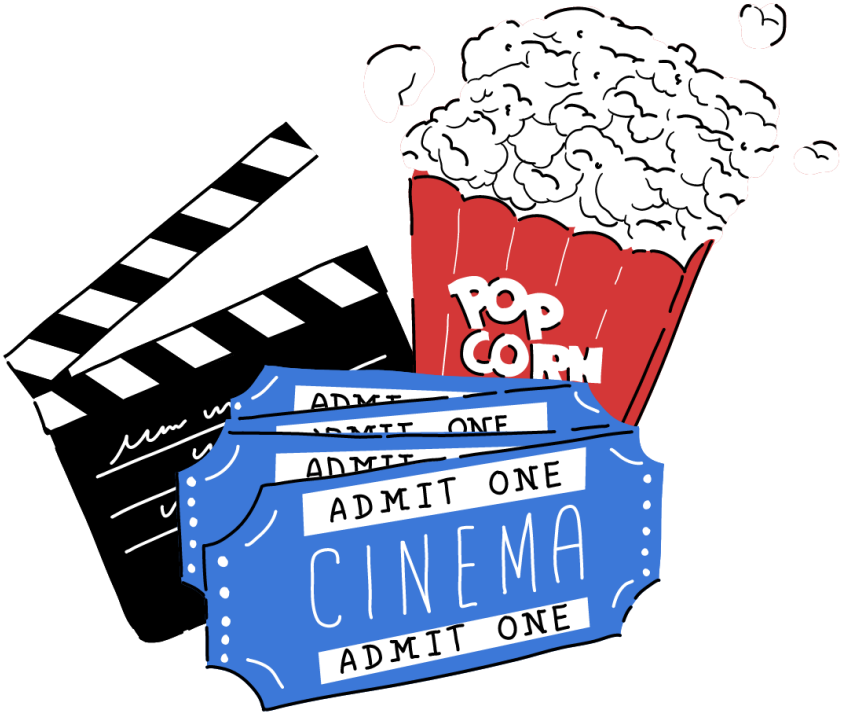
Análisis sobre la **recaudación** de las **salas de cine en España** (1988-2010)



Trabajo Académico MET II

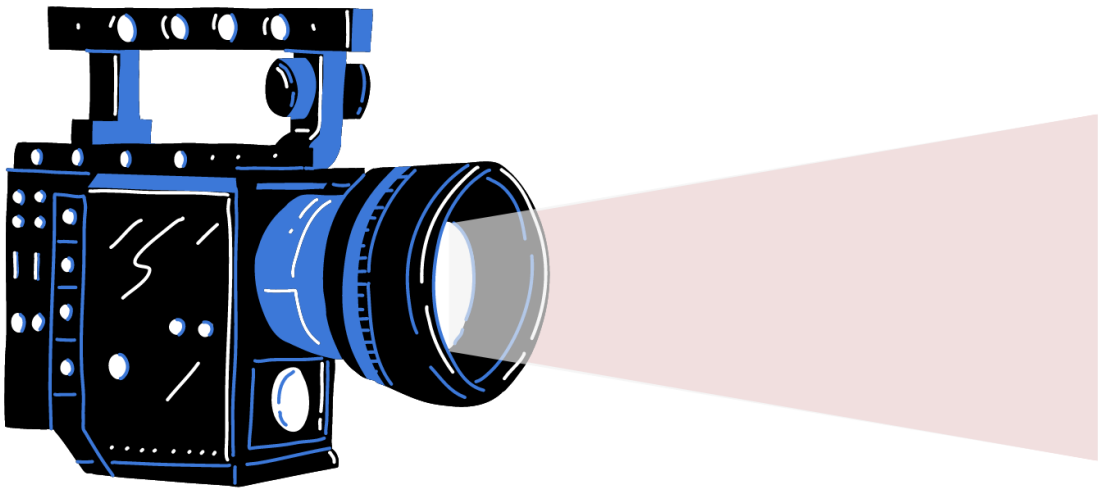
Realizado por Coral Montes, Juan Tomás y Júlia Vericat
Curso 2022/2023 - Ciencia de Datos

Índice

Introducción

En este trabajo buscamos explicar la recaudación de las salas de cine en nuestro país, mediante el número de pantallas existentes y el número de películas proyectadas. Para ello, disponemos de la recaudación anual producida (en millones de euros) desde 1988 hasta 2010, del número de salas de cine existentes en España y del número de películas proyectadas en dichas salas, y hemos utilizado varios modelos de regresión, con la ayuda del programa informático *StatGraphics*.

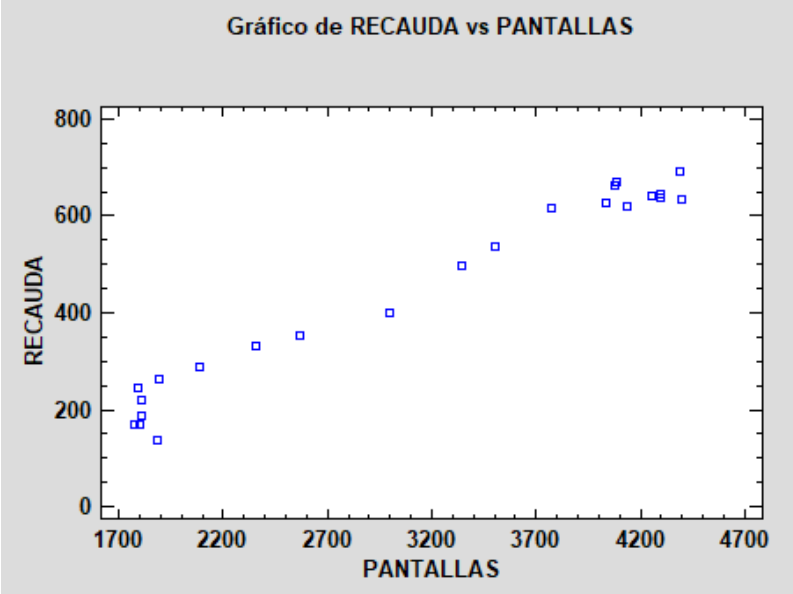
En primer lugar, propusimos un modelo lineal, como primer acercamiento a este problema, y a continuación establecimos un segundo modelo a partir de la información obtenida con el primero, con el que pudimos extraer las conclusiones que explicamos en el último apartado del informe.

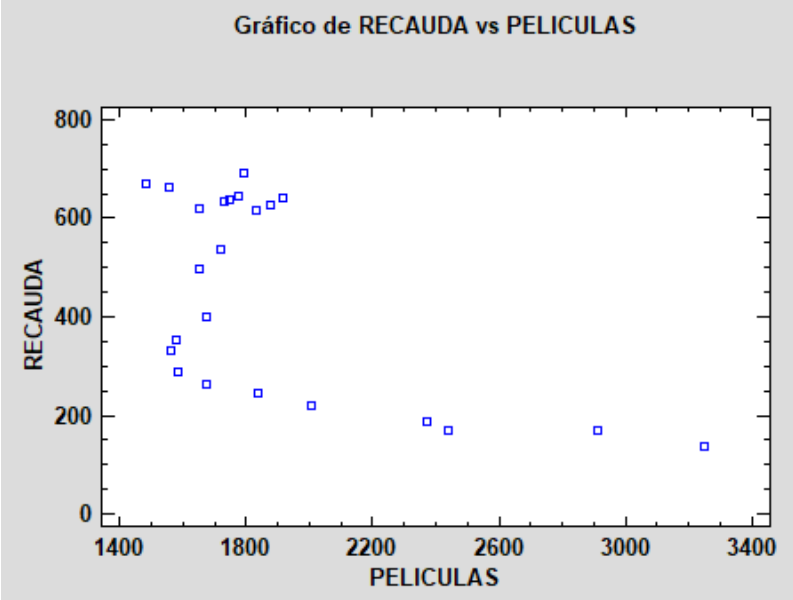


Primer modelo

Proposición del modelo. Variables

Viendo la relación existente entre la recaudación de las salas (RECAUDA) con el número de pantallas (PANTALLAS) y sus proyecciones (PELÍCULAS) mediante estos gráficos de dispersión, podemos proponer un primer modelo lineal:





Como podemos ver en el primer gráfico, entre la cantidad recaudada y el número de pantallas parece existir una relación lineal. En cambio, en el segundo gráfico no podemos apreciar dicha relación entre la recaudación y el número de películas. Por tanto, el **modelo lineal propuesto** es el siguiente:

$$RECAUDA = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$$

$$RECAUDA = \beta_0 + \beta_1(PANTALLAS - \overline{PANTALLAS}) + \beta_2(PELICULAS - \overline{PELICULAS}) + U$$

Los parámetros β del modelo de regresión cuantifican la relación existente entre cada variable explicativa (X_i) y la variable explicada (Y). En nuestro caso, las **variables** son las siguientes:

- **Primera variable explicativa**: X_1 es PANTALLAS menos su promedio $\overline{PANTALLAS}$, medido en unidades de pantallas.
- **Segunda variable explicativa**: X_2 es PELICULAS menos su promedio $\overline{PELICULAS}$, medido en unidades de películas.
- **Variable explicada**: Y es RECAUDA, la recaudación obtenida por la sala, medida en millones de euros.

Para este caso, nos interesa restarle el promedio respectivo a cada una de las variables para evitar absurdos a la hora de definir los parámetros. Buscaremos definir estos parámetros a partir de la situación promedio en vez de la "situación cero"; es decir, la recaudación que obtendríamos tras aplicar el modelo en el caso de que el número de salas y de películas fuese cero no tendría sentido. Es por esto que, a modo de control, utilizamos el promedio de las variables para operar.

Parámetros beta

Pasamos ahora a hablar sobre los **parámetros beta del modelo**, de los que hemos obtenido su estimación numérica a partir de la siguiente tabla de *StatGraphics*:

		Error	*Estadístico*	
Parámetro	*Estimación*	*Estándar*	*T*	*Valor-P*
CONSTANTE	445,041	4,9842	89,2904	0,0000
PANTALLAS-AVG(PANTALLAS)	0,171799	0,00544677	31,5415	0,0000
PELICULAS-AVG(PELICULAS)	-0,0658274	0,0131962	-4,98837	0,0001

- **Parámetro** β_0: Cuando todas las variables explicativas valen cero, el promedio de la variable explicada RECAUDA es 445,041
- **Parámetro** β_1: Cuando se produce un incremento unitario de la variable explicativa X_1, y se mantienen constantes los valores del restos de las variables explicativas, el incremento del promedio de la variable explicada es 0,1717
- **Parámetro** β_2: Cuando se produce un incremento unitario de la variable explicativa X_2, y se mantienen constantes los valores del restos de las variables explicativas, el incremento del promedio de la variable explicada es 0,0658

Ahora buscamos determinar si existe **significación de los parámetros**, mediante el uso del p-valor y del estadístico t. Para ello, establecemos una hipótesis nula y su alternativa:

$$H_0 : \beta_i = 0 \rightarrow \text{Hipótesis nula: El parámetro no es significativo}$$
$$H_1 : \beta_i \neq 0 \rightarrow \text{Hipótesis alternativa: El parámetro sí es significativo}$$

P-valor

Comprobando el p-valor de ambos parámetros en la tabla observamos que en ambos casos es menor que 0,05 y, por tanto, rechazamos la hipótesis nula; esto significa que los parámetros sí son significativos.

T-Student (estadístico t)

Teniendo en cuenta que $\alpha = 0,05$ y que los grados de libertad residuales:

$$\text{Si } \frac{b_i - \beta_i}{s_{b_i}} \equiv t_{n-k-1} \text{ aceptamos } H_0 ; \; t_{20}^{0,025} = 2,086; \; |t_{calc}| \leq t_{gdlr}^{\alpha/2}$$

Para ambos parámetros se cumple que 4,98837 > 2,086 y 31,5415 > 2,086 por lo que rechazamos la hipótesis nula, comprobando de nuevo la significación de los parámetros.

¿Es adecuado el modelo?

Después de ver que los parámetros establecidos son significativos, **determinamos si el modelo es adecuado** mediante la tabla ANOVA del modelo, y planteando nuevas hipótesis que resolveremos en este caso con el p-valor y el estadístico F:

H_0: $\beta_i = 0 \quad Si \ \forall i \geq 1 \quad \rightarrow$ Hipótesis nula: El modelo no es adecuado

H_1: $\beta_i \neq 0 \rightarrow$ Hipótesis alternativa: El modelo sí es adecuado

Análisis de Varianza

Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrado Medio*	*Razón-F*	*Valor-P*
Modelo	886766,	2	443383,	776,00	0,0000
Residuo	11427,4	20	571,372		
Total (Corr.)	898194,	22			

R-cuadrada = 98,7277 porciento
R-cuadrado (ajustado para g.l.) = 98,6005 porciento
Error estándar del est. = 23,9034
Error absoluto medio = 17,8292
Estadístico Durbin-Watson = 1,26344 (P=0,0092)
Autocorrelación de residuos en retraso 1 = 0,334487

P-valor

Si el p-valor es menor que el error de significación $\alpha = 0,05$, entonces aceptaremos la hipótesis nula. Al fijarnos en la tabla ANOVA, vemos que el p-valor es muy pequeño en comparación al error que hemos establecido, por lo que se rechaza la hipótesis nula; esto quiere decir que tomamos como adecuado el modelo.

F-ratio (estadístico F)

Si realizamos la misma prueba comparando el estadístico F con el F-Ratio:

$$Si \ F_{Ratio} \leq F_{glM,glR}\alpha \ \rightarrow \ Aceptamos \ H_0 \ (Modelo \ inadecuado)$$

El F-Ratio es de 776, un valor mucho mayor que el estadístico F obtenido a partir de los grados de libertad del modelo ($gl_M = 2$ y $gl_R = 20$), que es de 3'49. Por ello, se rechaza la hipótesis nula y nos reiteramos en que el modelo es adecuado.

Coeficientes de determinación

Fijándonos de nuevo en la tabla ANOVA del modelo, observamos que *StatGraphics* obtiene por nosotros diversos valores que pueden sernos útiles. Por ejemplo, podemos obtener los **coeficientes de determinación del modelo**, que en este caso aparecen como *R-cuadrada*:

Análisis de Varianza

Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrado Medio*	*Razón-F*	*Valor-P*
Modelo	886766,	2	443383,	776,00	0,0000
Residuo	11427,4	20	571,372		
Total (Corr.)	898194,	22			

R-cuadrada = 98,7277 porciento
R-cuadrado (ajustado para g.l.) = 98,6005 porciento
Error estándar del est. = 23,9034
Error absoluto medio = 17,8292
Estadístico Durbin-Watson = 1,26344 (P=0,0092)
Autocorrelación de residuos en retraso 1 = 0,334487

Por una parte, el coeficiente de determinación (98'728%) nos indica que esta será la variabilidad que tendrá la recaudación de las salas de cine españolas explicada por el modelo lineal que hemos planteado. En cambio, el coeficiente de determinación corregido (98'601%) nos da una variabilidad comparable a cualquier otro modelo lineal que tenga los mismos grados de libertad y variables independientes distintas, aunque esto no tenga un significado extrapolable por sí mismo al problema que nos ocupa.

Ejemplo práctico

A continuación exponemos un ejemplo en el que ofrecemos una estimación de la recaudación esperada, si el número pantallas fuera de 3800 unidades y el número de películas fuera de 1800 títulos.

Regression Results for RECAUDA

	Fitted	Stnd. Error	Lower 95,0%	Upper 95,0%	Lower 95,0%
Row	Value	CL for Forecast	CL for Forecast	CL for Forecast	CL for Mean
24	571,293	24,6478	519,879	622,708	558,753

	Upper 95,0%
Row	CL for Mean
24	583,834

Estimación puntual: 571'293
Intervalo de confianza al 95% para la media: [558'753, 583'834]
Intervalo de confianza al 95% para la estimación: [519'879, 622'708]

Según los datos que observamos, la estimación puntual para la variable RECAUDA resulta en 571,293. Esto nos indica que si ha habido 3800 salas en las que se hayan proyectado 1800 películas, se recaudarían aproximadamente 571.293.000 euros.

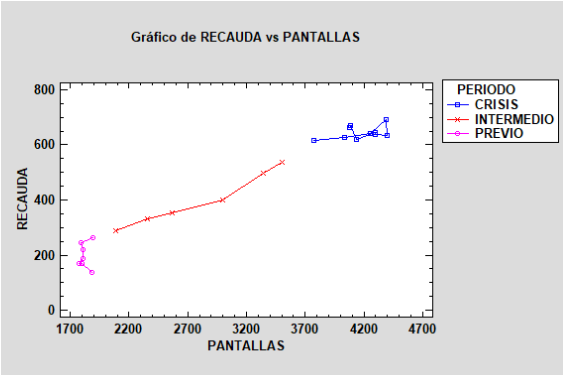
El valor de la estimación puntual se encuentra dentro del intervalo de confianza para la estimación, por lo que parece que las estimaciones de RECAUDA coinciden y son adecuadas, con un 95% de confianza.

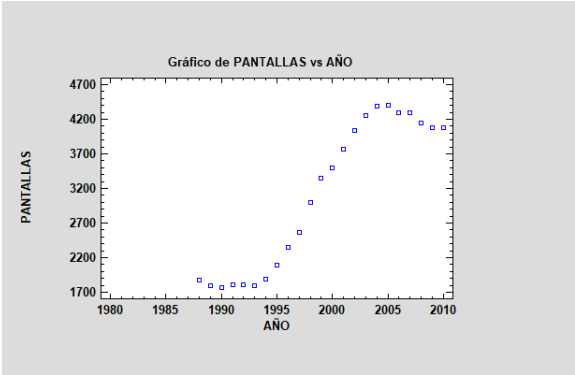
De igual forma, dicho valor también se encuentra dentro del intervalo de confianza para la media. Este intervalo indica el rango de valores posibles de la media del dinero recaudado. Por lo que si el valor estimado puntualmente se encuentra dentro de dicho intervalo, lo podríamos valorar como adecuado también con un 95% de confianza.

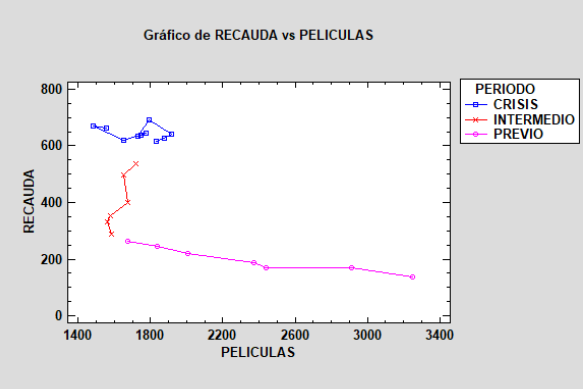
Se podría destacar que el rango del intervalo de confianza para la media es más reducido que el dicho intervalo para la estimación. Esto es lógico puesto que el intervalo para la estimación está basado en una predicción y el intervalo para la media está basado en datos del fichero, por lo que es más "fácil" que este último intervalo sea más reducido, y por tanto más preciso.

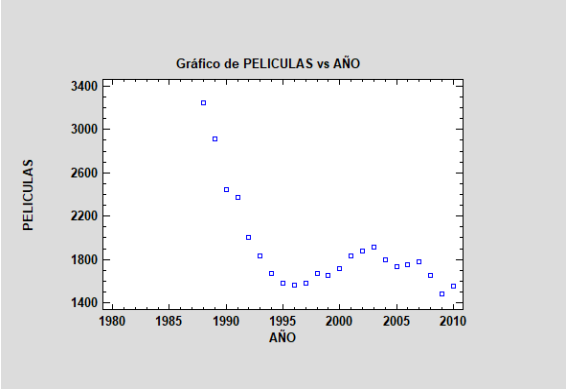
La crisis del 2001: necesidad de un nuevo modelo

Según los expertos, en el año 2001 comenzó "una caída libre acelerada por múltiples causas: la irrupción de internet, el P2P, las descargas ilegales, el cine en casa, los cambios de hábitos de los espectadores y la crisis económica". Realmente esta situación nos muestra que, aunque nuestro modelo sea adecuado, **necesitamos un nuevo modelo** que contemple esta temporalidad. Podemos justificar el cambio de modelo a partir de los dos primeros gráficos utilizados en este modelo, y la evolución del tiempo en cada variable explicativa (PANTALLAS y PELÍCULAS):









Como se puede observar, cada color representa un período y la relación que hay entre las películas que se han proyectado y el dinero que se ha recaudado durante ese período, al igual que las pantallas existentes y estas mismas recaudaciones. Por otro lado, tenemos la evolución en el tiempo del número de pantallas y del número de películas.

Período PREVIO (1988-1994)

- ➔ *Recaudaciones*: 150-250 *millones de euros*
- ➔ *Películas proyectadas*: 1700-3300 *películas*
- ➔ *Pantallas*: 1700-1900 *pantallas*

Se puede ver que el número de pantallas en España durante este período está por debajo de 2000. Tiene sentido que con un número pequeño de salas obtengamos unas recaudaciones menores en comparación a los otros períodos. En cambio, podemos observar un comportamiento bastante diferente en contraste con los otros dos períodos con respecto a la variable de películas proyectadas, puesto que con mucha menos recaudación, el número de películas proyectadas fue considerablemente grande, manteniéndose aproximadamente las recaudaciones con el paso de los años, aunque con una tendencia a la baja.

En 1990 se hizo la primera conexión a Internet en España, pero no fue hasta principios de los 2000 que Internet empezó a utilizarse en un nivel más cotidiano por la población general, por lo que entre 1988 y 1994, en España, si alguien quería ver una película tenía que ir al cine en cualquier caso. No obstante, a pesar de la gran cantidad de películas proyectadas, las recaudaciones son las más bajas de todos los períodos. Podemos dar respuestas a este hecho suponiendo que se debe a que el precio de las entradas durante este período era mucho más bajo que el de los otros períodos.

Período INTERMEDIO (1995-2000)

- ➔ *Recaudaciones*: 300-550 *millones de euros*
- ➔ *Películas proyectadas*: 1600-1700 *películas*
- ➔ *Pantallas*: 2100-3500 *pantallas*

De manera similar al período anterior, también concuerda que con un número de pantallas existentes intermedio, obtenemos unas recaudaciones acordes a dicha cantidad. Sin embargo, si comparamos la relación entre las recaudaciones y el número de pantallas parece haber otra contradicción. Se puede observar que hay un número de películas proyectadas relativamente pequeño, entre 1600 y 1700 películas. Este número de proyecciones generó entre 300 y 550 millones de euros, un número alto para las películas que se han proyectado y mayor que el del período temporal anterior.

En el año 1995 estaban operativos en España 30 proveedores de servicios de Internet y fue Telefónica la que introdujo Internet de manera residencial, si bien solo lo utilizaban grandes empresas e investigadores, no a nivel doméstico.

Este período de "transición", entre un período de relativa estabilidad económica tras la entrada de España en la Unión Europea (PREVIO) y un período de crisis económica en el sector (CRISIS), se observa cómo el número de películas proyectadas es bajo. No obstante, las recaudaciones son bastante altas. De igual manera que hemos hecho antes, podemos suponer que esto se debe a que los precios de las entradas han subido.

Período CRISIS (2001-2010)

- ➔ *Recaudaciones*: 600-700 *millones de euros*
- ➔ *Películas proyectadas*: 1500-1900 *películas*
- ➔ *Pantallas*: 3800-4400 *pantallas*

En este período también encaja que con el número más grande de salas existentes, se obtengan las mayores recaudaciones. En cuanto a películas, podemos ver que con un número "pequeño" de películas proyectadas, de nuevo por debajo de las 2000 películas, se ha obtenido el nivel de recaudaciones más alto que hay en el gráfico.

Aquí entran varios factores que pueden dar respuesta a esto, además del precio de las entradas, se considera relevante que en esta década se produjo el *boom* de Internet. A finales de los 90 y, sobre todo, principios de los 2000, Internet entró en nuestros hogares, dando lugar a las descargas ilegales y al cine en casa. Esto se sumó al cambio de hábitos de los consumidores, ya que el precio de las entradas llevaba subiendo desde 1995 aproximadamente, e iban mucho menos al cine.

Por último, cabe destacar en el gráfico del número de pantallas que, tal y como hemos comentado, el número de pantallas que aparece es plausible para las recaudaciones que se han obtenido en cada período. Parece que el número de salas es creciente en el tiempo y que no se ve afectado por la crisis; es decir, que aún habiendo una situación económica delicada, se seguían abriendo cines y/o salas de cine, aunque no al mismo ritmo inicial. Esto es inverso al comportamiento que hemos observado con las películas. Por ello, nos ha parecido curioso comentarlo.

Segundo modelo

Proposición y ajuste del modelo. Variables

En el modelo previo se ha podido observar el efecto de la crisis cinematográfica de 2001. Para introducir este hecho en el modelo, se crean las variables ficticias A2001 y A1995 para señalar el periodo de crisis y el periodo inmediatamente anterior, respectivamente. Tal y como hemos visto, hemos deducido que desde el año 2001 existió una crisis (variable A2001), además de un comportamiento anómalo desde 1995, atendiendo tanto al número de pantallas como al número de películas (variable A1995). Por todo esto, para incluir el comportamiento de este periodo debemos **plantear un nuevo modelo**:

$$RECAUDA = \beta_0 + \beta_1(PANTALLAS - \overline{PANTALLAS}) + \beta_2(PELÍCULAS - \overline{PELÍCULAS}) + \beta_3(A2001) + \beta_4(A1995) + \beta_5(A2001 * PANTALLAS - \overline{PANTALLAS}) + \beta_6(A2001 * PELÍCULAS - \overline{PELÍCULAS}) + \beta_7(A1995 * PANTALLAS - \overline{PANTALLAS}) \beta_8(A1995 * PELÍCULAS - \overline{PELÍCULAS}) + U$$

Para simplificar, podemos sacar factor común:

$$RECAUDA = \beta_0 + (PANTALLAS - \overline{PANTALLAS}) * (\beta_1 + \beta_5 * A2001 + \beta_7 1995) + (PELÍCULAS - \overline{PELÍCULAS}) * (\beta_2 + \beta_6 * A2001 + \beta_8 * 1995) + \beta_3(A2001) + \beta_4(A1995) + U$$

De todas formas, *StatGraphics* nos permite descartar aquellas variables que no resulten significativas, mediante selección paso a paso. Los parámetros tendrán igualmente el mismo significado que en el primer modelo. Así quedaría el **modelo ajustado**:

$$RECAUDA = \beta_0 + \beta_1(PANTALLAS - \overline{PANTALLAS}) + \beta_2(PELICULAS - \overline{PELICULAS}) + \beta_3(A2001) + \beta_4(A2001 * PANTALLAS - \overline{PANTALLAS}) + U$$

En la página siguiente adjuntamos la tabla con las variables independientes contempladas en el modelo tras haber aplicado el ajuste paso a paso.

Multiple Regression - RECAUDA
Dependent variable: RECAUDA
Independent variables:
A2001
A1995
PANTALLAS-AVG(PANTALLAS)
PELICULAS-AVG(PELICULAS)
A2001*PELICULAS-AVG(PELICULAS)
A2001*PANTALLAS-AVG(PANTALLAS)
A1995*PELICULAS-AVG(PELICULAS)
A1995*PANTALLAS-AVG(PANTALLAS)
Number of observations: 23

		Standard	T	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	98,79	102,4	0,9651	0,3473
A2001	475,9	136,4	3,49	0,0026
PANTALLAS-AVG(PANTALLAS)	0,1628	0,009837	16,55	0,0000
PELICULAS-AVG(PELICULAS)	-0,06722	0,01088	-6,181	0,0000
A2001*PANTALLAS-AVG(PANTALLAS)	-0,1083	0,03359	-3,223	0,0047

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	8,92E5	4	2,23E5	649,40	0,0000
Residual	6181,	18	343,4		
Total (Corr.)	8,982E5	22			

La clave que diferencia a este modelo lineal del anterior son las nuevas variables, A1995 y A2001, que toman valor 1 o 0 según si las películas son o no de dichos años. Viendo que la variable que indica las películas que corresponden al año 1995 no es significativa, debido a que no aparece en la tabla anterior, podemos establecer dos modelos distintos: un **modelo pre-2001** (que englobaría los periodos "Previo" (1988-1994) e "Intermedio" (1995-2000) que hemos observado en el primer modelo) y un **modelo post-2001** (que se corresponde con el periodo "Crisis", 2001-2010). De todas formas, nos ha parecido algo extraño que, aunque dicha variable (A1995) no sea significativa, sí que cause una anomalía observable en los gráficos.

Modelo pre-2001 (A2001=0)

$$RECAUDA_{INTER} = \beta_0 + \beta_1(PANTALLAS - \overline{PANTALLAS}) + \beta_2(PELICULAS - \overline{PELICULAS}) + U$$

Modelo post-2001 (A2001=1)

$$RECAUDA_{CRISIS} = \beta_0 + \beta_1(PANTALLAS - \overline{PANTALLAS}) + \beta_2(PELICULAS - \overline{PELICULAS}) + \beta_3(A2001) + \beta_4(A2001 * PANTALLAS - \overline{PANTALLAS}) + U$$

$$\downarrow\downarrow\downarrow$$

$$RECAUDA_{CRISIS} = \beta_0 + \beta_1(PANTALLAS - \overline{PANTALLAS}) + \beta_2(PELICULAS - \overline{PELICULAS}) + \beta_3 + \beta_4(PANTALLAS - \overline{PANTALLAS}) + U$$

$$\downarrow\downarrow\downarrow$$

$$RECAUDA_{CRISIS} = \beta_0 + [(\beta_1 + \beta_4)(PANTALLAS - \overline{PANTALLAS})] + \beta_2(PELICULAS - \overline{PELICULAS}) + \beta_3 + U$$

Cuestiones sobre el modelo: el cine post-2001

A continuación nos planteamos dos preguntas relacionadas con nuestro problema, que nos permitan confirmar que lo hemos planteado correctamente. Para contestar ambas cuestiones, utilizaremos el **modelo post-2001**, con un nivel de significación del 5%.

1. ¿Es aceptable pensar que se ha producido un cambio por la crisis del año 2001?

Para poder contestar a esta cuestión tenemos que plantear y comprobar las dos respuestas posibles. Por un parte, la primera contestación puede ser que la variable A2001 (es decir, el año 2001, inicio de la crisis) no tiene efecto sobre la recaudación, lo que significa que los parámetros beta β que acompañan a la variable A2001 son 0. Por otra parte, la otra contestación posible es que el año 2001 sí que tiene efecto sobre la recaudación, por lo que, como mínimo, habrá un parámetro beta β distinto de cero, y el año 2001 sí que tendría efecto sobre la recaudación.

Los parámetros beta β en este caso son: β_3 que acompaña a la variable A2001 (como variable individual) y β_4 que acompaña a la interacción de A2001 y la variable de las pantallas. Por ello, el planteamiento de las hipótesis será:

$$H_0\colon\ \beta_3 = \beta_4 = 0 \rightarrow \textit{Año 2001 no tiene efecto sobre la recaudación}$$
$$H_1\colon\ \textit{al menos un beta es distinto de cero} \rightarrow \textit{Año 2001 tiene efecto sobre la recaudación}$$

A partir del modelo que hemos ajustado, si sustituimos β_3 y β_4 por cero nos queda así:

$$RECAUDA_{CRISIS} = \beta_0 + \beta_1(PANTALLAS - \overline{PANTALLAS}) + \beta_2(PELICULAS - \overline{PELICULAS}) + U$$

Multiple Regression - RECAUDA
Dependent variable: RECAUDA
Independent variables:
PANTALLAS-AVG(PANTALLAS)
PELICULAS-AVG(PELICULAS)
Number of observations: 23

		Standard	T	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	445,041	4,9842	89,2904	0,0000
PANTALLAS-AVG(PANTALLAS)	0,171799	0,00544677	31,5415	0,0000
PELICULAS-AVG(PELICULAS)	-0,0658274	0,0131962	-4,98837	0,0001

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	886766,	2	443383,	776,00	0,0000
Residual	11427,4	20	571,372		
Total (Corr.)	898194,	22			

Para solucionar la cuestión, tenemos que calcular F_{calc} y F_{tabla}

$\star\ F_{calc} = \frac{(SCRr - SCRc)/s}{SCRc/(n-k-1)} = \frac{(11427 - 6181)/2}{CMRc} = \frac{2623}{343'4} = 7'638$

SCRc (sumas de cuadrados de residuos en el modelo completo) = 6181
SCRr (sumas de cuadrados de residuos en el modelo restringido) = 11427
s (número de restricciones) = 2

$gl_R\ modelo\ completo = 18$ 20 - 18 = 2 restricciones

$gl_R\ modelo\ con\ restricciones = 20$

n-k-1 (grados de libertad residuales modelos completo) = 18

$\star\ F_{tabla} = F_{s,\,n-k-1}\alpha = F_{2,18}\ (0'05) = 3'55$

Como el valor obtenido de F_{tabla} es menor que F_{calc} ($F_{tabla} = 3'55 < 7'638 = F_{calc}$) rechazamos la hipótesis nula. De esto concluimos que **el año 2001 sí que tiene efecto sobre la recaudación.**

Cabe mencionar que el modelo que nos ha resultado de sustituir β_3 y β_4 por cero, es el "mismo" que el modelo pre-2001. No obstante, es por **motivos diferentes**. En el modelo pre-2001, es así puesto que pertenece al período anterior a 2001, por ello A2001=0. Mientras que en el modelo con dichas betas sustituidas por cero (que hacen que la variable A2001 desaparezca, como si fuera A2001=0) es porque estábamos comprobando si el año 2001 tenía efecto sobre la recaudación.

---

2. ¿Es aceptable pensar que la recaudación no depende del número de salas a partir de 2001?

De igual forma que hemos hecho para la cuestión anterior, se deben plantear nuevas hipótesis. Por un lado, tenemos la posibilidad de que la recaudación realmente no dependa del número de salas a partir de 2001, por lo que esto significaría que los parámetros beta que acompañan a la variable de pantallas suman cero. De forma alternativa, si comprobamos que la recaudación sí que depende del número de salas a partir de 2001, esto implicaría que los parámetros beta β que acompañan a la variable de pantallas no suman cero.

Los parámetros beta β de este caso son: β_1 y β_4, ambos acompañan a la variable de pantallas. Por tanto, el planteamiento de las hipótesis será el siguiente:

$$H_0: \beta_1 + \beta_4 = 0 \rightarrow \textit{la recaudación no depende del número de pantallas a partir de } 2001$$
$$H_1: \beta_1 + \beta_4 \neq 0 \rightarrow \textit{la recaudación depende del número de pantallas a partir de } 2001$$

Entonces, a partir del modelo ajustado, sustituimos $\beta_1 + \beta_4$, lo que sumaría 0 y la variable de las pantallas no aparecerá:

$$RECAUDA_{CRISIS} = \beta_0 + [(\beta_1 + \beta_4)(PANTALLAS - \overline{PANTALLAS})] + \beta_2(PELICULAS - \overline{PELICULAS}) + \beta_3 + U$$

↓↓↓

$$RECAUDA_{CRISIS} = \beta_0 + \beta_2(PELICULAS - \overline{PELICULAS}) + \beta_3 + U$$

Regresión Múltiple - RECAUDA
Variable dependiente: RECAUDA
Variables independientes:
PELICULAS-AVG(PELICULAS)
A2001=1
Número de observaciones: 23

		Error	Estadístico	
Parámetro	*Estimación*	*Estándar*	*T*	*Valor-P*
CONSTANTE	311,334	20,2507	15,374	0,0000
PELICULAS-AVG(PELICULAS)	-0,157214	0,0361605	-4,34767	0,0003
A2001=1	307,526	31,6656	9,71167	0,0000

Análisis de Varianza

Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrado Medio*	*Razón-F*	*Valor-P*
Modelo	796744,	2	398372,	78,54	0,0000
Residuo	101449,	20	5072,46		
Total (Corr.)	898194,	22			

Para solucionar la cuestión, tenemos que calcular F_{calc} y F_{tabla}

$\star\ F_{calc} = \frac{(SCRr - SCRc)/s}{SCRc/(n-k-1)} = \frac{(101449 - 6181)/2}{CMRc} = \frac{47634}{343'4} = 138'753$

SCRc (sumas de cuadrados de residuos en el modelo completo) = 6181
SCRr (sumas de cuadrados de residuos en el modelo restringido) = 101449
s (número de restricciones) = 2

$gl_R\ modelo\ completo = 18$ 20 - 18 = 2 restricciones

$gl_R\ modelo\ con\ restricciones = 20$

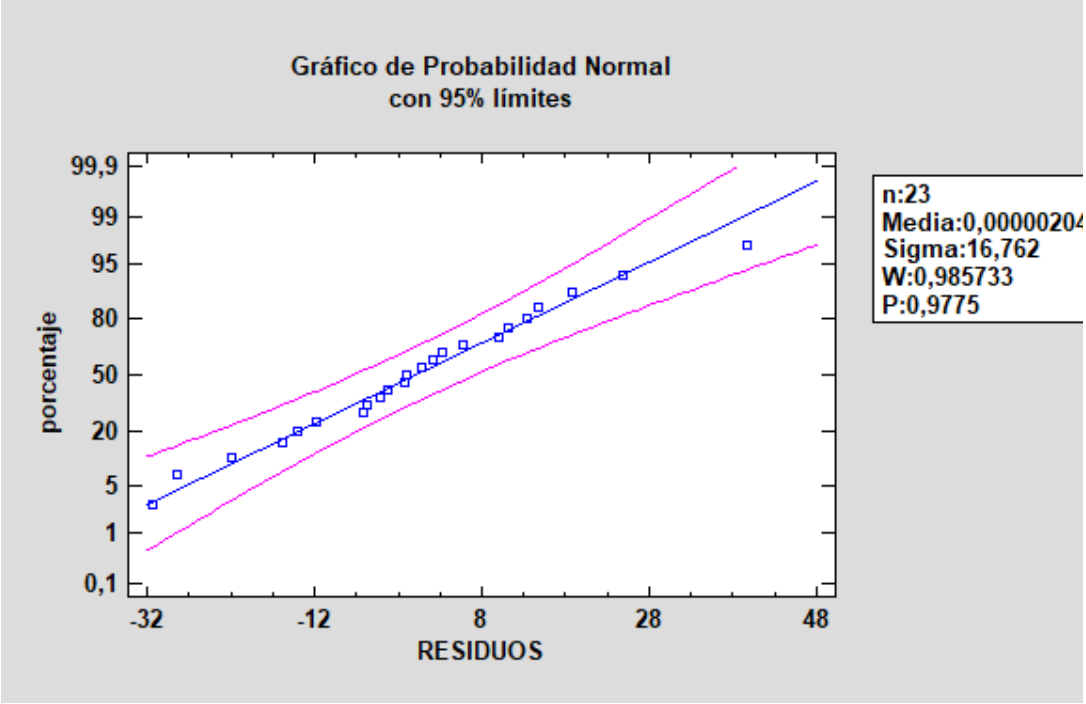
n-k-1 (grados de libertad residuales modelos completo) = 18

$\star\ F_{tabla} = F_{s,n-k-1}\alpha = F_{2,18}\ (0'05) = 3'55$

Como el valor obtenido de F_{tabla} es menor que F_{calc} ($F_{tabla} = 3'55 < 138'753 = F_{calc}$) rechazamos la hipótesis nula. De esto concluimos que **la recaudación sí depende del número de pantallas** a partir de la crisis del 2001.

Distribución del error del modelo

Mediante *StatGraphics*, hemos realizado un gráfico del **papel probabilístico normal** del modelo. En el gráfico de debajo vemos que los puntos de dispersión están alineados y dentro de los límites de error establecidos; esto nos permite afirmar que el error de nuestro modelo tiene una **distribución normal**.

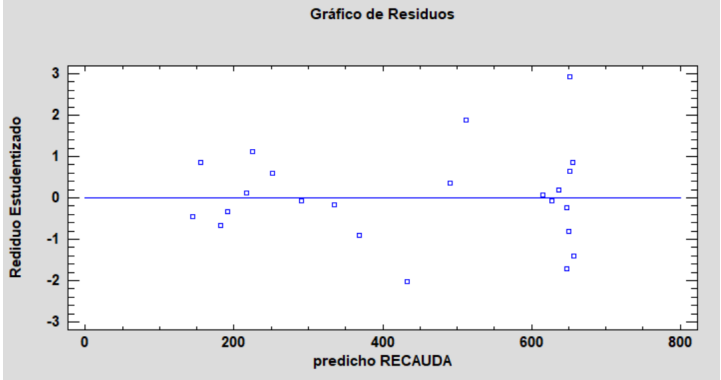


Residuos. Heterocedasticidad y autocorrelación

Para terminar con el análisis de este segundo modelo, utilizaremos los residuos del modelo para poder confirmar que este modelo se ajusta a lo que nosotros necesitamos analizar. En primer lugar, analizaremos y comentaremos los **gráficos de residuos**.

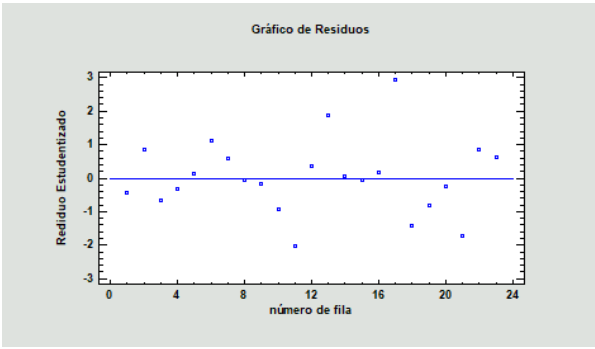
1. Residuos frente a estimación de la variable explicada de la recaudación (RECAUDA)

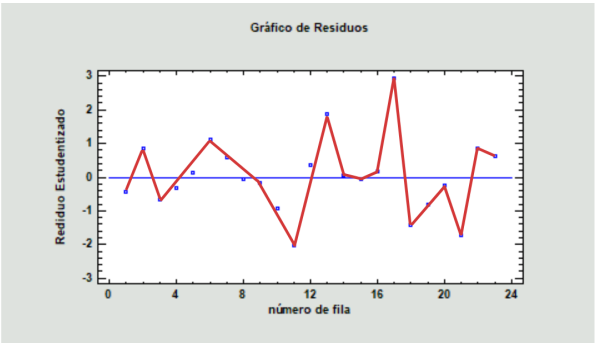
Este gráfico es algo confuso, y no nos permite observar si existiría heterocedasticidad en nuestro modelo. Pensamos que esto puede deberse a que el número de residuos es pequeño e intuimos que si fuese mayor se podría apreciar una nube de puntos.



---

2. Residuos frente al orden de las variables

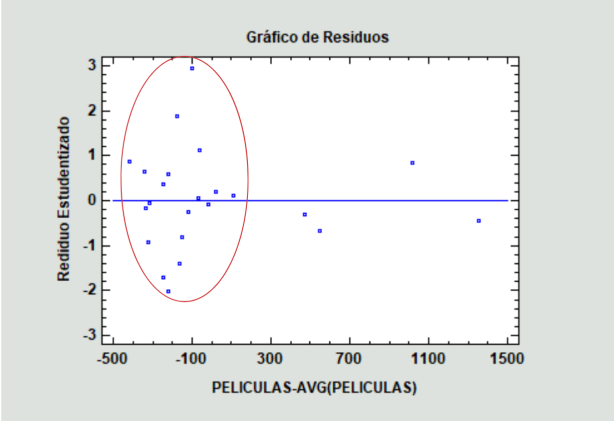




En el gráfico se observa que no hay una nube, por lo que parece que los residuos tienen un orden. Sin embargo, al igual que hemos dicho anteriormente, el número de residuos es pequeño por lo que comprobaremos más adelante si existe o no heterocedasticidad y autocorrelación.

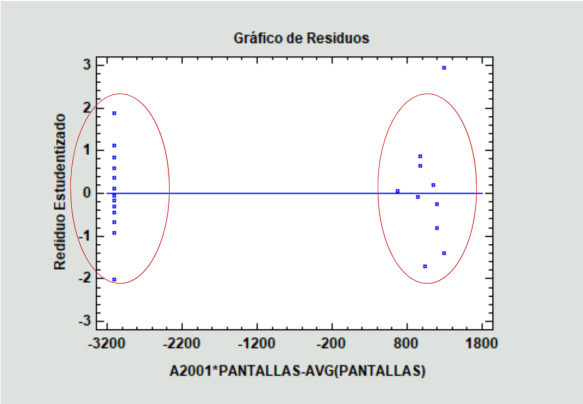
3. Residuos frente a la variables explicativas

En todos los siguientes gráficos de dispersión de las variables explicativas, no podemos ver con claridad que exista heterocedasticidad en los residuos de las varianzas, porque dichas varianzas no son de fuentes distintas sino que todas proceden de la misma fuente: las salas de cine españolas.

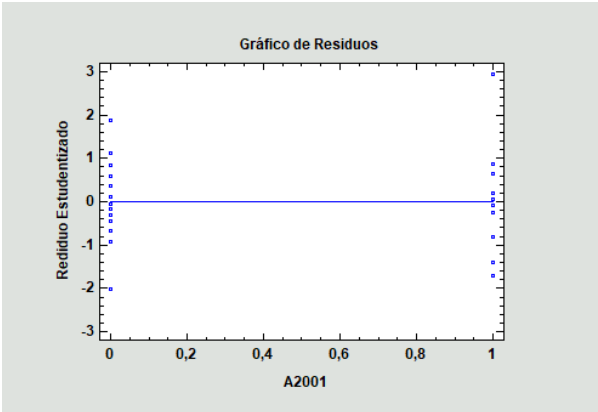


Para el gráfico de residuos de la variable películas podemos observar una nube marcada en rojo, lo que nos indicaría que hay una variabilidad parecida.

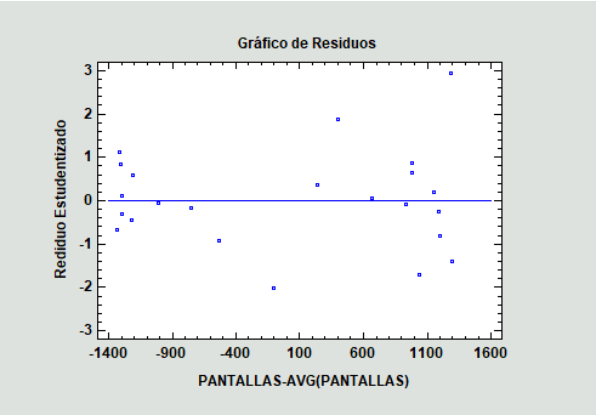
Hay algunos residuos fuera de dicha nube, pero, como conclusión general, podríamos decir que no se observa heterocedasticidad en este gráfico.



Cabe destacar el caso del gráfico de la interacción entre las variables del año 2001 y la de pantallas, donde parece observarse que hay dos poblaciones, por lo que el residuo parece depender de alguna variable explicativa. Sin embargo, al ser un gráfico de una interacción, esas dos poblaciones aparentes pueden deberse a dicha interacción.



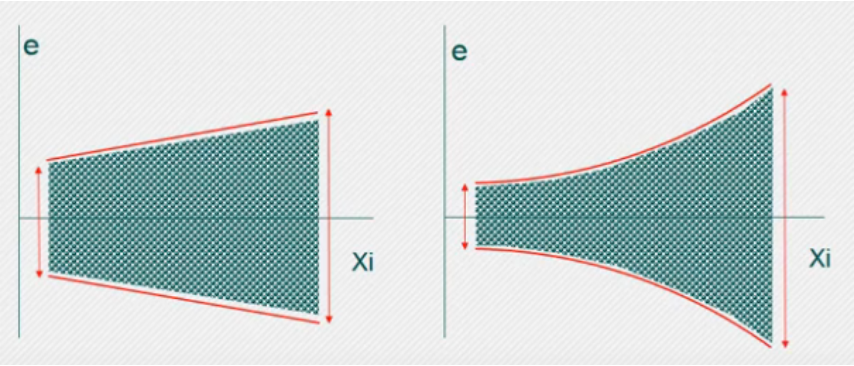
El único gráfico en el que se observa un comportamiento distinto es en el de la variable A2001, debido a que los valores solamente pueden ser 0 o 1. No obstante, tanto los valores en 0 como en 1 de A2001 se encuentran la mayoría en torno a la línea azul, en torno a cero, por lo que apunta a que no hay heterocedasticidad.



En este gráfico vemos un comportamiento parecido a los dos primeros gráficos (de la variable explicada y del orden), donde no encontramos una nube de puntos, pero igualmente deducimos que se debe a la baja cantidad de residuos

---

En un gráfico de residuos donde se puede observar que hay heterocedasticidad tendría un aspecto parecido a la siguiente imagen. Donde se aprecia que hay diferencia de varianzas, que hay heterocedasticidad. Aunque pudiera parecerlo en este último gráfico (el de la variable pantallas), esta forma de dispersión no se observa en ninguno de los gráficos anteriores.



En cualquier caso, aunque en los gráficos pudiera apreciarse heterocedasticidad o autocorrelación, a continuación intentaremos confirmar o desmentir su existencia numéricamente.

Más allá de esta interpretación gráfica, ahora buscamos demostrar mediante estadística la existencia o no de **heterocedasticidad**, utilizando de nuevo el planteamiento de hipótesis:

H_0: $\sigma^2 = cte \Rightarrow \gamma_i = 0 \; \forall \, i \geq 1$ — NO existe heterocedasticidad

H_1: $\sigma^2 \neq cte \Rightarrow$ *al menos uno es distinto de* 0 — Existe heterocedasticidad

Regresión Múltiple - RESIDUOS^2

Variable dependiente: RESIDUOS^2

Variables independientes:

PANTALLAS-AVG(PANTALLAS)
PELICULAS-AVG(PELICULAS)
A2001
A2001*PANTALLAS-AVG(PANTALLAS)

Número de observaciones: 23

		Error	Estadístico	
Parámetro	*Estimación*	*Estándar*	*T*	*Valor-P*
CONSTANTE	3490,57	2066,85	1,68884	0,1085
PANTALLAS-AVG(PANTALLAS)	0,244723	0,198633	1,23204	0,2338
PELICULAS-AVG(PELICULAS)	-0,00690723	0,219599	-0,0314538	0,9753
A2001	-4473,99	2753,29	-1,62496	0,1216
A2001*PANTALLAS-AVG(PANTALLAS)	0,992027	0,678286	1,46255	0,1608

Análisis de Varianza

Fuente	*Suma de Cuadrados*	*GI*	*Cuadrado Medio*	*Razón-F*	*Valor-P*
Modelo	904994,	4	226248,	1,62	0,2135
Residuo	2,52005E6	18	140003,		
Total (Corr.)	3,42504E6	22			

En este caso, utilizaremos el estadístico del p-valor. Gracias a la tabla obtenida mediante *StatGraphics*, podemos comprobar que en este análisis el p-valor del modelo es igual a 0,2135 > 0,05; por tanto, aceptamos la hipótesis nula y confirmamos que **no existe heterocedasticidad**.

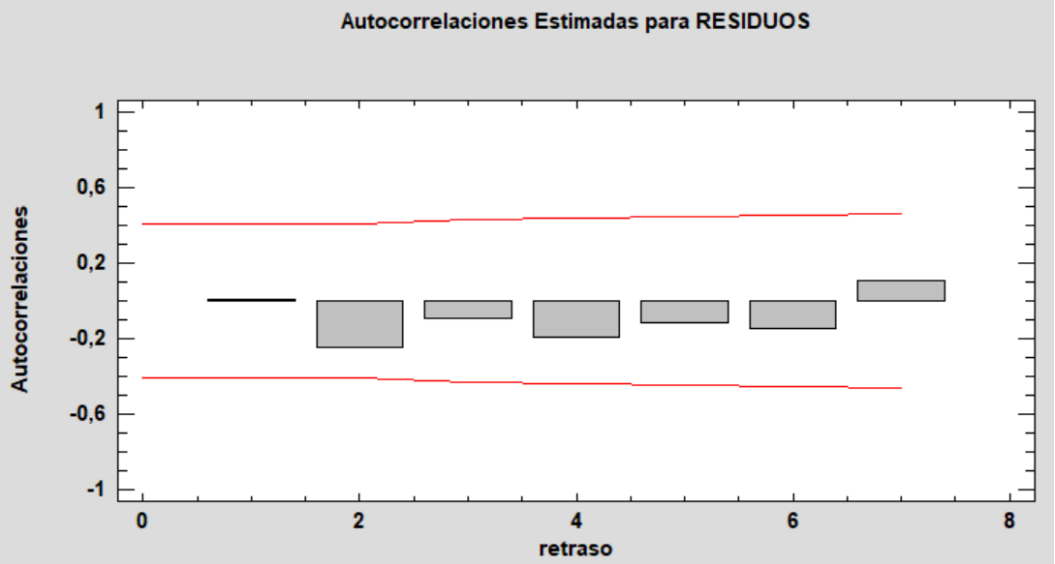
En tal caso, la varianza media prevista corresponde con el cuadrado medio residual que en el caso de nuestro modelo sería 343,4.

Por último, analizaremos si existe **autocorrelación** en el modelo, tanto simple (FAS) como parcial (FAP):

FAS: Se realiza la prueba para los coeficientes de autocorrelación simple (ρ_i) a partir de su estimación en la muestra (r_i) y de la comparación con un valor $r_{límite}$, que delimita la zona de aceptación de la prueba.

H_0: $\rho_i = 0$ → No hay autocorrelación

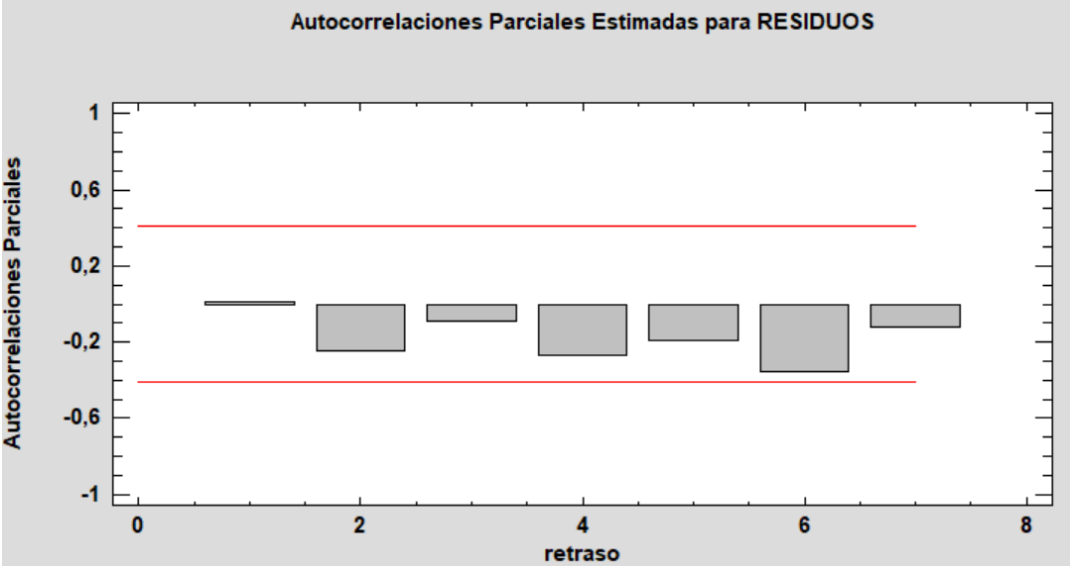
H_1: $\rho_i \neq 0$ → Hay autocorrelación Se acepta H_0 si $|r_i| < r_{límite}$



FAP: Se realiza la prueba para los coeficientes de autocorrelación parcial (α_i) a partir de su estimación en la muestra (a_i) y de la comparación con un valor $a_{límite}$, que delimita la zona de aceptación de la prueba.

H_0: $\alpha_i = 0$ → No hay autocorrelación
H_1: $\alpha_i \neq 0$ → Hay autocorrelación

Se acepta H_0 si $|a_i| < a_{límite}$



Como ninguno de los coeficientes de autocorrelación, tanto los simples como los parciales, sobrepasan los límites de la prueba de hipótesis, aceptamos la hipótesis nula y podemos decir que **no hay problemas de autocorrelación**.

Conclusiones

Tras haber realizado todos los análisis arriba descritos, y ajustar el modelo a las circunstancias temporales, podemos concluir que nuestro modelo es adecuado. Viendo que no existe heterocedasticidad ni autocorrelación (es decir, que los errores, los residuos observados son constantes y no dependen de su cercanía temporal a otros valores), confirmamos que el cambio observado en el año 2001, más allá de comprobar que ocurrió en la realidad, tiene sentido estadístico.

Hemos podido ver cómo las salas de cine en nuestro país fueron multiplicándose a finales del siglo XX, mientras que el número de películas iba reduciéndose con el paso de los años. Esta tendencia cambia con la entrada en el siglo XXI, con el desarrollo de Internet y su llegada a los hogares españoles, y de hecho perdura hasta el día de hoy: se nos hace difícil recordar la inauguración de un cine nuevo en los últimos años. Más allá de que la pandemia haya acelerado el declive del cine tradicional frente al cine en casa, este problema ya venía de principios de siglo. Tanto la producción de películas como el número de salas de proyección se estancaron a partir del año 2001, pero **la recaudación no se ha visto mermada tras esta crisis, ya que las salas de cine encontraron una solución: subir el precio de las entradas**.

Esta evolución en el precio es claramente visible: en 1990, una entrada de cine costaba unas 500 pesetas (aproximadamente 3 euros); en el año 2000, ya costaba de media 4 euros, y a finales de la década este precio ascendía a los 7 euros. Por tanto, el precio de las entradas de cine se ha multiplicado por más del doble en dos décadas.