Caso Hollywood Rules

Julián Mateo Arcos Moreno

Santiago Velásquez Villamil

Felipe Ángel Mogollón

Profesor:

Juan Nicolás Velásquez Rey

Pontificia Universidad Javeriana

Analítica de datos

Bogotá2025

“Hollywood Rules”

Introducción:

En el siguiente caso “Hollywood Rules: The View from Wall Street", se va a hacer una exploración estadística sobre el sector de películas y estudios de producción de Hollywood desde un punto de vista financiero. Por medio de pruebas de hipótesis y regresiones lineales se determinará la viabilidad de proyectos de cine como posible inversión de cobertura.

1. **Descripción general de los datos:**

   Para poder entender el comportamiento de la recaudación en taquilla de diferentes películas de Hollywood, se hizo un análisis estadístico descriptivo de las siguientes métricas: Mínimo, Cuartil 1, Mediana, Media, Cuartil 3 y Máximo. Estas en relación con la recaudación de las películas en su fecha de estreno, su recaudación total en US e internacionalmente y también el número de teatros donde fueron estrenadas. Los resultados quedaron consignados en la tabla 1, a continuación.

STAT	Opening.Gross	Total.U.S..Gross	Total.Non.U.S..Gross	Opening.Theatres
Min	4.120.497	13.090.630	0	852
1st Quartile	10.014.865	33.880.974	15.433.097	2.490
Median	14.503.650	52.330.111	42.950.069	2.880
Mean	17.468.466	59.620.651	59.560.983	2.766
3rd Quartile	21.569.368	74.345.586	75.985.298	3.209
Max	68.033.544	198.000.317	456.235.122	3.964

   Películas Por Categoría: A través de contadores realizados en el lenguaje de programación R, encontramos en la muestra de 75 películas que las categorías de comedia y películas de calificación R tienen las siguientes cantidades:

   Comedias: 23 Películas

   Calificación R: 15 Películas

2. **Justificación a la declaración de Michael London sobre el retorno de 12% por año.**

   Michael London hizo una declaración en The Hollywood Reporter en la cual asegura que en el negocio de los estudios históricamente se han generado retornos de por lo menos 12%. Para justificar la declaración, decidimos hacer un análisis de la muestra.

a. El primer paso para hacer el análisis fue calcular el ROI de las películas, por medio del lenguaje de programación R se hizo dicho calculo para cada una de las películas, y fue añadido a la base de datos global del caso.

b. El siguiente paso fue construir un intervalo de confianza para la media del US ROI de las películas utilizando los datos calculados anteriormente, esto a un nivel de confianza del 95%. Para ello, se utilizó la función t.test, en lenguaje de programación R. Los resultados del intervalo fueron los siguientes.

t.test U.S ROI	
0,1348149	0,4510486
Inferior Limit	Superior Limit
95%	

c. Finalmente, se planteó la siguiente prueba de hipótesis para determinar si la declaración de Michael London sobre el retorno de los estudios se puede soportar estadísticamente.

i. **H0** = $\mu \leq 0.12$

ii. **H1** = $\mu > 0.12$

La prueba de hipótesis se hizo a través de la función t.test de R, a un nivel de confianza del 95%. Los resultados fueron los siguientes:

t.test U.S ROI	
t	2,1792
df	74
p-value	0,01625
mean	0,2929317

El resultado de la prueba arrojo diferentes datos que son de mucha utilidad para la justificación de las declaraciones de Michael London. La primera es que el P-Value es de 0,01625, un valor menor al nivel de significancia (α) de 0,05. Esto nos indica que hay evidencia suficiente a un nivel de confianza del 95%, para rechazar la hipótesis nula y aceptar la hipótesis alternativa. Además, podemos evidenciar como la media de la muestra es de 0,2929317, claramente superior a el 0,12 del que habla Michael London. Esto significa que la estadística demuestra a un nivel de confianza del 95%, que el ROI promedio de las películas analizadas es en efecto, superior a 12%.

3. **Análisis estadístico sobre los géneros de películas.**

Griffith sospecha que algunos géneros de películas tienen mayor probabilidad de ser exitosos que otros. Para comprobar si puede usar eso a su favor, realizamos el siguiente análisis estadístico:

a. Primero, se va a comparar el recaudo en taquilla bruto en US del genero de comedia frente a otros géneros, esto para determinar si hay una diferencia significative entre el recaudo bruto de las comedias frente a los demás géneros. Para esto se planteó la siguiente prueba de hipótesis.

Recaudación Bruta:

i. H0 = μComedias = μDemas Generos
ii. H1 = μComedias ≠ μDemas Generos

Por medio de R y utilizando la función t.test, se hizo la prueba de dos muestras para determinar si hay una diferencia significative entre las medias del recaudo bruto en películas de genero de comedia frente a otros géneros. A un nivel de confianza de 95%, los resultados fueron los siguientes:

t.test Gross	
t	-1,3728
df	47,176
p-value	0,1763
mean Comedy	55.585.721
mean Other	68.743.100

En este caso, la prueba estadística arrojo como respuesta unos datos que nos conducen a no rechazar la hipótesis nula. Primero, el resultado del P-Value es de 0,1763, claramente superior a el nivel de significancia de (α) de 0,05. Por esto podemos concluir que a un nivel de confianza del 95%, no se puede rechazar la hipótesis nula ya que no hay evidencia estadística que soporte que hay una diferencia significative entre la media de recaudo bruto de películas de comedia y otro tipo de géneros.

b. A pesar de no encontrar un resultado que soporten las sospechas de Griffith, se va a hacer el mismo análisis pero desde la variable de ROI. Entonces por medio de

una prueba de hipótesis se va a determinar si hay una diferencia estadísticamente significative entre la media de ROI de películas de comedia y otros géneros. La prueba que se plantea es la siguiente:

ROI:

i. H0 = μComedias = μDemas Generos
ii. H1 = μComedias ≠ μDemas Generos

Por medio de R y utilizando la función t.test, se hizo la prueba de dos muestras para determinar si hay una diferencia significative entre las medias del ROI en películas de genero de comedia frente a otros géneros. A un nivel de confianza de 95%, los resultados fueron los siguientes:

t.test ROI GENRE	
t	-2,0471
df	38.965
p-value	0,04743
mean Comedy	0,5401722
mean Other	0,1835754

En este caso los datos obtenidos en la prueba de hipótesis son suficientes para rechazar la hipótesis nula. Para empezar, el P-Value es de 0,04743, inferior que el nivel de significancia de (α) de 0,05. Por esta razón podemos concluir con un nivel de confianza del 95% que se rechaza la hipótesis nula ya que hay evidencia estadística que soporta que existe una diferencia significativa entre las medias de ROI de películas de genero de comedia y otros géneros.

Tras realizar las pruebas podemos concluir lo siguiente: A pesar de que las comedias no recaudan significativamente más dinero en términos de taquilla bruta, son una inversión significativamente más rentable y eficiente, más allá del resultado de la prueba de hipótesis sobre el ROI, es más que evidente la diferencia abismal entre las medias de ROI de películas de comedia y otro tipo de géneros.

4. Impacto de la Clasificación MPAA en la Recaudación

Para este punto, se realizó una prueba t para comparar si existe una diferencia significativa en la recaudación total en Estados Unidos entre las películas con clasificación R y las que no la tienen.

Dado que el valor p es de 0.3979 (mayor que el nivel de significancia común de 0.05), no hay evidencia estadística suficiente para concluir que la clasificación R tenga un impacto significativo en la recaudación total en Estados Unidos. Aunque las películas no R tuvieron una recaudación promedio ligeramente superior, la diferencia no es lo suficientemente grande como para considerarse estadísticamente relevante.

t	df	p-value
0.85686	31.986	0.3979

5. Recaudación Total en EE. UU.

a. Modelo de Regresión Completo

Se construyó un modelo inicial para predecir la recaudación total en EE. UU. utilizando variables de preproducción: presupuesto, si era una comedia, clasificación R, si era una secuela y si estaba basada en una historia conocida.

Los resultados de este modelo indican que solo el presupuesto y si la película era una secuela tienen una relación estadísticamente significativa con la recaudación. Las variables Is_Comedy, MPAA_D y Known.Story no mostraron ser predictores significativos, ya que sus valores p son mayores al nivel de significancia de 0.05.

Min	1Q	Median	3Q	Max
-74206035	-19352591	-4197528	10124286	102024488

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)	Signif
(Intercept)	1,22E+10	1,05E+10	1.162	0.2492	
Budget	8,97E+02	1,65E+02	5.424	8.12e-07	***
Is_Comedy	1,48E+10	8,92E+09	1.655	0.1026	
MPAA_D	-4,16E+09	1,03E+10	-0.403	0.6881	
Sequel	2,92E+10	1,28E+10	2.283	0.0255	*
Known.Story	-9,98E+09	8,25E+09	-1.210	0.2304	

b. Modelo de Regresión Final

Basado en los hallazgos del modelo completo, se ajustó un modelo más simple y eficiente, incluyendo únicamente las variables significativas: presupuesto y secuela.

Este modelo final demuestra que:

- Por cada dólar adicional en el presupuesto, la recaudación total en EE. UU. aumenta en aproximadamente $0.87. Este efecto es altamente significativo.
- Ser una secuela se asocia con un aumento en la recaudación de aproximadamente $30.5 millones. Este efecto también es significativo.

El R-cuadrado ajustado del modelo final es de 0.2822. Esto significa que, en conjunto, el presupuesto y ser una secuela explican cerca del 28.22% de la variabilidad en la recaudación total en EE.UU.

Min	1Q	Median	3Q	Max
-73227072	-20666561	-4888979	13323011	102210359

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)	Signif
(Intercept)	1,35E+10	9,35E+09	1.438	0.1547	
Budget	8,71E+02	1,68E+02	5.184	1.91e-06	***
Sequel	3,05E+10	1,28E+10	2.384	0.0198	*

6. Recaudación de Apertura

a. Modelo Completo

El modelo inicial de regresión lineal fue creado para evaluar la capacidad predictiva de un conjunto amplio de variables sobre la recaudación de apertura. Los resultados indican que el presupuesto, si la película es una secuela y el número de cines de apertura tienen un impacto significativo. Por el contrario, variables como el género, la clasificación MPAA, o si el estreno coincidió con un día festivo, no demostraron una influencia estadística relevante.

Min	1Q	Median	3Q	Max
-16199353	-5387327	-1135383	3684677	27451341

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)	Signif

(Intercept)	-7,74E+09	4,46E+09	-1.736	0.08723	.
Budget	1,35E+02	4,32E+01	3.133	0.00260	**
Is_Comedy	1,13E+09	2,29E+09	0.495	0.62212	
MPAA_D	6,19E+08	2,60E+09	0.238	0.81299	
Sequel	9,30E+09	3,29E+09	2.828	0.00623	**
Known.Story	-2,61E+09	2,04E+09	-1.276	0.20667	
Summer	-4,13E+09	2,24E+09	-1.842	0.07002	.
Holiday	1,47E+08	3,56E+09	0.041	0.96710	
Christmas	-3,85E+09	3,49E+09	-1.102	0.27457	
Opening.Theatres	7,12E+06	1,56E+06	4.563	2.29e-05	***

b. Modelo Final

A partir de los hallazgos del modelo completo, y considerando un nivel de significancia del 10%, se construyó un modelo más conciso y eficiente. Este modelo final incluye las variables que demostraron ser predictores significativos: presupuesto, secuela, número de cines de apertura y si el estreno fue en verano. Este modelo refinado tiene un alto poder predictivo, explicando cerca del 46.86% de la variación en la recaudación de apertura de las películas.

Los coeficientes del modelo se interpretan de la siguiente manera:

- Por cada dólar adicional en el presupuesto, la recaudación de apertura aumenta en aproximadamente $0.1211.
- Las secuelas tienen una recaudación de apertura estimada de $9.37 millones más que las películas originales.
- Las películas que se estrenan en verano, en promedio, tienen una recaudación de apertura de $3.3 millones menos que las que no lo hacen.
- Cada cine adicional en el que una película se estrena se asocia con un incremento de $7,706 en la recaudación de apertura.

Min	1Q	Median	3Q	Max
-16513449	-4504585	-1519904	3204321	28232891

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)	Signif

(Intercept)	-9,81E+09	4,03E+09	-2.433	0.01751	*
Budget	1,21E+02	4,15E+01	2.916	0.00476	**
Sequel	9,37E+09	3,18E+09	2.944	0.00439	**
Summer	-3,30E+09	2,04E+09	-1.617	0.11032	
Opening.Theatres	7,71E+06	1,45E+06	5.332	1.13e-06	***

c. Conclusión del Modelo

Las cuatro variables incluidas en el modelo final son predictores robustos y significativos de la recaudación de apertura. El número de cines en los que se estrena una película se destaca como el factor más influyente para el éxito inicial en taquilla.

d. Impacto de los Cines de Apertura

Para ilustrar el efecto del número de cines, se calculó el impacto de aumentar los cines de apertura en 100. La estimación puntual para este cambio es de $770,612. Además, el intervalo de confianza del 95% para este efecto se encuentra entre $482,371.1 y $1,058,854. Esto confirma que, con alta confianza, un mayor número de cines se traduce directamente en una mayor recaudación de apertura.

7. **Relación entre la taquilla del primer fin de semana y la taquilla total en Estados Unidos.**
    a. Al realizar la regresión se obtuvieron los siguientes valores:

Min	1Q	Mean	3Q	Max
-39917880	-11784704	-4570762	6095607	75631670

Variable	Estimación	Error estándar	t value	Pr(>|t|)
Intercepto	5108000	4503000	1.134	0.26
Opening.Gross	3.121	0.218	14.31	<2e-16

$$Total\ US\ Gross = 5.108 \times 10^6 + 3.121 \times Opening\ Gross$$

La pendiente observada de 3.121 presenta un error estándar de 0.218 y p-value de $< 2 \times 10^{-16}$, lo que lleva a confirmar que hay una relación fuertemente significativa entre los datos. A esto hay que añadirle el hecho de que el R^2

obtenido fue de 0.737, de modo que cerca del 74% de la variación de la taquilla total estáR^2

b. Si fuese cierto el *age-old Hollywood Wisdom* de que el 25% de la taquilla total proviniese del fin de semana de estreno, el modelo seria el siguiente:

$$Total\ US\ Gross\ =\ \beta_0 + \beta_1 \times Opening\ Gross$$

Esta creencia equivale a que la taquilla total es igual a cuatro veces la del fin de semana de estreno, por lo que la pendiente (β_1) debe ser igual a 4. En este orden de ideas, se plantean las siguientes hipótesis:

- H_0: $\beta_1 = 4$
- H_1: $\beta_1 \neq 4$

c. Ya con las hipótesis establecidas, se realiza el modelo con el fin de comprobar si se acepta o no la hipótesis nula. Los resultados obtenidos fueron:

- T value = -4.03
- p-value = 0.00013

Con esto se pude decir que, con 73 grados de libertad, y un p-value obtenido de 0.00013, a cualquiera de los tres niveles de significancia tradicionales (10%, 5% y 1%) se rechaza H_0; por lo tanto, no puede sostenerse que, en promedio, el 25% de la taquilla se recaude en el primer fin de semana de la película.

d. El contraste presentado supone que la relación entre las variables es lineal y que los residuos tienen varianza constante y distribución que se acerca a una normal. La prueba muestra la existencia de valores atípicos lo que puede llevar a que el p-valué sea inexacto, donde, la muestra se restringe únicamente a películas con presupuestos entre los 20 y 100 millones de dólares, limitado así la posibilidad de generalizar el resultado a toda la industria, asimismo, pueden existir diferencias en los comportamientos según género, secuelas u otros factores que no se capturan en el modelo lineal.

e. Un modelo de sound aunque sí podría considerar variables adicionales como presupuesto, secuela o temporadas, dado que la variable opening gross explica el 74% de la variación como previamente se mencionó, se puede considerar que añadir más predictores solo aportaría mejoras marginales, por lo que el modelo lineal podría considerarse adecuado para validar la creencia popular. Sin embargo, se realiza a continuación para el desarrollo del caso.

f. Al desarrollar la nueva la nueva regresión sound se obtuvieron los siguientes resultados:

Métrica	Valor estimado
Proporción taquilla primer fin de semana	0.32

De igual manera a lo sucedido con la regresión simple el p-value y pendiente fueron prácticamente los mismos, 0.000134 y 3.121 respectivamente, por lo que se sigue refutando que el 25% de la taquilla se recauda el primer fin de semana.

g. Como ya se mencionó, el coeficiente R^2 arrojo un valor de 0.737 representando que al rededor del 74% de la variación total en la taquilla total de Estados Unidos es explicada por la variación en la taquilla del fin de semana de estreno.

8. Influencia de la taquilla de apertura y de las críticas en la recaudación total en Estados Unidos.

a. y b. Con modelo de regresión realizado para este punto se obtuvieron los siguientes resultados:

Variable	Estimación	Error estándar	t value	Pr(>|t|)
Intercepto	-25100000	7485000	-3.353	0.00128
Opening.Gross	3.025	0.1926	15.705	<2e-16
Critics.Opinion	630100	132600	4.753	9.96e-06

Métrica	Valor
R-cuadrado	0.8
R-cuadrado ajustado	0.7944

Este modelo predice la taquilla total en Estados Unidos usando la información disponible luego del primer fin de semana, mostrando que las variables signficativas son la previmente analizada *Opening Gross* y la puntuación de los críticos (*Critics' Opinion*). La ecuación estimada es la siguiente:

$$Total\ US\ Gross = -25.1\ millones + 3.025 \times Opening\ Gross + 630100 \times Critics'\ Opinion$$

Tanto la recaudación de apertura como la opinión de los críticos presentaron una considerada alta significancia en el modelo (p<0.001), donde el R^2 en presentado fue del 0.8 y 0.7944 cuando ajustado, lo que demuestra que 80% del modelo es explicado por estas variables.

c.

Predicción (fit)	Límite inferior 95%	Límite superior 95%
55672091	18103052	93241130

Para la película *Flags of Our Fathers*, el modelo predice una taquilla total de aproximadamente 55.7 millones de dólares, con un intervalo de predicción

al 95% entre 18.1 millones y 93.2 millones, reflejando alta incertidumbre en la estimación individual de la película.

d.

Métrica	Valor estimado
Impacto en taquilla de +10 pts de críticos	6301000

El valor estimado o coeficiente de la opinión de los críticos de 630100, indica que un aumento en 10 puntos en la puntuación de estos se asocia en promedio con un incremento de aproximadamente:

$$10 \times 630100 \approx 6.3\ millones\ de\ dolares$$

en taquilla total. Reflejando que esta sería la ganancia esperada en recaudación que podría justificar cuanto estaría dispuesto a invertir Griffith para mejor en diez puntos la opinión de los críticos.

9. Efecto diferencial de las críticas en comedia frente a otros géneros.

a. Al evaluar si las malas críticas afectan menos a las comedias que a otros géneros, se estimó un modelo con interacción entre la opinión de los críticos y la variable que identifica las comedias. La ecuación obtenida fue:

$$Total\ US\ Gross = -32.35\ millones + 2.965 \times Opening\ Gross + 743900 \times Critics'\ Opinion$$

Variable	Estimación	Error estándar	t value	Pr(>|t|)
Intercepto	-32350000	8927000	-3.624	0.000545
Opening.Gross	2.965	0.1946	15.232	<2e-16
Critics.Opinion	743900	157300	4.731	1.13e-05
Is_Comedy	17830000	15300000	1.165	0.248
Critics.Opinion:Is_Comedy	-212900	310800	-0.685	0.496

Métrica	Valor
R-cuadrado	0.8087
R-cuadrado ajustado	0.7978

El coeficiente de la interacción, -212900, no es estadísticamente significativo (p=0.5) indicando que no hay evidencia suficiente de que la influencia de las críticas sobre la taquilla total sea diferente para las comedias a otros géneros. Por lo que la hipótesis de Griffith de que las malas criticas perjudican menos a las películas de comedia no se puede corroborar con los datos analizados.

10. Star Power

a. Para que la afirmación de Griffith de que no es el gran presupuesto en sí mismo el que impulsa la taquilla total en Estados Unidos, sino la cantidad de actores de

primera línea, que el determina como. la variable *star power*, Al incorporar a la regresión del punto anterior esta nueva variable debería cumplirse las siguientes condiciones:

i. El coeficiente de la nueva variable debe ser positivo y estadísticamente significativo. Esto llevaría a que, manteniendo las demás variables constantes, un mayor número de actores reconocidos estaría asociado con un aumento en la taquilla total.
ii. El coeficiente de presupuesto debe disminuir en magnitud o dejar de ser tan significativo como lo es ahora, al incluir la variable *star power*. Pues si antes presupuesto aparecía con un efecto positivo, al controlas por la cantidad de actores reconocidos ese efecto tendría que reducirse en gran medida o hasta de pronto dejar de ser significativo.