Reporte final de venta de televisores

Amy Murakami Tsutsumi - A01750185

2022-11-14

Inteligencia Artifical para la Ciencia de Datos II (Grupo 501)

Módulo 5: Estadística Avanzada para ciencia de datos

1. Resumen

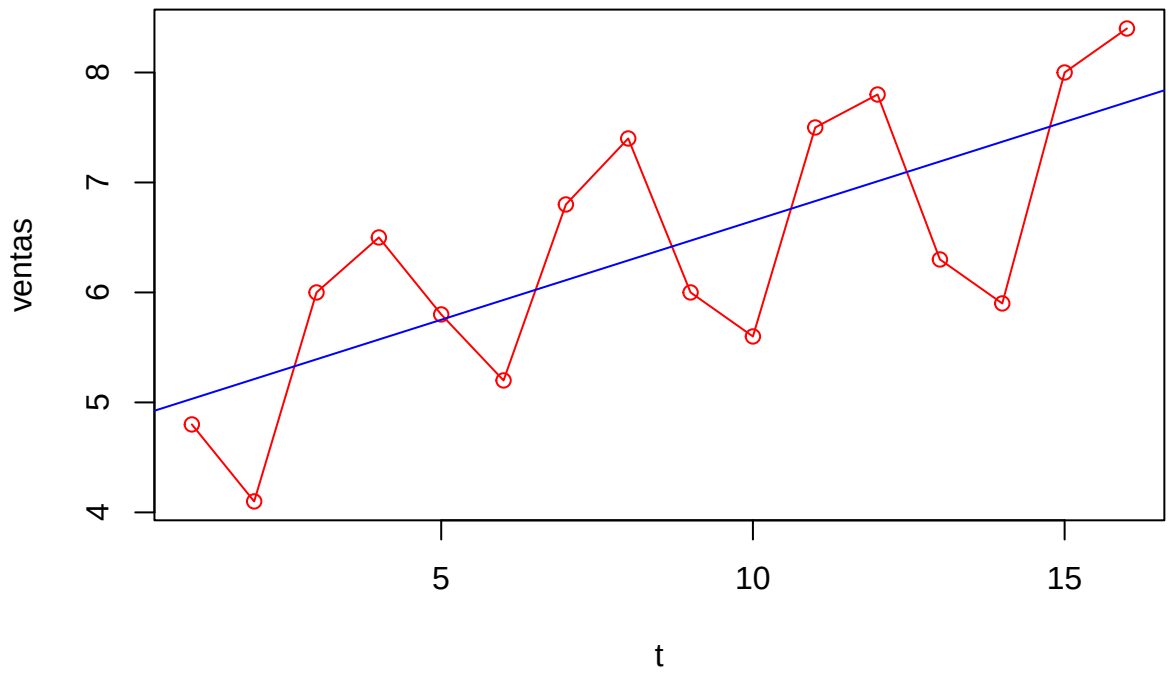
La problemática a resolver es realizar un análisis de tendencia de una serie de tiempo que representa las ventas de televisores. Los métodos y técnicas estadísticas utilizadas son la función decompose, regresión lineal, análisis de verificación del modelo lineal (significancia de beta1, variabilidad explicada por el modelo, análisis de los residuos y prueba de normalidad), método de promedios móviles y cálculo del CME y EPAM. Después de este proceso de análisis se obtuvo un modelo de regresión lineal adecuado (con dependencia, normal, media simétrica y homocedasticidad) al igual que un modelo creado por promedios móviles que se asemeja a los datos de las ventas reales.

2. Introducción

El portafolio de análisis tiene como propósito utilizar herramientas estadísticas vistas en el módulo cinco para poder analizar e interpretar los resultados de las predicciones de los modelos. El problema consiste en elaborar un análisis de tendencia de una serie de tiempo no estacionaria para poder realizar el pronóstico del siguiente año. Este análisis es importante ya que se utilizarán las herramientas aprendidas para poder generar las predicciones más cercanas a los datos reales considerando que son series de tiempo que cambian en el tiempo, es decir, no oscilan alrededor de un valor constante.

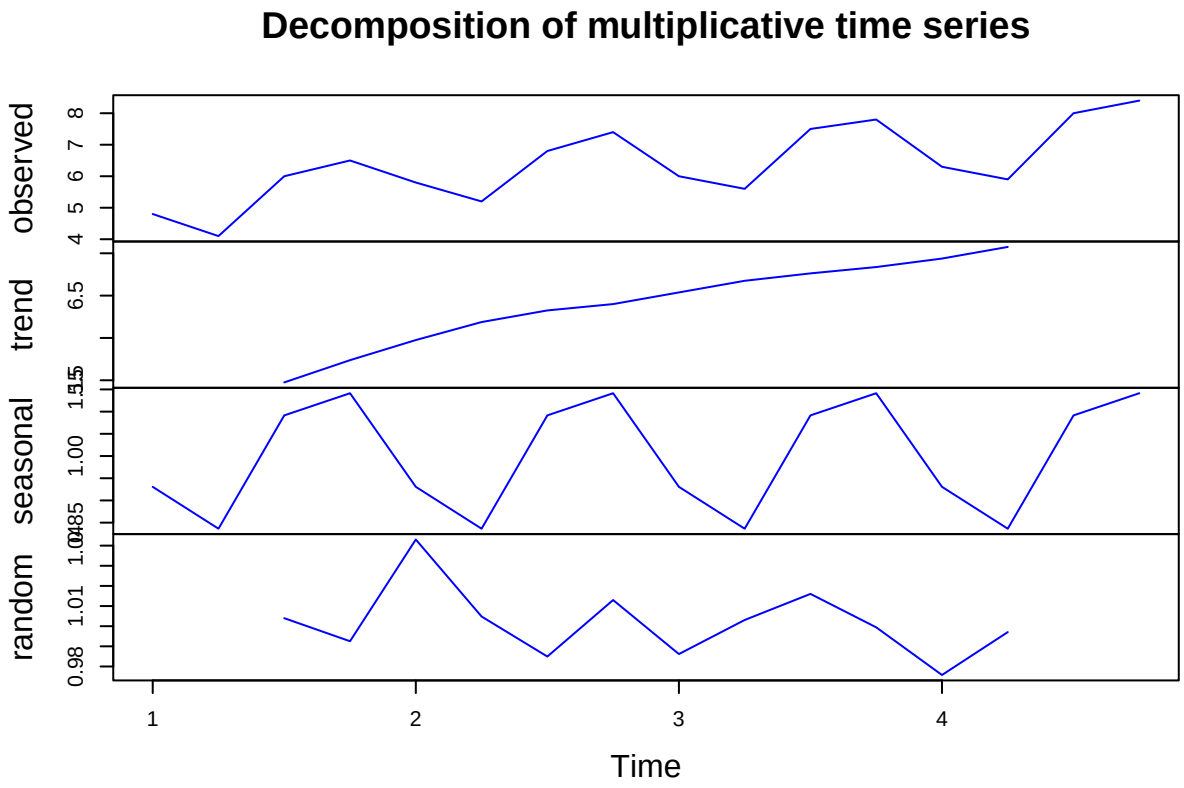
3. Análisis de resultados

Realiza el gráfico de dispersión. Observa la tendencia y los ciclos. Realiza el análisis de tendencia y estacionalidad



Esta primera gráfica muestra una serie de tiempo no estacionaria de color rojo, se puede notar una tendencia a crecer a largo plazo y que la media incrementa con el paso del tiempo lo que significa que la serie no oscila alrededor de un valor constante. Además, de color azul se muestra una tendencia lineal (ventas = 4.8525 + 0.1799 t). Sin embargo, esta tendencia no considera la estacionalidad, es decir, los ciclos a lo largo de los años.

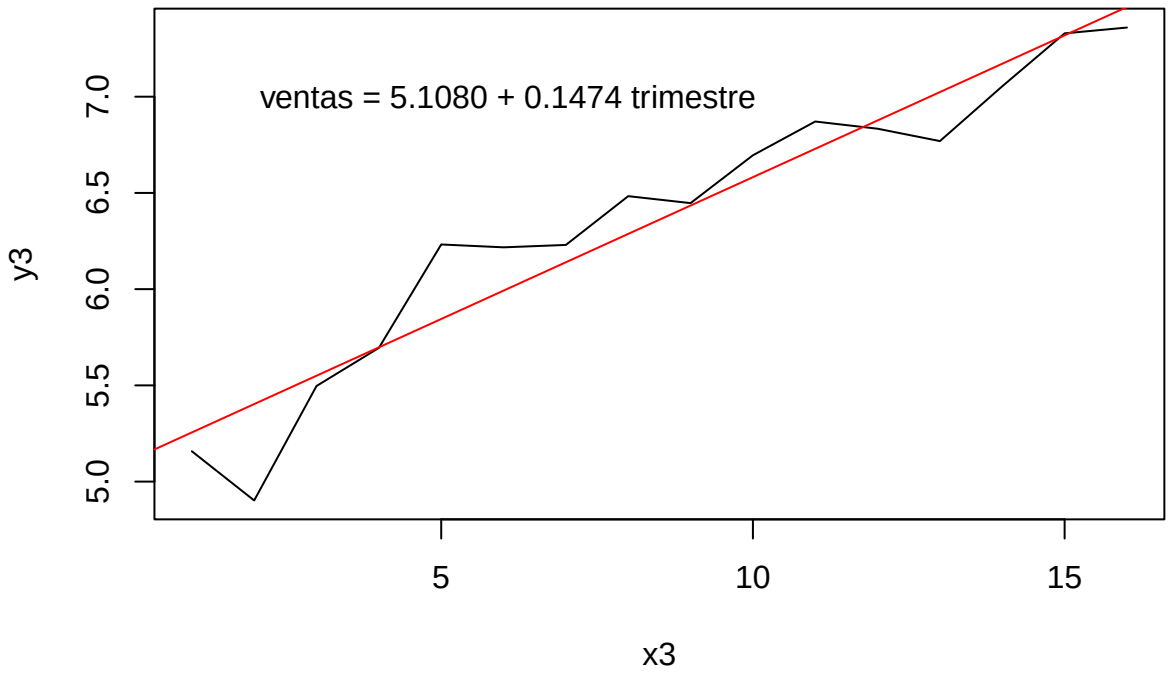
Descompón la serie en sus 3 componentes e interprétalos



Al descomponer la serie en los tres componentes se tiene en la parte superior el observado que es la suma de los componentes de tendencia, el efecto estacional y los residuos. El siguiente es el de tendencia que muestra un patrón gradual y consistente de las variaciones, es decir, es la variación a largo plazo. Esta gráfica se muestra creciente a lo largo del tiempo. Más adelante se tiene la variación estacional y se puede observar que es una variación cíclica ya que se comporta con la misma intensidad durante los distintos meses del año de manera periódica. Por último, está la variación irregular que es una variación imprevisible y no recurrente en el tiempo que describe las variaciones a corto plazo, por lo tanto, se puede notar que no tiene la misma duración que los demás componentes y no se puede predecir sus valores.

Analiza el modelo lineal de la tendencia:

Realiza la regresión lineal de la tendencia (ventas desestacionalizadas vs tiempo). Dibuja la recta junto con las ventas desestacionalizadas.



Utilizando los valores que están en la columna de Estimate y los coeficientes se tiene que el valor del intercepto es $\hat{\beta}_0 = 5.10804$ y el valor de la pendiente es $\hat{\beta}_1 = 0.1474$. Por lo tanto, la ecuación de regresión de mejor ajuste para los ventas es: $y = 5.10804 + 0.1474x$. Se puede notar una gran mejora en está gráfica utilizando la regresión lineal de la tendencia a comparación de las primeras gráficas.

Analiza la pertinencia del modelo lineal: Ahora se realizará el proceso para verificar el modelo de regresión lineal.

Significancia de beta1

a) Hipótesis Se desea comprobar que β_1 es significativamente diferente de 0 para asegurar que el modelo lineal es el apropiado. Por lo tanto, se tienen dos hipótesis estadísticas:

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

b) Regla de decisión Se rechaza H_0 si:

- Regla clásica: Si $|t^*| > |t_0|$

- Regla valor p: Si $p < \alpha$ con $\alpha = 0.05$

Entonces se calculará la tstudent:

```
## [1] "t0: -2.08596344726586"
```

c) Análisis del resultado

```
##                 2.5 %    97.5 %
## (Intercept) 4.8684475 5.3476363
## x3          0.1226037 0.1721603
```

```
##
##  One Sample t-test
##
## data:  N3$residuals
## t = -1.4751e-16, df = 15, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 97 percent confidence interval:
##  -0.1233257  0.1233257
## sample estimates:
##     mean of x
## -7.589415e-18
```

d) Conclusión No se rechaza H_0 porque:

- $|t^*| = 1.4751e-16$ es menor que $|t_0| = 2.08596344726586$
- Valor $p = 1$ es mayor que $\alpha = 0.05$

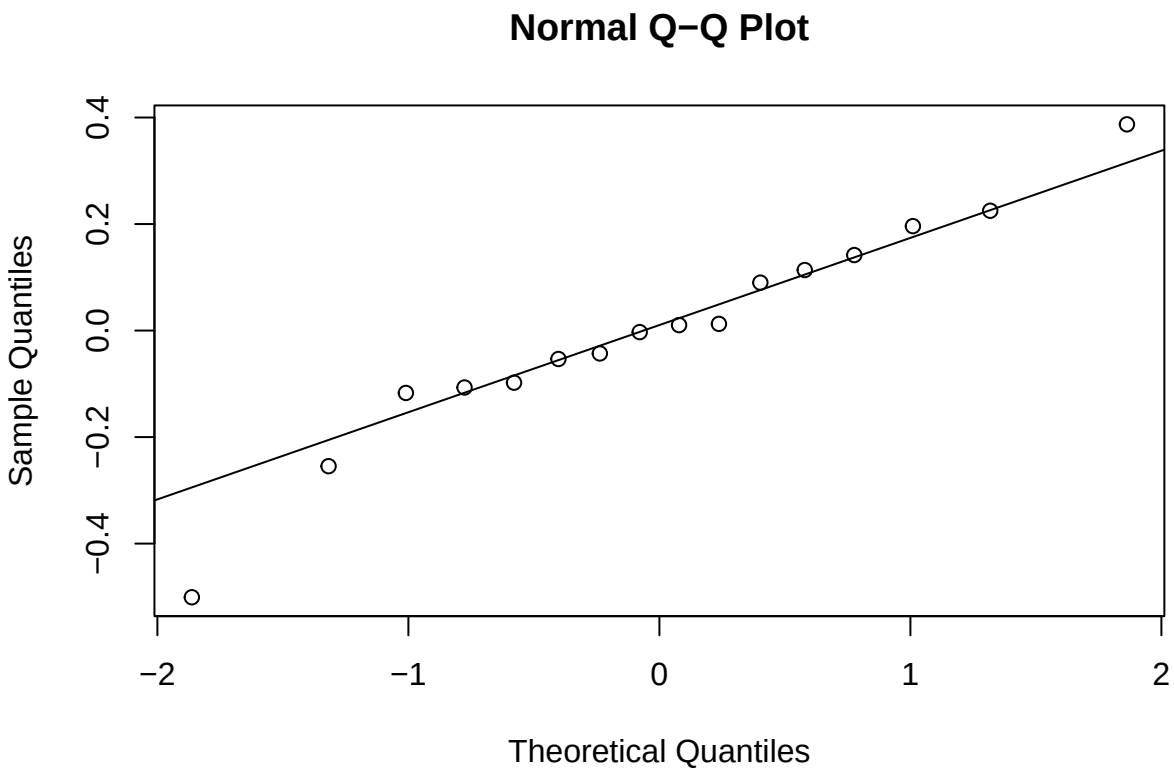
Por lo tanto, hay dependencia porque β_1 es significativamente diferente de 0. En otras palabras, hay efecto de X en Y.

Variabilidad explicada por el modelo (c.d) Podemos observar que el coeficiente de determinación para los datos es de 0.9208, es decir, la variable predicadora explica el 92.08% de la variación. Esto indica que tiene una proporción alta de que la variable Y se explique por el valor de la variable X al ser un valor cercano a 1. En otras palabras, una vez que se conoce X se tiene menos incertidumbre sobre el valor de Y.

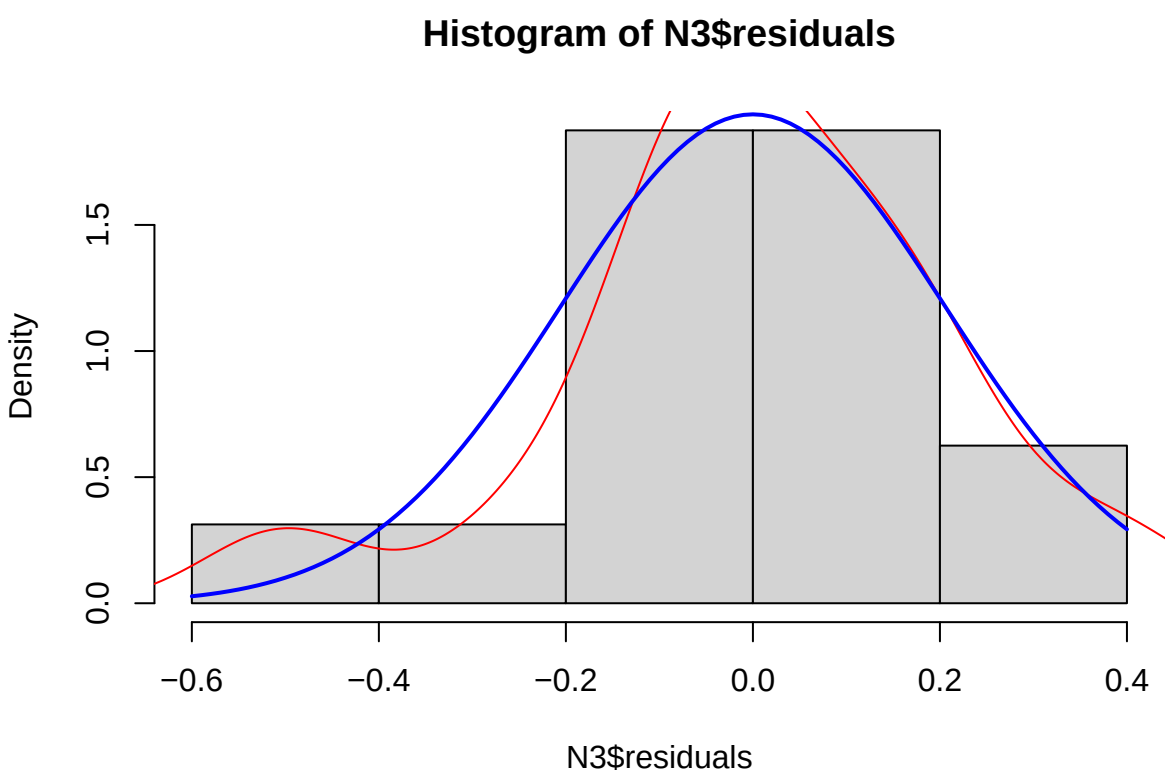
Análisis de los residuos

a) Normalidad de los residuos

```
##
##  Shapiro-Wilk normality test
##
## data:  N3$residuals
## W = 0.96379, p-value = 0.7307
```



Se puede observar que en la prueba de Shapiro-Wilks el valor de p es mayor al nivel de significancia de 5%, por lo tanto, se puede decir que la muestra proviene de una población con distribución normal. Utilizando la gráfica del Q-Q plot, podemos observar que es ideal y normal ya que se aproxima a la línea y=x.



Además, en el histograma, se puede observar que la regresión lineal con los datos de las ventas es casi simétrica ya que tiene un ligero sesgo a la izquierda.

b) Verificación de media cero

```
##
##  One Sample t-test
##
## data:  N3$residuals
## t = -1.4751e-16, df = 15, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.1096629  0.1096629
## sample estimates:
##     mean of x
## -7.589415e-18
```

Ya que los residuos se distribuyen como una normal, se puede aplicar la prueba de hipótesis para las medias.

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

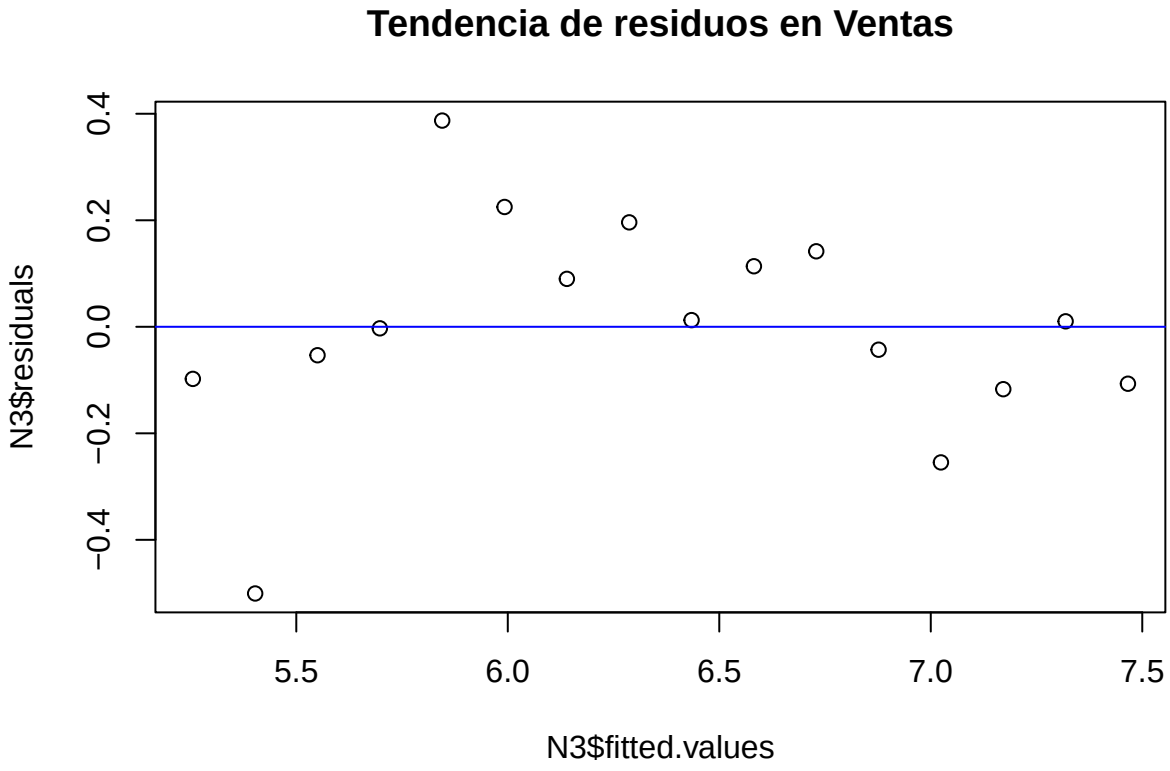
Para rechazar H_0 se deben cumplir lo siguiente:

- $|t^*|$ es mayor que $|t_0|$

- Valor p es menor que α

Sin embargo, al realizar la prueba de significancia de β_1 se probó que no se cumplen las condiciones, por lo tanto no se rechaza H_0.

c) Homocedasticidad La gráfica de tendencia de residuos de la regresión lineal con los datos de las ventas muestran que se tienen residuos aleatorios, es decir, no se puede distinguir una estructura evidente. Además, la dispersión de los residuos es constante en toda la gráfica, es decir que cumple con los supuestos. Esto quiere decir que el modelo de regresión es adecuado, tiene linealidad y homocedasticidad.



Calcula el CME y el EPAM (promedio de los errores porcentuales) de la predicción de la serie de tiempo.

Se realizará el cálculo del CME y el EPAM con dos métodos diferentes, el primero será la regresión lineal y el segundo por promedios móviles.

Los resultados utilizando la regresión lineal son:

```
## [1] "CME: 0.126637431638943"
```

```
## [1] "EPAM: 4.44643010531475"
```

Ahora utilizando promedios móviles, los resultados son:

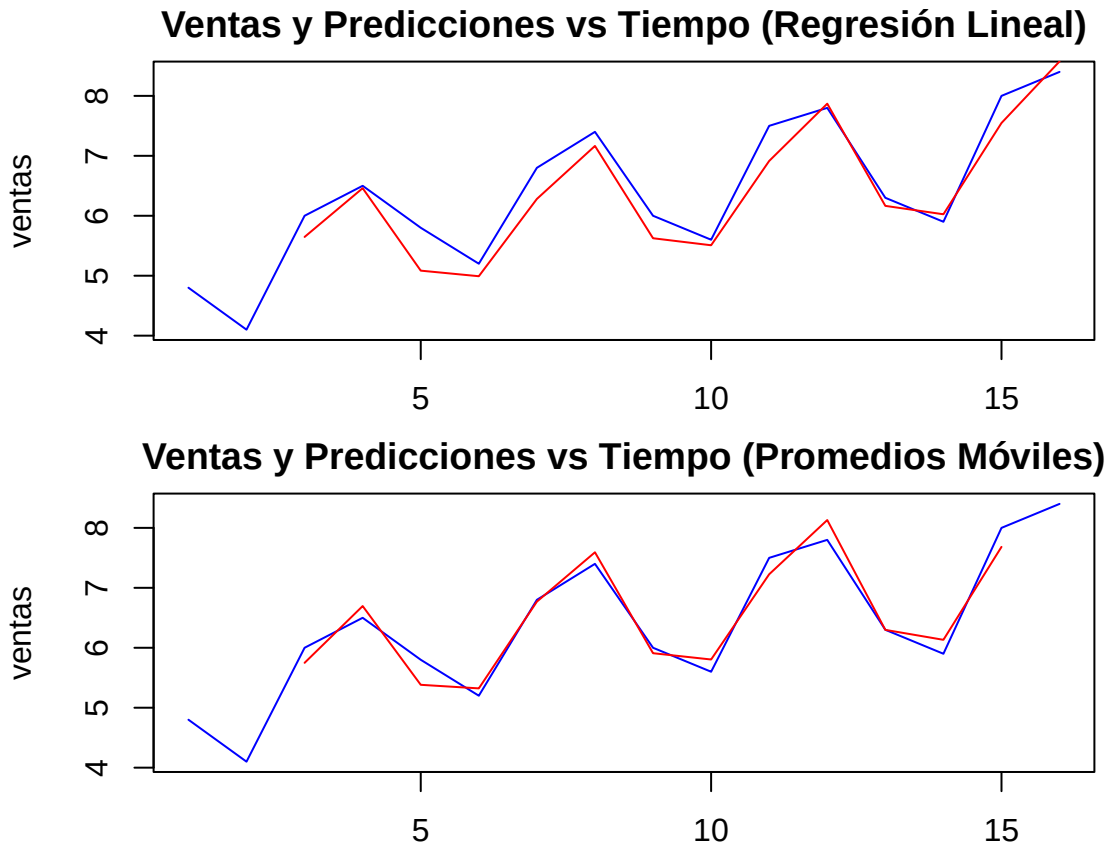
```
## [1] "CME: 0.0552123190119455"
```

```
## [1] "EPAM: 3.13744121568221"
```

Se puede observar que los valores obtenidos para el promedio de los cuadrados de los errores (CME) y el promedio de los errores porcentuales (EPAM) para los dos métodos son bastante pequeños. Sin embargo, el método de promedios móviles ofrece un menor CME (0.05521) y EPAM (3.1374) que el de la regresión lineal (CME = 0.12663 y EPAM 4.44643).

Dibuja el gráfico de los valores de las ventas y las predicciones vs el tiempo

Ahora se graficarán las ventas (en azul) y las predicciones (en rojo) a lo largo del tiempo utilizando los dos métodos mencionados anteriormente (regresión lineal y promedios móviles).



Al comparar las dos gráficas anteriores se puede notar que ambas se asemejan a los valores reales, pero la que más se aproxima es la que se generó con promedios móviles.

Realiza el pronóstico para el siguiente año.

```
## [1] 7085.872
```

```
## [1] 6491.284
```

```
## [1] 8632.585
```

```
## [1] 9195.263
```

4. Conclusión

Durante esta actividad se generó un modelo de regresión lineal para los datos de las ventas de televisores. Luego, se verificó la validez del modelo utilizando la significancia de $\hat{\beta}_1$, el coeficiente de determinación y análisis de residuos (normalidad, media cero y homocedasticidad). En este proceso de verificación se comprobó que hay dependencia, la probabilidad normal es ideal, la media es simétrica y hay homocedasticidad lo que indica que el modelo de regresión es adecuado. Incluso se realizó la predicción de las ventas de televisores utilizando la regresión lineal al igual que los promedios móviles; esto dió como resultado una gráfica de predicción que se asemeja bastante a los datos de las ventas reales. Por último, se calcularon los pronósticos del siguiente año con la regresión lineal.

5. Referencias bibliográficas

No se utilizaron referencias bibliográficas para este reporte.

6. Anexos

https://github.com/A01750185/E1-PortafolioAnalisis.git