4 valores son nulos debido a que a partir de ellos se obtiene el primer promedio móvil.

```
 [1]    NA    NA    NA    NA 5.350 5.600 5.875 6.075 6.300 6.350 6.450 6.625 6.725 6.800 6.875 7.000
[17] 7.150
```

Figura 5: Promedios móviles de los datos.

Ahora, para obtener los promedios móviles centrados, se obtiene el promedio de los promedios (cada 2), quedando de la siguiente forma:

```
 [1]     NA     NA     NA     NA 5.4750 5.7375 5.9750 6.1875 6.3250 6.4000 6.5375 6.6750 6.7625 6.8375
[15] 6.9375 7.0750
```

Figura 6: Promedios móviles centrados de los datos.

Teniendo los promedios móviles centrados, se pasará a obtener los valores estacionales irregulares. Estos se calculan dividiendo el valor de las ventas entre su respectivo promedio móvil centrado.

```
Los valores estacionales irregulares son:

 1.09589 1.132898 0.9707113 0.840404 1.075099 1.15625 0.917782 0.8389513 1.109057 1.140768 0.9081081
0.8339223
```

Figura 7: Valores estacionarios irregulares.

Ya obtenidos dichos valores, ahora se agrupan por su similitud, es decir por su respectivo trimestre. Adicionalmente, se obtiene el promedio de cada uno de los grupos para calcular el índice estacional. Los índices estacionales ayudan a predecir el comportamiento de los ciclos estacionales. Esto quiere decir que se está tomando en cuenta la estacionalidad (ciclos a lo largo del año). Dichos índices se pueden observar en la figura 8.

```
Índice estacional del primer trimestre: 0.9322005
Índice estacional del segundo trimestre: 0.8377592
Índice estacional del tercer trimestre: 1.093349
Índice estacional del cuarto trimestre: 1.143305
```

Figura 8: Índices estacionales de los trimestres.

Con los índices estacionales, finalmente se puede obtener las ventas desestacionalizadas dividiendo las ventas entre el índice correspondiente a su trimestre. Estos resultados se muestran en la figura 9.

```
 [1] 5.149107 4.894008 5.487726 5.685272 6.221837 6.207034 6.219423 6.472464 6.436384 6.684498 6.859658
[12] 6.822327 6.758203 7.042596 7.316968 7.347121
```

Figura 9: Ventas desestacionalizadas.

Con estos nuevos valores, se puede graficar la serie de tiempo desestacionalizada que se ve de la siguiente forma en la figura 10:

Figura 10: Serie de tiempo con ventas desestacionalizadas.

Como se observa, es un suavizamiento diferente al que se obtendría con promedios móviles, y se tiene menos variación por decirlo de alguna forma, por lo que aplicar sobre estos datos un modelo de regresión lineal es mucho mejor que aplicarlo sobre los datos originales. Una vez hecho lo anterior, ahora si se obtendrá el modelo de regresión lineal que se obtienen a través de las ventas desestacionalizadas. El modelo obtenido se muestra en la figura 11.

```
Call:
lm(formula = y3 ~ x3)

Coefficients:
(Intercept)           x3
     5.0996       0.1471
```

Figura 11: Modelo de regresión lineal aplicado a las ventas desestacionalizadas.

Por lo tanto, el modelo quedaría de la siguiente forma:

$$y = 5.0996 + 0.1471t$$

En donde t simboliza el trimestre a calcular. Por ejemplo, si se quisiera calcular las ventas del primer trimestre del quinto año, se seguiría la siguiente fórmula:

$$y = 5.0996 + 0.1471(17) = 7.6003$$

Se pone $t = 17$ debido a que corresponde al 17^{o} trimestre que se está calculando.

Dicho valor, posteriormente se multiplicaría por el índice estacional correspondiente, es decir el del primer trimestre en este caso, obteniendo las siguientes ventas:

$$y = 7.6003(0.9322005) = 7.085003$$

Dicho resultado esta escalado, por lo que se tendría que multiplicar por 1000 para obtener las ventas reales, ya que la escala con la que se estuvo trabajando, es en miles. Por lo tanto, en el trimestre 17, se pronostica tener 7085.003 ventas de televisores.

Graficando la recta de ajuste junto con las ventas desestacionalizadas, se obtiene la figura 12.

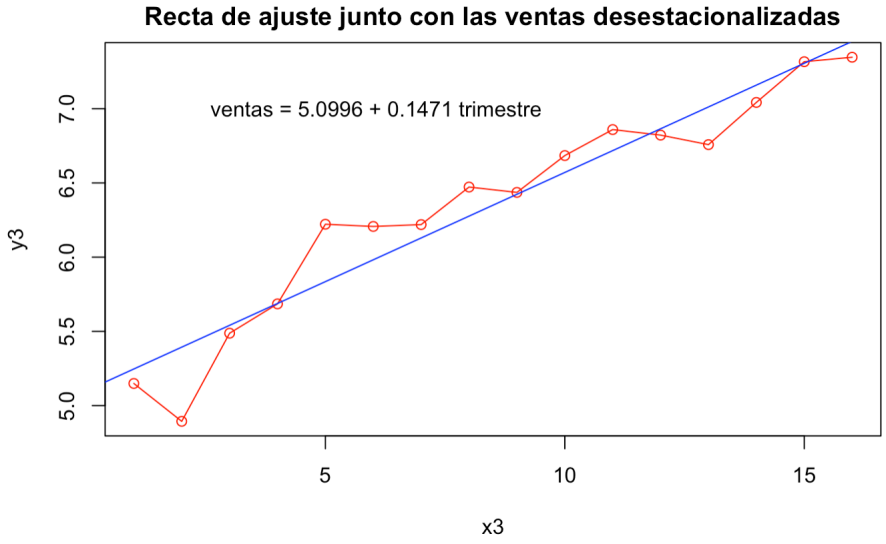

Figura 12: Modelo de regresión lineal aplicado a las ventas desestacionalizadas.

Claramente se observa que la recta de predicción no es la mejor estimación para la serie de tiempo, sin embargo se aproxima bastante a ella, por lo que los siguientes pasos serán analizar justamente la pertinencia del modelo generado.

Significancia de β_1

Una vez construido el modelo, es importante verificar la significancia de β_1 dado a que este no puede ser igual a 0, ya que como se multiplica con la variable independiente, si es 0, el producto de igual forma daría 0 indicando así independencia entre las variables. Para verificar justamente que el valor del estimador β_1 sea significativo para el modelo, realizamos una prueba de hipótesis de la siguiente forma:

- $H_0 : \beta_1 = 0$
- $H_1 : \beta_1 \neq 0$

En este caso, la regla de decisión sería que el valor del estadístico de prueba $|t*|$ sea mayor que $|t_0|$ y que el valor p sea menor que 0.05, dado a que es el nivel de significancia para la prueba. Cabe mencionar que se está utilizando el estadístico $t*$ debido a que se tienen menos de 30 observaciones y se trabajaron con 14 grados de libertad.

A realizar la prueba de hipótesis se obtuvo un valor de $t_0 = -2.144787$, por lo tanto, el valor de $|t*|$ debe de ser mayor que $|-2.144787|$. En la figura 13 se observa el resumen del modelo, en donde se puede observar el estadístico mencionado.

```
Call:
lm(formula = y3 ~ x3)

Residuals:
     Min       1Q   Median       3Q      Max
-0.49988 -0.09991  0.00369  0.12051  0.38653

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.09961    0.11153   45.73  < 2e-16 ***
x3           0.14714    0.01153   12.76 4.25e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2127 on 14 degrees of freedom
Multiple R-squared:  0.9208,    Adjusted R-squared:  0.9151
F-statistic: 162.7 on 1 and 14 DF,  p-value: 4.248e-09
```

Figura 13: Resumen del modelo de regresión lineal.

Según el resumen anterior, el valor de $|t*| = 12.76$, mientras que el $p = 4.25e-09$ dejando así claro que se rechaza la hipótesis nula.

Análisis de la pertinencia del modelo lineal

De igual forma, del resumen anterior (figura 13) se puede visualizar que el modelo tiene un coeficiente de determinación de 0.92, lo cual indica que es bastante adecuado para predecir lo valores de la variable predictora (ventas). De igual forma, el coeficiente de determinción ajustado es muy similar, indicando así que la variable es de gran significancia para el modelo.

Prueba de normalidad

En la prueba de normalidad, para ver cómo se distribuyen los residuos, se puede plantear una prueba de hipótesis, en donde:

- H_0 : los datos provienen de una población normal
- H_1 : los datos no provienen de una población normal

Aplicando la prueba de normalidad de Shapiro, se obtiene el siguiente QQ-plot de la figura 14

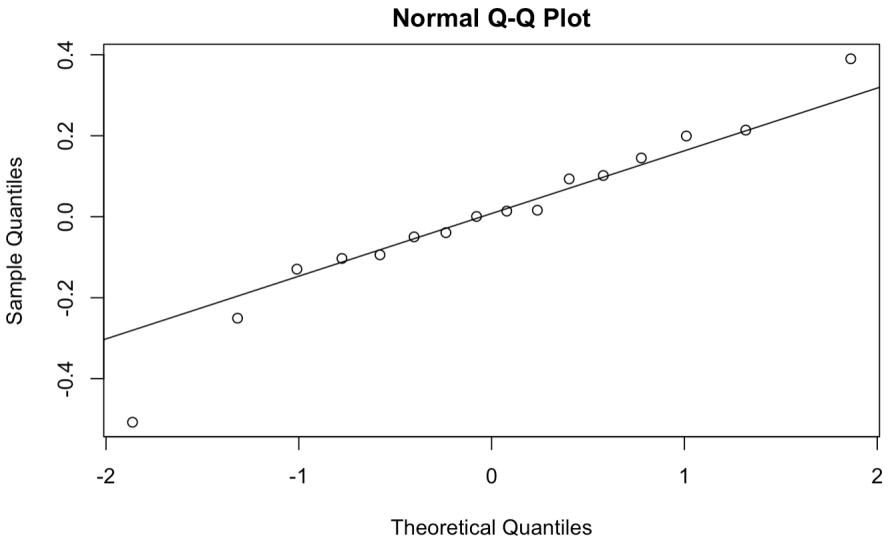

Figura 14: Diagrama QQ-plot.

Se observa en el Q-Q plot que los residuos están en una línea recta, siendo así la forma ideal de los residuos cuando pertenecen a una población normal.

Media de los residuos igual a 0

Para la suma de los residuos, se puede aplicar de igual forma una prueba de hipótesis, u obtener un intervalo de confianza. Como se puede observar en la figura 15, dentro del intervalo de confianza creado se encuentra el 0, por lo tanto se dice que la media de los residuos es equivalente a 0.

```
        One Sample t-test

data:  N3$residuals
t = 1.3931e-16, df = 15, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.1094818  0.1094818
sample estimates:
   mean of x
7.155734e-18
```

Figura 15: Análisis de la media de los residuos.

Homocedasticidad

Finalmente, para la homocedasticidad, se observa en la figura 16 que los residuos siguen un patrón como tal, o más bien no muestran una estructura evidente. Por lo tanto el modelo es adecuado ya que se muestra independencia.

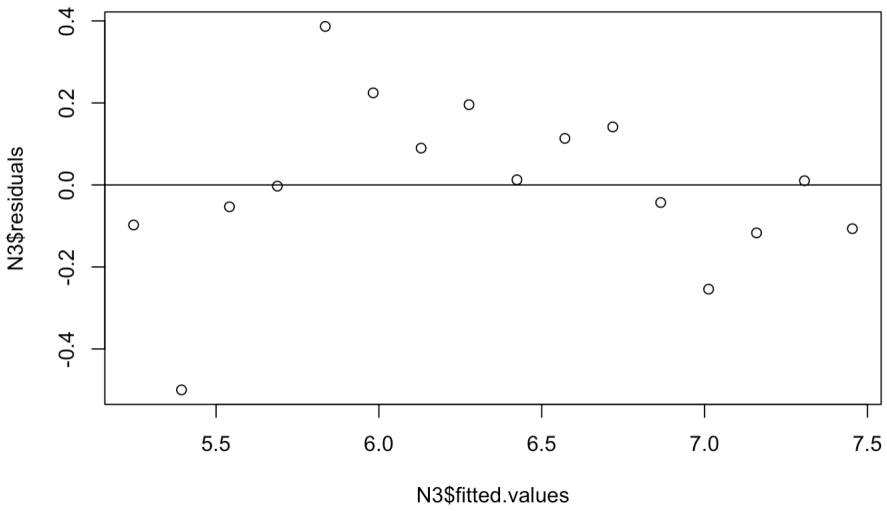

Figura 16: Análisis de homocedasticidad.

Cálculos del CME y el EPAM

Un punto importante para evaluar qué tan acertados son los pronósticos, es obtener el promedio de los cuadrados de los errores (CME) y el promedio de los errores porcentuales (EPAM). La primer métrica se refiere al promedio de los errores de pronóstico al cuadrado, los cuales se calculan restanto de valor real, el valor pronosticado. Por otro lado, el EPAM se obtiene del promedio del valor absoluto del los errores porcentuales, que se obtienen dividiendo el error de pronóstico entre el valor real.

Para calcularlos, primero se debe de utilizar el modelo generado para realizar las predicciones, las cuales se pueden visualizar de forma gráfica en la figura 17.

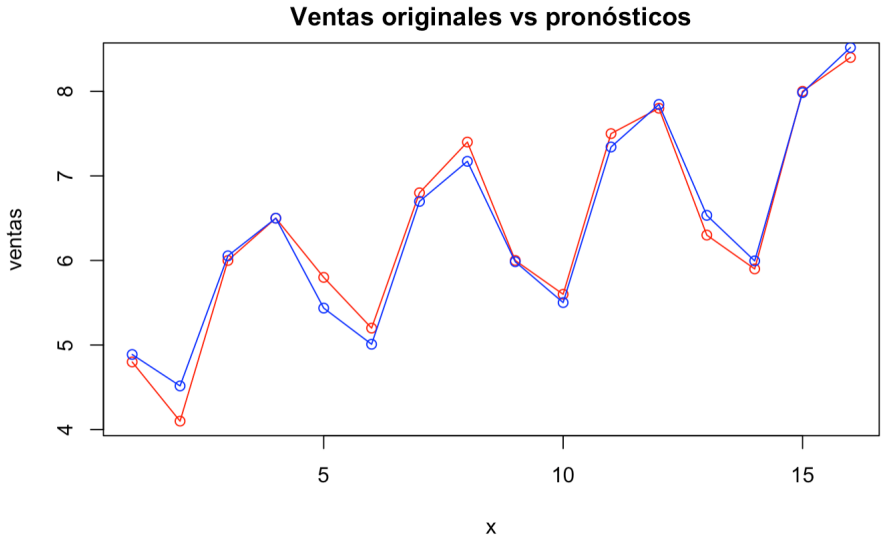

Figura 17: Predicciones realizadas vs valores reales.

Haciendo los cálculos correspondientes, se obtiene que el $CME = 0.03308245$, mientras que el $EPAM = 0.02438393$, los cuales son muy pequeños, lo cual es bueno, pues indica la eficiencia del modelo para predecir.

Pronóstico para el siguiente año

Finalmente, se aplica el modelo de regresión lineal construido y validado para predecir las ventas de los trimestres del siguiente año. En la figura 18 se muestra dichos resultados.

```
Pronóstico del primer trimestre del quinto año: 7085.003
Pronóstico del segundo trimestre del quinto año: 6490.456
Pronóstico del tercer trimestre del quinto año: 8631.444
Pronóstico del cuarto trimestre del quinto año: 9194.001
```

Figura 18: Pronóstico de ventas para el siguiente año.

Los resultados anteriores son las ventas en miles que se estiman para los siguientes cuatro trimestres. Todos esto fue utilizando el modelo de regresión lineal construido tomando como base las ventas desestacionalizadas.

Conclusiones

En conclusión, el modelo lineal generado es bastante bueno, ya que además de superar con éxito todos los supuestos, tiene un alto valor de coeficiente de determinación. Claramente el modelo se aplicó a las ventas desestabilizadas por lo tanto al trabajar con series de tiempo siempre es buena idea aplicar métodos de suavizamiento y posteriormente los modelos correspondientes.

De igual forma, con los valores de CME y EPAM, se observa que son bastante bajos, lo cual indica que los pronósticos se están realizando de manera satisfactoria.

Gracias a las técnicas implementadas sobre la serie de tiempo, fue posible cumplir con el objetivo de la actividad, el cual era encontrar una forma de pronosticar las ventas de televisores en los siguientes años. Además, gracias al análisis de igual forma ahora se conoce que el trimestre con mayores ventas es el cuarto. Esto se puede deber a que a finales del año se tienen varios eventos como el buen fin, o las fiestas navideñas que pueden aumentar la venta de estos artículos.

Anexos

Liga a la carpeta de Google Drive que contiene el documento de R en donde se realizó el análisis.
`https://drive.google.com/drive/folders/10GSomkaam_ZgFvZnwFY_H4PQi41S_OIJ?usp=sharing`

Referencias

[1] S. de Minitab. ¿qué es una serie de tiempo? [Online]. Available: https://support.minitab.com/es-mx/minitab/21/help-and-how-to/statistical-modeling/time-series/supporting-topics/basics/what-is-a-time-series/

[2] R. M. Granadosn. Variables no estacionarias y cointegracion. [Online]. Available: https://www.ugr.es/~montero/matematicas/cointegracion.pdf