A8 - Series de Tiempo

Juan Bernal

2024-11-14

Para los datos de las ventas de televisores analiza la serie de tiempo más apropiada:

Año	Trimestre	Ventas_miles
1	1	4.8
1	2	4.1
1	3	6.0
1	4	6.5
2	1	5.8
2	2	5.2

1. Realiza el análisis de tendencia y estacionalidad:

* Identifica si es una serie estacionaria

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

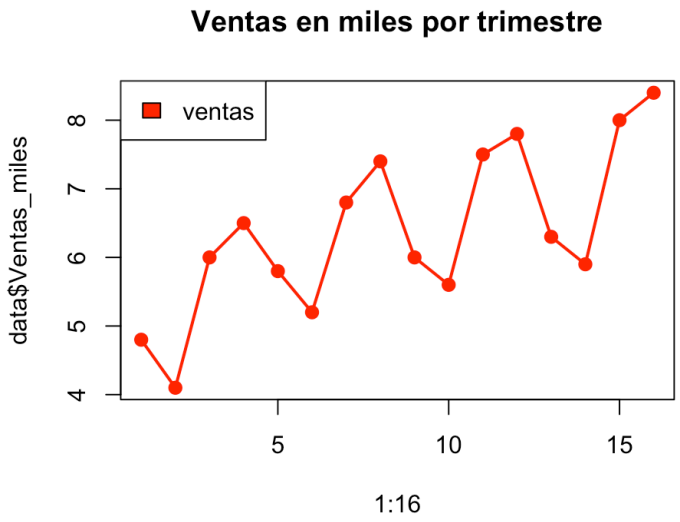
```
##
##  Augmented Dickey-Fuller Test
##
## data:  data$Ventas_miles
## Dickey-Fuller = -2.7111, Lag order = 2, p-value = 0.3015
## alternative hypothesis: stationary
```

Prueba de estacionalidad:

- $H_0: $ Es estacional, por lo que no es estacionaria.
- $H_1: $ Es estacionaria, por lo que no es estacional.

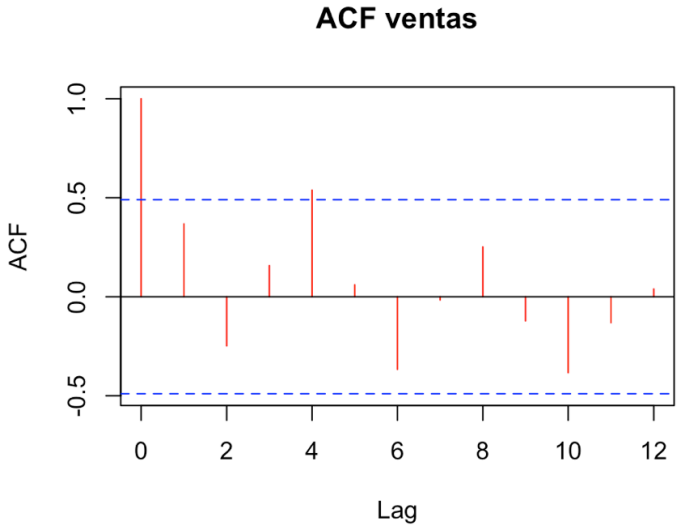
Dado el p-value obtenido en la prueba de Dickey-Fuller y suponiendo un nivel de significancia estándar de $\alpha = 0.05$, entonces podemos decir que no contamos con suficiente evidencia para rechazar la hipótesis inicial, por lo que la serie es estacional y no estacionaria.

* Grafica la serie para verificar su tendencia y estacionalidad



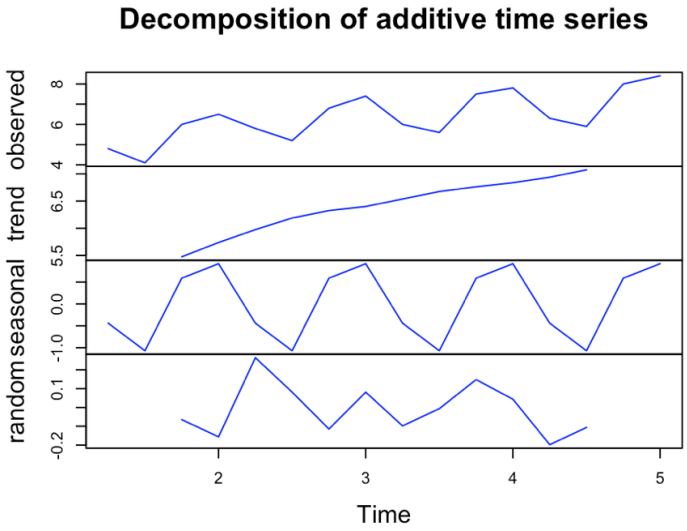
La gráfica muestra las ventas en miles por trimestre. Observando la tendencia general, parece que hay un incremento en las ventas a lo largo del tiempo con cierta variabilidad estacional entre los trimestres. Este patrón sugiere que, aunque las ventas aumentan de manera general, existen fluctuaciones regulares que podrían corresponder a la estacionalidad de la serie.

* Analiza su gráfico de autocorrelación

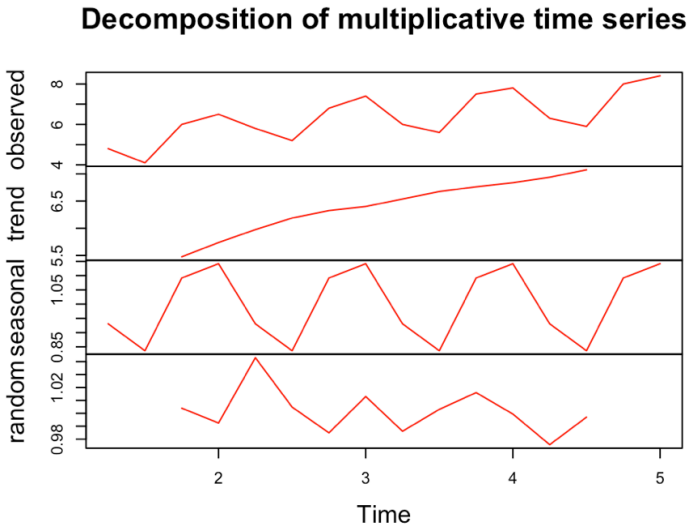


La gráfica de autocorrelación (ACF) de las ventas muestra una alta correlación en el lag 1, indicando que las ventas de un trimestre están fuertemente relacionadas con las del trimestre anterior. También hay correlaciones en los lags 2 y 4, lo que sugiere un posible patrón estacional anual.

* Identifica si el modelo puede ser sumativo o multiplicativo (puedes probar con ambos para ver con cuál es mejor el modelo)



Observemos que en la gráfica azul los componentes de estacionalidad y error parecen mantenerse en un rango similar a través del tiempo.



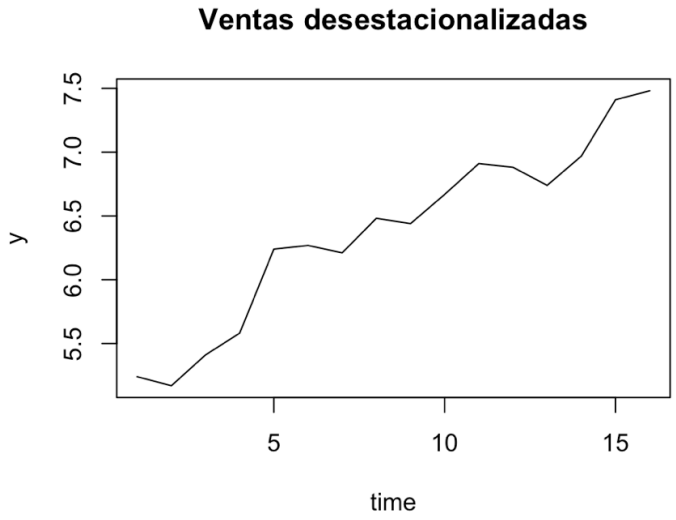
En la gráfica del modelo multiplicativo, el componente estacional y el aleatorio parecen mantener proporciones relativas al valor observado.

Dado que el componente estacional no muestra cambios significativos en su amplitud en ambas gráficas, el modelo aditivo parece ser más adecuado en este caso, ya que la estacionalidad y los errores no aumentan proporcionalmente con el nivel de la serie.

2. Calcula los índices estacionales y grafica la serie desestacionalizada

```
##         Qtr1       Qtr2       Qtr3       Qtr4
## 1            -0.4395833 -1.0687500  0.5895833
## 2  0.9187500 -0.4395833 -1.0687500  0.5895833
## 3  0.9187500 -0.4395833 -1.0687500  0.5895833
## 4  0.9187500 -0.4395833 -1.0687500  0.5895833
## 5  0.9187500
```

Los trimestres con mejores ventas son el primero y el cuarto, mientras que los trimestres de menor actividad son el segundo y, en particular, el tercero, que muestra una baja significativa respecto al promedio. Esto sugiere un patrón estacional donde las ventas son relativamente más altas a comienzos y finales del año, y caen a mitad del año.



Se observa un crecimiento general en las ventas a medida que avanza el tiempo, con algunas oscilaciones que reflejan incrementos y estabilizaciones en ciertos puntos. Esto sugiere una tendencia positiva en las ventas tras eliminar los efectos de la estacionalidad.

3. Analiza el modelo lineal de la tendencia

* Realiza la regresión lineal de la tendencia (ventas desestacionalizadas vs tiempo)

```
##
## Call:
## lm(formula = y ~ time)
##
## Coefficients:
## (Intercept)         time
##      5.1392       0.1461
```

El modelo de regresión lineal de la tendencia obtenido es:

- $ventas = 5.1392 + 0.1461 \times Trimestre$

Observando que el aumento de las ventas por trimestre es gradual.

* Analiza la significancia del modelo lineal, global e individual

```
##
## Call:
## lm(formula = y ~ time)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2992 -0.1486 -0.0037  0.1005  0.3698
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  5.13917    0.10172   50.52 < 0.0000000000000002 ***
## time         0.14613    0.01052   13.89         0.0000000014 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.194 on 14 degrees of freedom
## Multiple R-squared:  0.9324, Adjusted R-squared:  0.9275
## F-statistic:   193 on 1 and 14 DF,  p-value: 0.000000001399
```

Prueba de significancia de los coeficientes:

*$H_0 : \beta_i = 0$. Los coeficientes no son significantes para el modelo.

*$H_1 : \exists \beta_i \neq 0$. Al menos un coeficiente es significante para el modelo.

Dado el p-value de los coeficientes y suponiendo un nivel de significancia de $\alpha = 0.05$, entonces contamos con suficiente evidencia para rechazar la hipótesis inicial, por lo que todos los coeficientes son significantes.

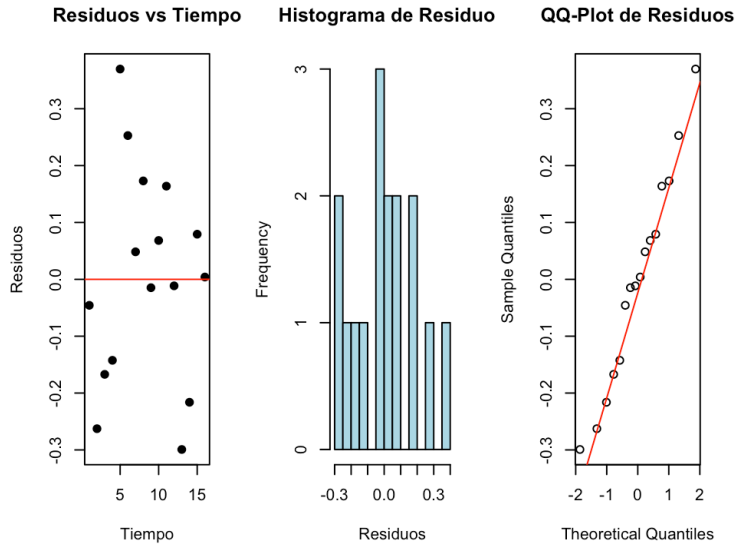
Prueba de significancia del modelo:

*$H_0 : \beta = 0$. Los coeficientes no son significantes para el modelo.

*$H_1 : \beta \neq 0$. Al menos un coeficiente es significante para el modelo.

Dado el p-value del modelo y suponiendo un nivel de significancia de $\alpha = 0.05$, entonces contamos con suficiente evidencia para rechazar la hipótesis inicial, por lo que el modelo es significante.

* Haz el análisis de residuos



En la primera gráfica los residuos parecen estar distribuidos de manera algo aleatoria alrededor de la línea roja en 0, pero hay algunos puntos que están alejados más del promedio y un posible patrón no completamente aleatorio. Esto puede indicar que hay ciertos factores en el modelo que no están siendo bien capturados, como un patrón temporal o autocorrelación.

En la segunda gráfica el histograma muestra una distribución de residuos centrada cerca de 0, lo que es deseable. Sin embargo, parece haber una ligera asimetría en la distribución. Si bien la mayoría de los valores están cerca de la media, algunos residuos están un poco más lejos de lo esperado si el supuesto de normalidad es importante en este análisis.

Por último, en la qqplot la mayoría de los puntos se alinean bien con la línea de referencia, lo que indica que los residuos aproximadamente siguen una distribución normal. Sin embargo, hay algunas desviaciones en los extremos, especialmente hacia la parte inferior, que sugieren posibles valores atípicos o desviaciones del supuesto de normalidad.

```
## Loading required package: carData
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.3762583      1.243485   0.042
##  Alternative hypothesis: rho != 0
```

El resultado sugiere que los residuos del modelo están autocorrelacionados, lo que implica que las suposiciones del modelo de independencia de los residuos podrían no cumplirse. Esto podría indicar que el modelo necesita ajustes adicionales, como el uso de términos de regresión adicionales o modelos específicos para manejar la autocorrelación.

El análisis sugiere que el modelo ajusta razonablemente bien, pero puede haber áreas que necesitan ajustes. Hay indicios de posibles patrones en los residuos que sugieren que ciertos elementos podrían no estar siendo capturados. La ligera desviación de la normalidad y algunos puntos que se alejan de la línea en el QQ-plot indican que puede ser necesario ajustar o revisar el modelo para mejorar el ajuste, posiblemente añadiendo variables o ajustando su estructura.

4. Calcula el CME y el EPAM de la predicción de la serie de tiempo

```
## CME: 0.6971073
```

```
## EPAM: 12.66042 %
```

El modelo tiene un CME de 0.6971, lo que indica que el error cuadrático medio de las predicciones es moderado, reflejando un nivel razonable de precisión. El EPAM de 12.66 % señala que el modelo, en promedio, se desvía un 12.66 % de los valores reales, lo que representa un error relativamente aceptable según el contexto. En general, el modelo muestra un desempeño decente, pero puede haber margen para mejoras en la precisión.

5. Explora un mejor modelo, por ejemplo un modelo cuadrático. Para ello transforma la variable ventas (recuerda que la regresión no lineal es una regresión lineal con una tranformación).

```
##
## Call:
## lm(formula = y ~ time + timesquare)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30333 -0.13440 -0.01928  0.11368  0.33301
##
## Coefficients:
##              Estimate Std. Error t value          Pr(>|t|)
## (Intercept)  4.930833   0.155679  31.673 0.000000000000108 ***
## time         0.215572   0.042149   5.115          0.000199 ***
## timesquare  -0.004085   0.002410  -1.695          0.113918
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1822 on 13 degrees of freedom
## Multiple R-squared:  0.9446, Adjusted R-squared:  0.9361
## F-statistic: 110.8 on 2 and 13 DF,  p-value: 0.000000006805
```

```
## CME del modelo cuadrático: 0.6970679
```

```
## EPAM del modelo cuadrático: 12.66041 %
```

Prueba de significancia de los coeficientes:

*H_0 : $\beta_i = 0$. Los coeficientes no son significantes para el modelo.

*H_1 : $\exists \beta_i \neq 0$. Al menos un coeficiente es significante para el modelo.

Dado el p-value de los coeficientes y suponiendo un nivel de significancia de $\alpha = 0.05$, entonces contamos con suficiente evidencia para rechazar la hipótesis inicial en los coeficientes del intercepto y el tiempo, por lo que el coeficiente cuadrático no es significante.

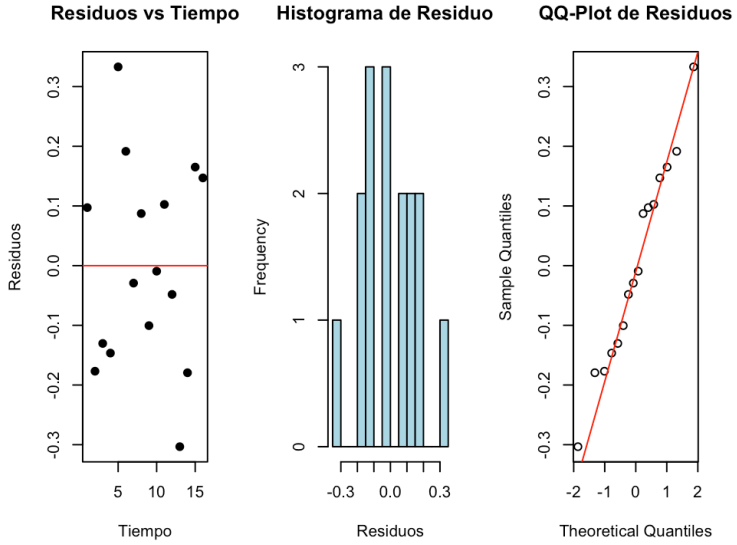
Prueba de significancia del modelo:

*H_0 : $\beta = 0$. Los coeficientes no son significantes para el modelo.

*H_1 : $\beta \neq 0$. Al menos un coeficiente es significante para el modelo.

Dado el p-value del modelo y suponiendo un nivel de significancia de $\alpha = 0.05$, entonces contamos con suficiente evidencia para rechazar la hipótesis inicial, por lo que el modelo cuadrático es significante.

El modelo tiene un CME de 0.6971, lo que indica que el error cuadrático medio de las predicciones es moderado, reflejando un nivel razonable de precisión. El EPAM de 12.66 % señala que el modelo, en promedio, se desvía un 12.66 % de los valores reales, lo que representa un error relativamente aceptable según el contexto. En general, el modelo muestra un desempeño decente, pero puede haber margen para mejoras en la precisión.



En la primera gráfica los puntos de los residuos parecen estar distribuidos de manera aleatoria alrededor de la línea roja de referencia (en 0). Esto sugiere que no hay un patrón obvio en los residuos a lo largo del tiempo, lo cual es deseable y respalda la suposición de que los residuos son independientes. No obstante, hay algunos puntos ligeramente alejados que pueden indicar pequeños ajustes necesarios.

El histograma muestra que la mayoría de los residuos están centrados cerca de 0, lo cual es un buen indicio. Sin embargo, parece haber una ligera concentración hacia un lado, pero en general, su distribución es aproximadamente simétrica, indicando que el modelo ajusta bastante bien los datos.

El QQ-Plot muestra que los puntos se alinean bastante bien con la línea de referencia, lo que sugiere que los residuos siguen una distribución aproximadamente normal. Hay algunos puntos en los extremos que se desvían de la línea, lo que puede indicar posibles valores atípicos o que hay ligeras desviaciones de la normalidad.

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1895177      1.548932   0.126
##  Alternative hypothesis: rho != 0
```

Prueba de independencia:

*H_0 : Los residuos no están correlacionados

*H_1 : Los residuos están correlacionados

Dado el p-value obtenido en la prueba de independencia de Durbin-Watson y suponiendo un nivel de significancia de $\alpha = 0.05$, entonces podemos decir que no tenemos suficiente evidencia para rechazar la hipótesis nula, por lo que el modelo cumple con el supuesto de independencia de los residuos.

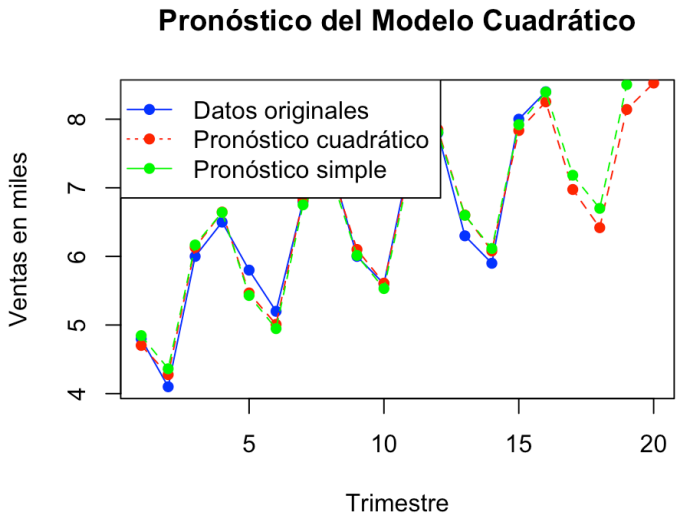
El modelo cuadrático parece ajustarse razonablemente bien a los datos, ya que los residuos no muestran un patrón evidente en el tiempo, y su distribución es aproximadamente normal. No obstante, hay algunos indicios menores de posible asimetría o valores atípicos que podrían investigarse para mejorar el ajuste.

6. Concluye sobre el mejor modelo

El modelo cuadrático (y ~ time + timesquare) es el mejor modelo para describir la relación entre la variable dependiente “ventas en miles” y el tiempo en trimestres. Esto se debe a que ofrece un mejor ajuste a los datos, evidenciado por su R-cuadrado ligeramente mayor al del modelo simple (93.61% frente a 92.75%), aunque su residuo estándar es ligeramente mayor (0.194 frente a 0.1822) y, además, el término cuadrático no es estadísticamente significativo. Sin embargo, el modelo cuadrático cumple con el supuesto de los residuos de mejor manera que el modelo simple, por lo que es mejor para manejar la autocorrelación, aún y cuando el comportamiento sea aparentemente no lineal.

Por lo tanto, el modelo cuadrático describe mejor el comportamiento de los datos y cumple con el supuesto de validez, aunque es más complejo que el modelo simple, su término cuadrático no es significante y parece hacer un sobreajuste a los picos con curvas.

7. Realiza el pronóstico para el siguiente año y grafícalo junto con los pronósticos previos y los datos originales.



El pronóstico cuadrático (rojo) sigue de manera bastante precisa las subidas y bajadas de las ventas reales. Al usar un término cuadrático, puede adaptarse a la forma ondulada de los datos y captar las variaciones más bruscas en las ventas. El pronóstico simple (verde) también sigue la forma general de los datos, pero es menos preciso en los puntos extremos (picos y valles). Su respuesta a los cambios en los datos es más gradual, por lo que no captura con tanta precisión las oscilaciones rápidas.

En conclusión, la gráfica sugiere que el modelo cuadrático proporciona un pronóstico más ajustado a los datos reales en comparación con el pronóstico simple.