# ¿Por qué tienen **éxito** las series?

**Laura Castro Cano**
**Marc Fernández i Muñoz**
**Júlia Martínez Huerta**
**Gerard Ventura i Colomer**
**Sílvia Vidal Abarca Acosta**

***Treball Seminar paper***
***2n ADE/ECO, curs 2019-20***

**Facultat de Ciències Econòmiques i Empresarials**
Universitat Pompeu Fabra

# ÍNDICE

Pág.

## 1. Resumen ejecutivo

Durante décadas hemos presenciado el éxito de muchas producciones filmográficas que aún hoy siguen presentes en nuestras vidas. Pero con la llegada de nuevas plataformas de streaming hemos podido ver cómo, en los últimos años, el auge y alcance de nuevas e incluso antiguas producciones han sido extraordinarias. Y esto se debe a la aparición de nuevos medios de streaming como Netflix, HBO e incluso la mismísima Disney no ha querido quedarse atrás en este nuevo formato de entretenimiento.

Uno de los contenidos que más se han puesto de moda son las series. Pero, ¿cuáles son los factores que hacen que una serie tenga éxito?, ¿Qué tienen en común todas estas series exitosas para llegar al "top"? ¿Qué entendemos por éxito?.
Debido a estas preguntas hemos querido realizar éste trabajo, que tiene como objetivo analizar los factores de éxito de las series.

Éste estudio ha sido realizado con webscraping[1] a partir de una de las bases de datos más grandes de contenido audiovisual (IMBd) y analizado dicha base de datos con técnicas econométricas llegando a conclusiones muy interesantes.

Así mismo, para empezar a trabajar tuvimos que definir objetivamente qué nos define el éxito de dichas series. Dado que muchas productoras mantienen sus audiencias privadas, no podíamos recurrir a este dato. De esta manera, decidimos que la manera más objetiva de definir éxito sería una combinación entre el número de votos y su puntuación media respectiva a estos.

Después de adaptar y trabajar nuestra base de datos de manera extensa, cuyo proceso se especifica más adelante, procedimos a analizar los resultados, llevándonos a conclusiones determinantes, en la medida de los datos a los que hemos podido acceder. En ellas, de

[1] El *webscraping* consiste en crear un programa para conseguir la información de una página web disponible en el código HTML. La estructura de la página puede cambiar, lo que haría incorrecto el código escrito en el apéndice. Por otra parte, al tratarse de una automatización, existe cierto riesgo de error en los datos.
Para consultas sobre el código, se ha utilizado entradas de preguntas de StackOverflow y Reddit.com/r/RStudio y la documentación de R.

manera más significativa, hemos podido observar como el hecho de estar disponible en una gran plataforma de streaming y en segundo lugar, el estar hecha por una gran productora son factores altamente influyentes en el éxito de las series.
Además de otros muchos más resultados.

## 2. Introducción

### *2.1 Motivación y objetivo*

Con este trabajo, nuestro principal objetivo, era crear un modelo que nos ayudara a entender y poner en común qué hace que una serie sea exitosa o no, y por consecuencia, qué factores influyen en más o menos medida en este éxito.
De esta manera nos centramos en encontrar posibles factores, que a nuestro parecer, de inicio podrían tener peso en la fama de una serie.
Como hemos dicho, el hecho de tener una página como IMDb, con tal cantidad de información nos ha ayudado a determinar qué factores podrían ser posibles candidatos a ser causas determinantes.

Finalmente nos inclinamos por las siguientes variables : número de temporadas, género, distribuidora, productora, minutos de media de cada capítulo, año de inicio de la serie, país de origen.

Con este estudio, queremos comprobar si las hipótesis que posteriormente planteamos se ajustan a los verdaderos factores que determinan el éxito de una serie.

### *2.2 Hipótesis previas al estudio*

Previamente a realizar el estudio, hemos planteado diferentes hipótesis respecto a las variables que determinan el éxito de una serie con el objetivo de poder contrastarlas. Nuestras hipótesis principales son que, de entre las variables que estudiaremos, aquellas que sirven como determinantes del éxito serán el género, el país de orígen (en especial, aquellas producidas en Estados Unidos), la productora y la distribuidora; mientras que las que no

afectan al éxito serán el año de inicio, la duración (media) de los episodios y el número de temporadas.

- ***Año de inicio de la serie***

Consideramos que esta variable no es un factor determinante del éxito ya que nosotros, a priori, observamos series exitosas en diferentes periodos. Creemos que el año de inicio no determina si la serie va a tener éxito, ya que eso depende del momento en el que sale. Una serie puede ser exitosa en su momento y no en un futuro; o seguirlo siendo durante mucho tiempo en cualquier época.

- ***Minutos en media de cada episodio***

Esta variable tampoco creemos que sea determinante del éxito ya que la durada de cada capítulo suele estar entre los 30 y los 60 minutos en todas las series, de manera que consideramos que no sea una variable relevante para nuestro estudio.

- ***Géneros***

En esta variable esperamos que sí sea determinante, sobretodo los géneros de drama y comedia, ya que creemos que en general, la gente se suele "enganchar" más a este tipo de serie, sea por la trama en el caso de Drama, por ejemplo, o per el guión y capacidad de entretener como en el caso de la comedia, más que en otros géneros.

- ***País***

Creemos que el hecho de que la serie venga de países como Estados Unidos o Inglaterra, que son grandes influenciadores en el mundo actual, podría ser un factor de peso a la hora de determinar el éxito de una serie.

- ***Número de temporadas***

En este caso, opinamos que a posteriori, una vez estrenada la serie, el número de temporadas sí que puede ser, en muchos casos, un derivado del éxito de esta. Pero este número no hace que la serie tenga éxito, sino que el éxito hace que tenga más temporadas.

- ***Productor***

Consideramos que es lógico pensar que según quien produzca la serie, va a tener más éxito o no. El hecho de que una gran productora como por ejemplo Warner Bros se ocupe de su producción hace mucho más probable que la serie tenga éxito que cuando lo hace una pequeña productora.

- ***Distribuidora***

En la misma línea que la variable Productor, creemos que las grandes distribuidoras son un gran escaparate para que la gente mire las series distribuidas por plataformas como Netflix. Mucho más que si no están en ellas.

## 3. Cuestiones previas al análisis de las variables

En primer lugar, vamos a analizar aspectos claves referentes a las variables que han sido utilizadas durante el estudio. Este análisis trata la obtención de las variables, da información acerca de cada una de ellas y explica los filtros aplicados para la realización del estudio.

### *3.1 Obtención de datos*

La obtención de datos ha resultado ser la parte más complicada del estudio. Hay muchos factores que determinan el éxito de las series y muchos de ellos no nos son posibles de cuantificar, factores tales como: la calidad de el guión, el estilo de grabación, o la popularidad de los actores en el momento de hacer la serie son probablemente muy importantes pero difíciles de medir. Otros factores, si bien sencillos de cuantificar, son de muy difícil acceso; así pues, variables como la inversión en publicidad o, más notablemente para la definición de éxito, el número de visualizaciones son factores potencialmente muy importantes pero se trata de datos privados de cada productora y/o distribuidora.

Además de estos dos grandes problemas, nos hemos encontrado con un gran obstáculo: la disponibilidad de datos. Si bien es cierto que en ocasiones grandes empresas publican información privada (por ejemplo, Netflix a veces publica datos sobre visualizaciones de algunas de sus mejores series), esta información está disponible para un muy reducido número de observaciones y su obtención no es uniforme (es decir, no tenemos todos estos datos en una sola página web). Así pues, para nuestro estudio, los datos tienen que cumplir con una de las siguientes dos condiciones: (1) Estar disponibles en una base de datos fiable; o (2) poderlos encontrar de una forma uniforme, es decir, todas en la misma página web – para que estén recogidos de la misma forma y poder automatizar el proceso de recopilación de datos.

Para la obtención de datos hemos utilizado la página IMDb. El hecho de extraer los datos de dicha página tiene varias ventajas:

(1) Es la mayor base de datos en cuanto a contenido audiovisual, tanto en lo referente a títulos disponibles como el gran tráfico que tiene – lo que implica un mayor número de observaciones, proporcionando así una mayor objetividad;
(2) Tienen bases de datos a disposición al público de manera gratuita y su información es fiable;
(3) Tienen una gran cantidad de información disponible sobre cada título, proporcionando así un mayor número de variables que potencialmente pueden ser incorporadas al estudio;
(4) Utilizan un identificador único para cada título, disponible en la base de datos, que nos permite trabajar con las variables y obtener información adicional de manera sencilla.
(5) El hecho de ser fundada en 1990 nos da un horizonte temporal relativamente amplio para el análisis de las series.

Los datos del estudio han sido obtenidos de maneras distintas. En un primer lugar, algunas variables de estudio (especificadas más adelante) han sido obtenidas de tres bases de datos disponibles en IMDb. Sin embargo, estas bases de datos contienen información insuficiente para un estudio de los factores de éxito de las series, por lo que hemos recurrido al *webscraping* como herramienta para la obtención de algunas de las variables del estudio. Aunque este método no tiene una fiabilidad absoluta, todas las variables obtenidas de este modo tienen un margen de error (datos incorrectos) de entre 0 y 0.02, obtenido a partir de la selección aleatoria de aproximadamente 120 observaciones para cada variable. (incluyendo NAs [un total de 37 para las variables obtenidas mediante este método]).

### *3.2 Variables*[2]

| **Nombre**[3] | **Clase de variable (R)** | **Descripción** |
|---|---|---|
| Éxito* | Double | Variable dependiente. Índice que representa el éxito de una serie. Extraída a partir de la ponderación de las variables "numVotes" (número de votos) y "averageRating" (valoración media ponderada) normalizadas sin valores atípicos. La ponderación utilizada ha sido 0.7 para el número de votos y 0.3 para la valoración media –Valoramos en mayor medida el número de votos (proxy que representa la audiencia) que la valoración media que una serie ha obtenido. Sigue una distribución $\sim N(5,1)$(Apéndice II, figura 1), la hemos normalizado con media 5 para que la variable tome únicamente valores positivos. |
| tconst** | Character | Identificador alfanumérico único del título. |
| TitleType** | Character | Tipo de título (película, corto, serie, etc). |
| RuntimeMinutes** | Integer | Duración del título. En series, duración media de los episodios. |

[2] El nombre de las variables corresponde al que tiene en la base de datos. Únicamente explicamos aquí las variables utilizadas para el estudio, hemos obviado variables que no analizaremos durante el trabajo.

[3] * Elaboración a partir de datos de la base de datos de IMDb

** Disponible en la base de datos de IMDb

*** Obtenida mediante *webscraping*.

| | | |
|---|---|---|
| Genres** | Character | Géneros que se pueden asociar con el título hasta un máximo de tres. |
| PaisPrincipal*** | Factor | País de producción del título (puede diferir del país de grabación). |
| NumeroTemporadas*** | Integer | Número de temporadas que tiene la serie. |
| Binaria_Productoras*** | Integer | 1 si está producida por una gran productora y 0 en caso contrario.<br>Definimos gran productora como aquella que ha producido 20 o más de los títulos de la lista filtrada de series. |
| Binaria_Distribuidoras*** | Integer | 1 si está disponible en Netflix, HBO o Amazon Prime; 0 en caso contrario.<br>Únicamente hemos escogido estas tres plataformas debido a su presencia territorial. |
| USA, UK, Canada, Japan, Otros*** | Double | 5 variables dicotómicas, una para cada país. 1 si el país de producción (recogido en "PaisPrincipal") coincide con el del nombre de la variable y 0 en caso contrario. |
| Drama,Comedy, Romance,Crime, Action, Thriller*** | Double | 6 variables dicotómicas, una para cada género. 1 si uno de los géneros (variable "genres") coincide con el del nombre de la variable y 0 en caso contrario. |

| | | |
|---|---|---|
| startYear** | Integer | Año de emisión o lanzamiento. Esta variable la hemos convertido en 30 dicotómicas, una para cada año, de tal forma que podremos analizar qué años tienen más influencia y hacer así un análisis más detallado. |

### *3.3 Código en R*

El código (Apéndice I) está separado en dos Scripts. En uno de ellos se encuentra el código ejecutable para el webscraping; en el otro, el relativo a la manipulación de los datos.

Webscraping: Hemos utilizado el paquete *rvest* para la recopilación de datos de la página IMDb. Mediante este paquete, es posible crear un programa mediante el cual, utilizando código CSS indicamos un elemento de una página web que queremos guardar en R. Para este proceso, es especialmente útil la clasificación que tiene IMDb de las series, puesto que cada una de ellas tiene un número de referencia único disponible en la base de datos (variable "tconst"). El proceso seguido para la obtención de las variables ha sido el siguiente: (1) especificar el enlace del que recopilar datos, cambiando la referencia con cada repetición; (2) especificar el elemento que el programa tiene que guardar; (3) Guardar el valor en un vector; (4) Crear una latencia entre repeticiones, de tal manera que la página no nos bloquee el acceso[4].

Variables: En este Script está todo el código relacionado con el filtrado de datos y demás pasos utilizados para llegar a la base de datos final con la que hacer el estudio. En el Script está explicado cada paso.

[4] Al tratarse de un programa informático, las acciones se realizan de forma muy rápida. En el ordenador utilizado, aproximadamente cada medio segundo estaríamos visitando una página del servidor de IMDb. Al detectar que todas estas entradas se realizan desde una misma dirección IP, el servicio podría pensar que se trata de un ataque de sobrecarga de servidores (DDoS) y bloquear el acceso. Para evitar esta situación y como buena práctica es importante crear dicha latencia.
En el caso de querer replicar el código del *Script* "webscraping", hay que tener en cuenta que cada variable tarda aproximadamente una hora y media para crearse y la conexión del dispositivo no puede fallar en ningún momento del proceso. Por otro lado, la estructura de la página puede cambiar y no garantizamos que el código sea válido más tarde del 05/07/2020, día en que se hizo el *webscraping* de variables.

### *3.4 Filtros aplicados*

En un principio, la base de datos con la que hemos trabajado, "title.basics", contaba con 6.721.152 observaciones. Evidentemente no todas han sido útiles para nuestro estudio, por tanto, estos son los filtros que hemos aplicado en el orden debido (número de observaciones tras filtrar a la derecha):

| | |
|---|---|
| (1) Únicamente series. Hemos descartado cualquier otro tipo de título, como películas o cortos. | 182.041 |
| (2) Géneros de interés: únicamente hemos querido analizar los 6 géneros más relevantes (los que más se repiten), esto no quiere decir que únicamente haya 6 géneros presentes, más bien, que todas las observaciones forman parte de uno de los siguientes géneros: drama, comedia, romance, crimen, acción o thriller. Hemos descartado géneros que indican que la observación no es una serie, sino un programa televisivo. Estos son: documentales, concursos, noticias y telerrealidad (Reality TV). | 36.924 |
| (3) Títulos creados a partir del 1990. De esta manera prevenimos en cierta manera el hecho que un título tenga un mayor número de valoraciones por la existencia de la página en el momento que se emitió por primera vez. | 30.885 |
| (4) Títulos con al menos 4.855 valoraciones. De esta manera conseguimos dos puntos: descartar los títulos cuya valoración sea elevada debido al bajo número de votos, obteniendo así una valoración más representativa u objetiva; y analizar el 5% de títulos que tienen un mayor éxito –estos títulos tienen, además, una correlación entre el número de votos y la puntuación del 0,21. | 1.723 |
| (5) Quitamos los puntos atípicos. Descartamos las observaciones que tienen un éxito muy elevado y las que tienen una valoración muy baja. La razón por la que hemos quitado los valores atípicos es que, de lo contrario, la normalización de las variables "numVotes" y "averageRating" se ve afectada. | 1.313 |

(6) Quitamos las series con un valor en la variable duración de episodios ("runtimeMinutes") mayor a 120, puesto que estos datos son incorrectos. 1.275

### *3.5 ¿Por qué no incluimos la variable número de temporadas en el modelo?*

Tal y como podemos observar, una de las variables que forma la base de datos es la referente al número de temporadas que tiene una serie ("*numeroTemporadas*" en la base de datos). Esta se trata de una de las variables obtenidas mediante webscraping. Este hecho es debido a que, al inicio de la realización del estudio, pensamos que esta podría ser una de las variables determinantes del éxito de una serie y por lo tanto, que debía de formar parte de nuestro modelo. Pero, a medida que fuimos avanzando, nos dimos cuenta de que dicha variable más que ser un factor determinante del éxito (que es lo que nosotros queríamos estudiar) era una consecuencia de este, es decir, el éxito que alcanzaba una serie motivaba la creación e inversión por parte de las productoras en nuevas temporadas.

Es por este motivo, que finalmente tomamos la decisión de no incluir dicha variable en la creación de nuestro modelo estimado tal y como se puede observar en el comando reg explicado más adelante.

## 4. Análisis de las variables

### *4.1 Metodología*

Para poder llevar a cabo nuestro estudio, hemos decidido realizar un modelo de regresión múltiple donde hemos incluido los siguientes regresores: runtimeminutes (media en minutos de duración del capítulo), binaria_productora, binaria_distribuidora, pais ( que se separan en usa, uk, japan, canada y otros países), género (que se separa en drama, comedy, romance, crime, action y thriller) y startyear (que se separan en una dummy por cada posible año).

Antes de empezar con el análisis, hemos construido nuestro modelo estimado que es el siguiente:

```
. reg exito runtimeminutes binaria_distribuidoras binaria_productoras usa uk japan canada drama comedy romance crime
>  action x1990dicotomica x1991dicotomica x1992dicotomica x1993dicotomica x1994dicotomica x1995dicotomica x1996dicot
> omica x1997dicotomica x1998dicotomica x1999dicotomica x2000dicotomica x2001dicotomica x2002dicotomica x2003dicotom
> ica x2004dicotomica x2005dicotomica x2006dicotomica x2007dicotomica x2008dicotomica x2009dicotomica x2010dicotomic
> a x2011dicotomica x2012dicotomica x2013dicotomica x2014dicotomica x2015dicotomica x2016dicotomica x2017dicotomica
> x2018dicotomica x2019dicotomica, r
```

Linear regression

| | |
|---|---|
| Number of obs | = 1,272 |
| F(42, 1229) | = 3.04 |
| Prob > F | = 0.0000 |
| R-squared | = 0.0903 |
| Root MSE | = .79155 |

| exito | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| runtimeminutes | -.005602 | .0014973 | -3.74 | 0.000 | -.0085396 | -.0026644 |
| binaria_distribuidoras | .385395 | .0595252 | 6.47 | 0.000 | .2686128 | .5021772 |
| binaria_productoras | .112328 | .0500189 | 2.25 | 0.025 | .014196 | .2104599 |
| usa | .0278178 | .075422 | 0.37 | 0.712 | -.1201522 | .1757879 |
| uk | .1692676 | .0898346 | 1.88 | 0.060 | -.0069786 | .3455137 |
| japan | .068908 | .1228386 | 0.56 | 0.575 | -.1720884 | .3099045 |
| canada | -.0733546 | .1112754 | -0.66 | 0.510 | -.2916653 | .1449561 |
| drama | .2337092 | .0619131 | 3.77 | 0.000 | .1122422 | .3551763 |
| comedy | -.0750195 | .0604321 | -1.24 | 0.215 | -.193581 | .043542 |
| romance | -.0284049 | .0708492 | -0.40 | 0.689 | -.1674037 | .1105939 |
| crime | .0030216 | .0621579 | 0.05 | 0.961 | -.1189258 | .124969 |
| action | .1641403 | .0660124 | 2.49 | 0.013 | .0346309 | .2936497 |
| x1990dicotomica | .044938 | .251257 | 0.18 | 0.858 | -.4480022 | .5378782 |
| x1991dicotomica | .1657658 | .2469591 | 0.67 | 0.502 | -.3187422 | .6502739 |
| x1992dicotomica | .0065371 | .2461806 | 0.03 | 0.979 | -.4764438 | .4895179 |
| x1993dicotomica | .3707123 | .2695245 | 1.38 | 0.169 | -.1580668 | .8994915 |
| x1994dicotomica | -.0145739 | .2721408 | -0.05 | 0.957 | -.5484858 | .519338 |
| x1995dicotomica | .2348084 | .2647192 | 0.89 | 0.375 | -.2845432 | .7541601 |
| x1996dicotomica | .4205069 | .2625755 | 1.60 | 0.110 | -.0946391 | .9356528 |
| x1997dicotomica | .3901033 | .2273209 | 1.72 | 0.086 | -.0558767 | .8360833 |
| x1998dicotomica | .2443536 | .2891328 | 0.85 | 0.398 | -.3228949 | .8116022 |
| x1999dicotomica | .4266652 | .286474 | 1.49 | 0.137 | -.135367 | .9886973 |
| x2000dicotomica | .1470714 | .2600933 | 0.57 | 0.572 | -.3632046 | .6573473 |
| x2001dicotomica | .284768 | .2496585 | 1.14 | 0.254 | -.2050361 | .7745721 |
| x2002dicotomica | .2413307 | .2840789 | 0.85 | 0.396 | -.3160026 | .798664 |
| x2003dicotomica | .5886763 | .2884412 | 2.04 | 0.041 | .0227847 | 1.154568 |
| x2004dicotomica | .4691587 | .2460561 | 1.91 | 0.057 | -.0135778 | .9518952 |
| x2005dicotomica | .1404894 | .2459336 | 0.57 | 0.568 | -.3420066 | .6229855 |
| x2006dicotomica | .2246747 | .2452522 | 0.92 | 0.360 | -.2564846 | .7058339 |
| x2007dicotomica | .3921574 | .2476284 | 1.58 | 0.114 | -.0936638 | .8779787 |
| x2008dicotomica | .131257 | .2368561 | 0.55 | 0.580 | -.33343 | .5959441 |
| x2009dicotomica | .4519322 | .2388603 | 1.89 | 0.059 | -.0166869 | .9205513 |
| x2010dicotomica | .2393349 | .2239672 | 1.07 | 0.285 | -.2000655 | .6787353 |
| x2011dicotomica | .2126314 | .2288061 | 0.93 | 0.353 | -.2362625 | .6615252 |
| x2012dicotomica | .1305091 | .2195872 | 0.59 | 0.552 | -.3002982 | .5613164 |
| x2013dicotomica | .1070284 | .2287753 | 0.47 | 0.640 | -.341805 | .5558619 |
| x2014dicotomica | .139649 | .2176711 | 0.64 | 0.521 | -.2873991 | .5666971 |
| x2015dicotomica | .2114638 | .2193252 | 0.96 | 0.335 | -.2188296 | .6417571 |
| x2016dicotomica | -.0068814 | .2153988 | -0.03 | 0.975 | -.4294715 | .4157086 |
| x2017dicotomica | .0645153 | .2238518 | 0.29 | 0.773 | -.3746588 | .5036893 |
| x2018dicotomica | .1303722 | .2213972 | 0.59 | 0.556 | -.3039861 | .5647306 |
| x2019dicotomica | -.0990395 | .2188431 | -0.45 | 0.651 | -.5283869 | .330308 |
| _cons | 4.714106 | .2254517 | 20.91 | 0.000 | 4.271793 | 5.156419 |

**Figura 1**: *Modelo de regresión múltiple estimado.*

Dicho modelo lo hemos conseguido utilizando el comando *reg* que se puede observar en la parte superior. A parte de las betas estimadas usadas para crear nuestro modelo, este comando también nos proporciona información respecto al p_value, el estadístico de contraste y el intervalo de confianza correspondientes a cada una de las variables independientes.

Una vez teníamos toda esta información, hemos comenzado a analizar de manera individual el efecto de cada una de las variables regresoras en nuestra variable dependiente. Para ello, en función del tipo de variable de estudio hemos realizado un contraste de significación individual, que resolveremos con los valores obtenidos anteriormente, o conjunta, que resolveremos con la ayuda adicional del comando test.

El principal objetivo es poder resolver la pregunta de si dichas variables influyen o no sobre el éxito.

### *4.2 Estudio de las variables*

#### *4.2.1 Runtimeminutes*

La primera variable que analizaremos será los minutos de media que dura cada capítulo con el fin de saber si es un factor que determina el éxito. Para ello llevaremos a cabo el siguiente contraste:

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

Una vez introducido en STATA el comando *reg* anterior, hemos obtenido con un 5% de significancia un p-value asociado a runtimeminutes más pequeño que éste, lo cual nos lleva a rechazar la hipótesis nula. De manera que la variable analizada influye en nuestra variable dependiente. Si contextualizamos el resultado, esto quiere decir que la media de los capítulos es un factor determinante del éxito. El resultado obtenido no apoya nuestra hipótesis previa al estudio, en la cual considerábamos que no era un factor determinante debido a que la mayoría de capítulos tienen una durada similar.

#### *4.2.2 Binaria distribuidora*

$H_0 : \beta_2 = 0$

$H_1 : \beta_2 \neq 0$

En cuanto a la variable distribuidora queríamos saber si una serie que es distribuida por una gran plataforma hace que en media tenga más éxito.
Observamos que se rechaza la $H_0$ debido a que el p-valor es 0 (más pequeño que 0.05) y también porqué el t-estadístico es 6.47 (más grande que 1.96). Por lo tanto, sabemos que la variable distribuidora sí que influye en el éxito de una serie.

Si interpretamos la $\beta$ obtenida podemos confirmar que se produce un aumento del éxito cuando una serie no está distribuida por una gran distribuidora a sí estarlo de 0.38 de media. Nos parece muy razonable el resultado obtenido ya que creemos que, si una serie es distribuida por alguna gran plataforma como Netflix, HBO o Amazon Prime, esta pueda alcanzar un mayor público y tener más éxito. Por lo tanto nuestra hipótesis se confirma.

#### *4.2.3 Binaria productoras*

La siguiente variable a analizar es la productora de la serie, la cual es una variable binaria que toma valor 1 cuando la responsable de realizarla es una gran productora, mientras que toma valor 0 en el resto de casos. Nuestro principal objetivo es estudiar si afecta más en el éxito que haya sido realizada por una gran productora o no.

$H_0 : \beta_3 = 0$

$H_1 : \beta_3 \neq 0$

Para ello hemos usado el soporte de STATA introduciendo el comando *reg,* presentado anteriormente.

Al tratarse de una variable binaria el análisis es el siguiente:

- El valor de la $\beta_0$, es decir, la constante corresponde a la media de éxito que puede alcanzar una serie realizada por una productora que no esté considerada como grande.
- El valor de la $\beta_1$, es la diferencia de medias entre la media de éxito cuando la variable toma valor 1, es decir, la serie está realizada por una gran productora y la media de éxito cuando no es el caso. De manera que se interpreta como el aumento de éxito cuando se pasa en la variable binaria del valor 0 al 1.

Resolviendo el contraste, como la $\beta_3$=0 de nuestra hipótesis nula no está incluido en el intervalo de confianza al 95%, llegamos a la conclusión que rechazamos la hipótesis nula de manera que esta variable sí influye en el éxito de una serie.

Si interpretamos las $\beta$ 's obtenidas con el comando anterior podemos confirmar que se produce un aumento del éxito cuando pasamos de productoras regulares a grandes de 0,13.

### *4.2.4 País principal*

$H_0$ : $\beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$

$H_1$ :al menos una de las $\beta$ diferente a 0

La siguiente variable de estudio es el país principal de producción del título. El objetivo principal del estudio de esta variable es saber si en función del país encargado de la producción una serie está predeterminada a tener más éxito o no, es decir, saber si el país productor influye o no en el éxito de una serie.

Previamente a comenzar con el análisis hemos tenido que crear una variable binaria para cada una de las subcategorías que componen la variable país principal: USA, UK, Japan, Canada y Otros.

A continuación, hemos realizado un contraste de significación conjunta con la ayuda del comando test en STATA y donde hemos utilizando la variable binaria "Otros" como nuestra variable de referencia.

Una vez ya teníamos toda esta información, hemos realizado el análisis de la variable llegando a la conclusión de que no rechazamos la hipótesis nula con un al 95% de confianza. Por lo tanto, llegamos a la conclusión de que el país productor no es un factor que determine el éxito de una serie.

```
4 . test usa uk japan canada

 ( 1)  usa = 0
 ( 2)  uk = 0
 ( 3)  japan = 0
 ( 4)  canada = 0

       F(  4,  1229) =    1.53
            Prob > F =    0.1905
```

**Figura 2**: *Contraste de significación conjunta de la variable "País principal".*

### *4.2.5 Géneros*

$H_0$ : $\beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = 0$

$H_1$ :al menos una de las β diferente a 0

Para analizar la variable de género de las series hemos hecho un contraste de significación conjunta. Hemos omitido la variable "Thriller" ya que es la variable de referencia. Una vez obtenidos los resultados de la regresión múltiple de nuestro modelo hemos realizado la prueba "test" en STATA para saber si se rechaza la hipótesis nula.

Analizando el "test" rechazamos la hipótesis nula y concluimos que la variable género sí que influye en el éxito de una serie ya que el p-value es 0. Por lo tanto nuestra hipótesis era verdadera.

Si analizamos las β del modelo de regresión múltiple podemos ver cómo las variables "drama" y "action" son las dos variables que sí que influyen en el éxito de una serie (rechazamos hipótesis nula). En cambio en las otras variables no se puede rechazar la hipótesis nula en un intervalo de confianza del 95%.
Por lo tanto, observamos que si una serie pertenece a los géneros "drama" o "action" de media tendrán 0.233 y 0.164 más de éxito que la variable de referencia.

```
6 . test drama comedy romance crime action

   ( 1)  drama = 0
   ( 2)  comedy = 0
   ( 3)  romance = 0
   ( 4)  crime = 0
   ( 5)  action = 0

         F(  5,  1229) =    6.47
              Prob > F =    0.0000
```

**Figura 3**: *Contraste de significación conjunta de la variable "Género".*

### *4.2.6 Startyear*

En último lugar, hemos analizado la variable año de inicio (startyear), con el propósito de saber si el año de inicio de la serie afecta al éxito o no. Previamente al análisis, hemos

convertido cada uno de los posibles años de inicio en una variable binaria con el principal objetivo de diferenciar el efecto individual de cada año en el éxito de una serie.

A la hora de hacer la regresión, hemos omitido la variable "x2020dicotomica" para evitar multicolinealidad perfecta. Como consecuencia, dicha variable se ha convertido en nuestra variable de referencia.

A continuación, hemos realizado el siguiente contraste de significación conjunta:

$H_0$ : $\beta_{13} = \beta_{14} = \ldots = \beta_{40} = \beta_{41} = \beta_{42}$
$H_1$ : al menos una de las β diferente a 0.

Para resolverlo, hemos utilizado el comando *test* en STATA el cual nos proporciona la información necesaria para resolverlo. Una vez ya teníamos a nuestra disposición toda la información, hemos llegado a la conclusión de rechazar la hipótesis nula con un 95% de confianza ya que el p_value que hemos obtenido es inferior al valor de alfa correspondiente. Si aplicamos el resultado a nuestro estudio, lo que nos está indicando este resultado es que la variable año de inicio, es decir, el conjunto de las diferentes dummies, sí que influye conjuntamente en el nivel de éxito que puede llegar a alcanzar.

Si a continuación analizamos las diferentes betas obtenidas en el comando reg, podemos observar que no hay ningún año que destaque significativamente en nuestra variable independiente.

En resumen, en conjunto la variable año de inicio si que afecta al éxito que puede llegar a alcanzar una serie ya que por ejemplo, en los últimos años se ha podido observar un incremento en la publicidad (debido al incremento en el uso de las redes sociales) o el surgimiento de nuevas plataformas que hacen más accesible el consumo de series pero, si analizamos de manera individual cada año, observamos que no hay ninguno que tenga una influencia significativa en el éxito.

```
. test x1990dicotomica x1991dicotomica x1992dicotomica x1993dicotomica x1994dicotomica x1995dicotomica x1996dicotomica x1997
> dicotomica x1998dicotomica x1999dicotomica x2000dicotomica x2001dicotomica x2002dicotomica x2003dicotomica x2004dicotomica
>  x2005dicotomica x2006dicotomica x2007dicotomica x2008dicotomica x2009dicotomica x2010dicotomica x2011dicotomica x2012dico
> tomica x2013dicotomica x2014dicotomica x2015dicotomica x2016dicotomica x2017dicotomica x2018dicotomica x2019dicotomica

 ( 1)  x1990dicotomica = 0
 ( 2)  x1991dicotomica = 0
 ( 3)  x1992dicotomica = 0
 ( 4)  x1993dicotomica = 0
 ( 5)  x1994dicotomica = 0
 ( 6)  x1995dicotomica = 0
 ( 7)  x1996dicotomica = 0
 ( 8)  x1997dicotomica = 0
 ( 9)  x1998dicotomica = 0
 (10)  x1999dicotomica = 0
 (11)  x2000dicotomica = 0
 (12)  x2001dicotomica = 0
 (13)  x2002dicotomica = 0
 (14)  x2003dicotomica = 0
 (15)  x2004dicotomica = 0
 (16)  x2005dicotomica = 0
 (17)  x2006dicotomica = 0
 (18)  x2007dicotomica = 0
 (19)  x2008dicotomica = 0
 (20)  x2009dicotomica = 0
 (21)  x2010dicotomica = 0
 (22)  x2011dicotomica = 0
 (23)  x2012dicotomica = 0
 (24)  x2013dicotomica = 0
 (25)  x2014dicotomica = 0
 (26)  x2015dicotomica = 0
 (27)  x2016dicotomica = 0
 (28)  x2017dicotomica = 0
 (29)  x2018dicotomica = 0
 (30)  x2019dicotomica = 0

       F( 30,  1229) =    1.53
            Prob > F =    0.0342
```

**Figura 4**: *Contraste de significación conjunta de la variable "Startyear".*

## 5. Interpretación de las medidas del buen ajuste

Una vez realizada la regresión múltiple de nuestro estudio en STATA, hemos podido observar que el valor del coeficiente de determinación tiene un valor de 0.0903. Como podemos observar este valor es muy cercano a 0, factor que nos indicaría que la predicción que nosotros hemos realizado con el modelo es poco fiable. Dado nuestro tema de estudio, es entendible que obtengamos un valor tan pequeño ya que hay variables que no están incluidas en el modelo o bien por que son confidenciales o porque no son cuantificables. Por ello, aunque a priori parezca un valor relativamente pequeño, dados los datos con los que contábamos para realizar el modelo podríamos decir que es un valor significativamente grande. En el apéndice II (figura 2) encontramos la distribución de los residuales del modelo, con desviación estándar 0,778 y media 0.

Un ejemplo de variable no incluida en el modelo sería el gasto en publicidad que pensamos que puede tener un gran peso a la hora de considerar el éxito, ya que esta inversión que se realiza puede ser de gran provecho para acercar la serie significativamente al usuario y que sea mucho más seguida a nivel mundial.

Por otro lado, como ejemplo de variable no cuantificable que puede influir en el éxito es el nivel de fama de los actores previo a la realización de la serie. El principal problema de esta variable es que es muy complicado cuantificar la fama de los protagonistas. Hay muchos actores que consideramos famosos actualmente pero que en el momento de la emisión no tenían tal reputación.

Todo lo argumentado anteriormente se puede plasmar en el elevado valor que toma el Root-MSE (0.79), el cual nos informa de la desviación típica estimada del error. Como podemos observar el error tiene un gran peso en nuestro modelo ya que incluye todas las variables que estamos omitiendo y afectan al éxito.

## 6. Análisis de las series con mayor éxito (puntos atípicos)

Al contrario del análisis hecho con STATA, en el que hemos eliminado los puntos atípicos para hacer nuestro modelo de regresión múltiple, en este análisis queríamos analizar estos puntos atípicos para ver qué factores compartían. Creíamos que los factores comunes entre las series con más éxito serían las variables que sí que influyen en el modelo que hemos realizado anteriormente.

Tal y como podemos observar en los gráficos, podemos sacar conclusiones bastante claras.

En primer lugar, vemos que un 63% de las series están en una gran distribuidora (Netflix, HBO, Amazon Prime).

Nuestro modelo de regresión múltiple también predice que la variable distribuidora es una de las variables que más influye.

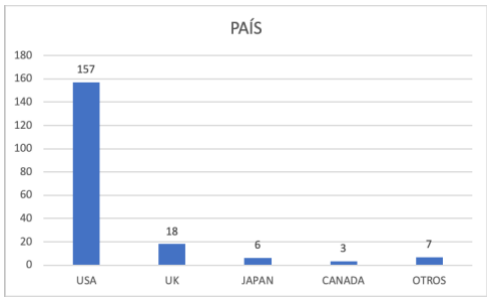

**Figuras 5 y 6**: *Gráficos de las variables "Distribuidora" y "Productora".*

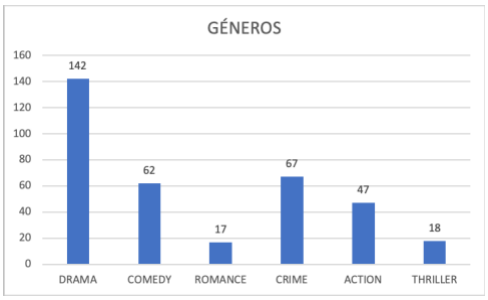
En segundo lugar, también observamos que un 68% de las series están rodadas por una gran productora. Ésta también era una de las variables que más influía en nuestro modelo.

Podemos ver, además, cómo la mayoría de las series son originarias de Estados Unidos. En nuestro modelo la variable país de origen no es significante. Estados Unidos no es significante aunque muchas series de éxito sean de allí ya que, seguramente las series con menos éxito también son originarias del país en cuestión, factor que acaba provocando una compensación.

En cuanto a la variable género, el que más se repite y que predomina en las series con mayor éxito es el drama (142), seguido por crimen (67) y comedy (62).

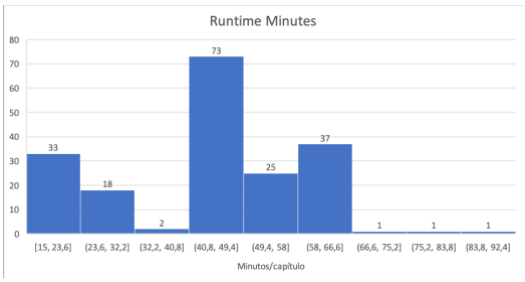
Nuestro modelo predice que las variables "drama" y "action" son las dos variables de género que influyen en el éxito de una serie. En este caso vemos como drama sí que predomina entre las series con más éxito, pero acción no.

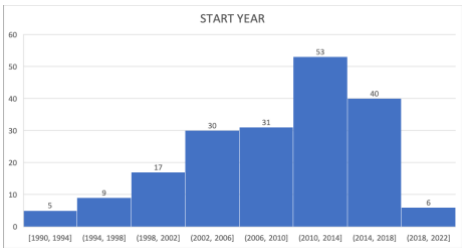



**Figuras 7 y 8**: *Gráficos de las variables "País" y "Género".*

Si observamos las variables año de inicio de la serie y media de minutos por capítulo podemos observar como la mayoría de las series con más éxito se distribuyen alrededor de 2010 y 2018. Suponemos que las series con más éxito se concentran en estos años ya que coincide con la llegada y popularidad de plataformas de streaming (la variable que más influye en nuestro modelo). También creemos que es pertinente decir que las series más recientes (2018-actualidad) no son demasiadas por el hecho de que aún no han tenido el tiempo suficiente de ser exitosas. Los minutos de las series se distribuyen de forma un poco desigual. La mayoría de las series duran entre 40 y 50 minutos. La media de minutos de las series que estamos analizando es de 43 minutos y medio.




**Figuras 9 y 10**: *Gráficos de las variables "Start Year" y "Runtime Minutes".*

## 7. Conclusiones

Para terminar el trabajo comentaremos los puntos y conclusiones más importantes. En primer lugar decir que bastantes variables que hemos estudiado influyen en el éxito de una serie.

Las que influyen en mayor medida en el éxito son, primeramente, el hecho de estar disponible en una gran plataforma de streaming (probablemente debido a la disponibilidad) y en segundo lugar, el estar hecha por una gran productora (debido a la gran cantidad de dinero disponible para invertir en una serie). Dichos resultados concuerdan con las hipótesis previas que realizamos antes de empezar al estudio.

También hemos podido observar que la variable de año de inicio y tiempo medio de capítulos, que a priori teníamos bastante claro que serían variables que no afectarían, han resultado ser factores que sí que marcan el éxito de la serie.

Otro resultado sorprendente de nuestro estudio ha estado respecto al país principal productor el cual pensábamos, y así se ve reflejado en las hipótesis, que sí que sería un factor determinante y ha resultado que no es así. Este resultado pensamos que puede estar relacionado con que el país en si no es lo que determina el éxito sino todo el trabajo por parte de productoras, distribuidoras, publicidad…

La última variable ha sido el género el cual sí que afecta al éxito, tal y como esperábamos, sobre todo aquellas que estén categorizadas como series de drama o acción.

Por otro lado, como ya hemos comentado anteriormente, es oportuno destacar que nuestro modelo se ajusta bastante teniendo en cuenta la gran cantidad de variables que influyen en el éxito de las series y que no hemos podido encontrar y/o cuantificar. Es por este motivo que creemos que este trabajo podría seguir en un futuro intentando recopilar más datos sobre otras variables e intentar hacer un modelo que se ajuste aún más. Con el propósito de crear un modelo mejor, también creemos que se podría continuar un análisis más individualizado según géneros, debido a que lo que puede ser bueno para un género, igual no lo es tanto para otro y esto dificulta la modelización de los factores de éxito. Una de las variables que se podría

explicar mejor según el género es la duración media de los episodios (Apéndice III). En el apéndice se observa, a priori, que la duración media característica de los episodios varía según género. A modo de ejemplo, los dramas (Apéndice III, figura 6) necesitan una duración más larga que las comedias (Apéndice III, figura 3).

En general, todo este estudio realizado podría ser utilizado por grandes productoras para saber qué perfiles de series son las que tienen más éxito y explotar los resultados del estudio con el fin de aumentar los beneficios.

## 8. Agradecimientos

El más sincero agradecimiento a nuestro tutor José Garcia Montalvo, doctor en economía por la Harvard University y profesor de economía en la Universitat Pompeu Fabra, por guiarnos y ayudarnos rápidamente en todo momento, a pesar de las difíciles circunstancias en las que nos encontramos.

También agradecer a todos los miembros del equipo por la dedicación, tiempo y esfuerzo que han sido necesarios para llevar a cabo el estudio y por amoldarse de manera tan efectiva a la difícil situación.

## 9. Bibliografía

**Libro:**

J.H. Stock and M.W. Watson, *Introduction to Econometrics* (third edition), Addison-Wesley

**Páginas Web:**

IMDb. (14 de Abril de 2020). *IMDb*. Obtenido de https://www.imdb.com

IMDb. (15 de Abril de 2020). *IMDb datasets*. Obtenido de https://datasets.imdbws.com

Pacholski, L. (03 de Mayo de 2020). *CSS Diner*. Obtenido de http://flukeout.github.io/#

## 10.Apéndice I

*Código en R*

La base de datos junto con los scripts del código están disponibles en el enlace a continuación: https://bit.ly/2AXIYPd

En el documento "README" se encuentra un resumen de las variables, anotaciones importantes a tener en cuenta y los archivos necesarios para la utilización del código.

## 11.Apéndice II

*Distribución de la variable dependiente "Éxito" y de los residuales del modelo.*

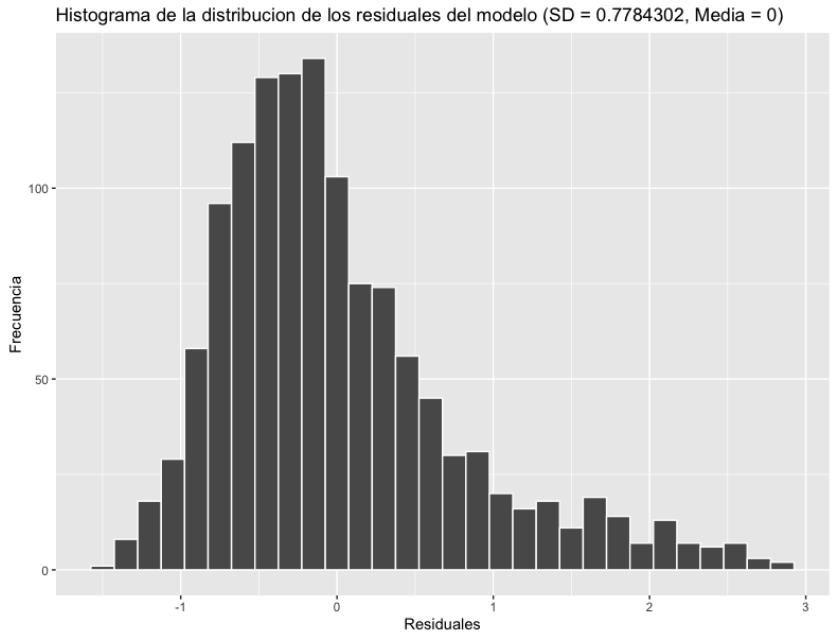



**Figura 1 :** *Distribución de los residuales del modelo.*

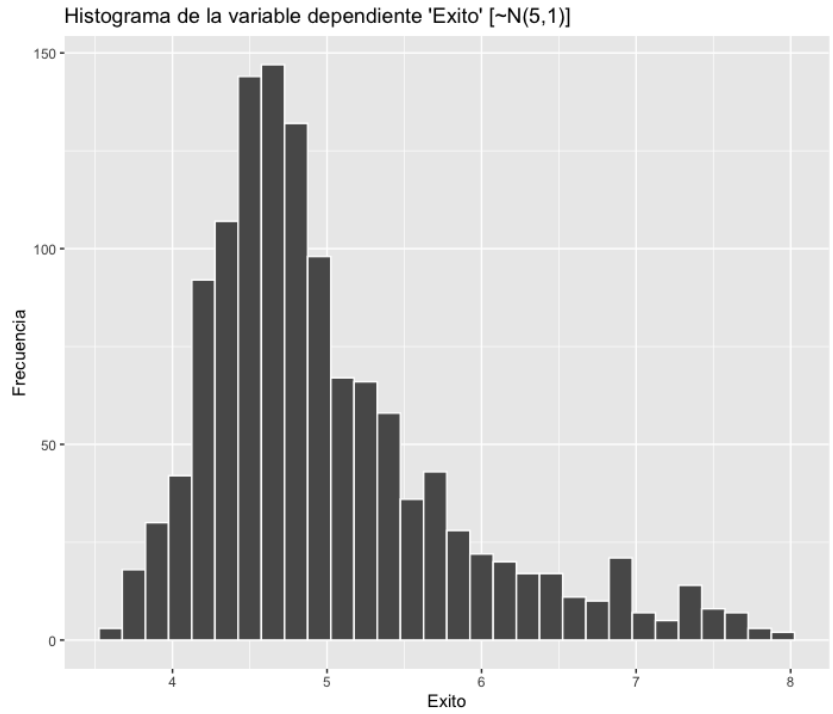



**Figura 2 :** *Histograma de la variable dependiente "Éxito".*

## 12.Apéndice III

*Distribución de la variable "RuntimeMinutes" según género.*

Nota: el modelo mostrado se trata de un GAM con un 95% de confianza, no el modelo lineal utilizado

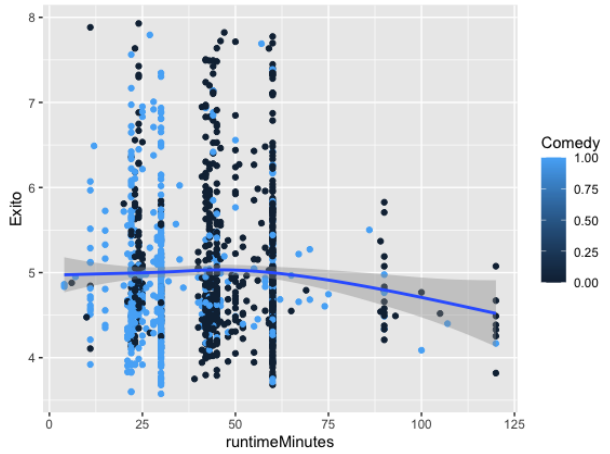



**Figura 3 :** *Distribución "RuntimeMinutes" en Comedia*

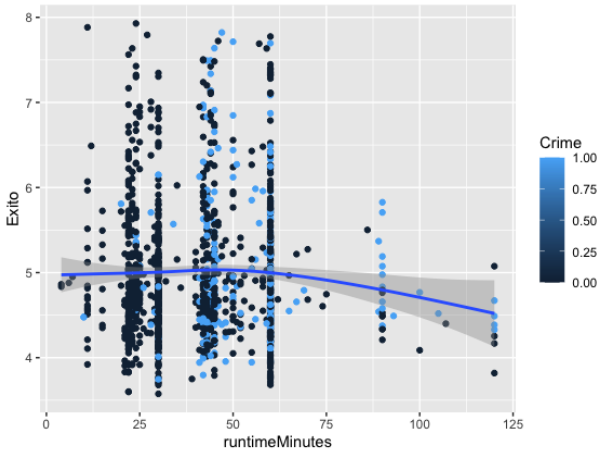



**Figura 4 :** *Distribución "RuntimeMinutes" en Crimen*

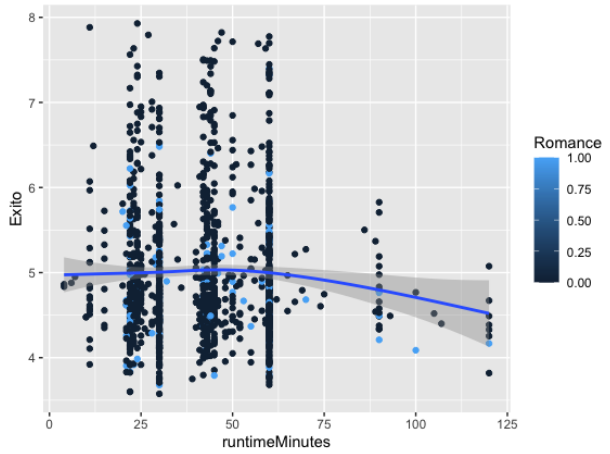



**Figura 5 :** *Distribución "RuntimeMinutes" en Romance*

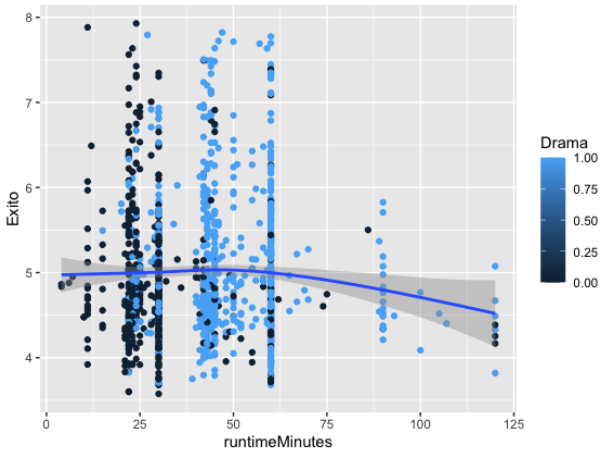



**Figura 6 :** *Distribución "RuntimeMinutes" en Drama*

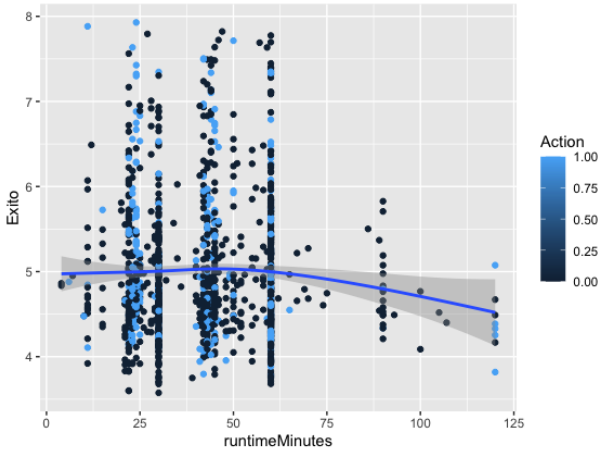



**Figura 7 :** *Distribución "RuntimeMinutes" en Acción*

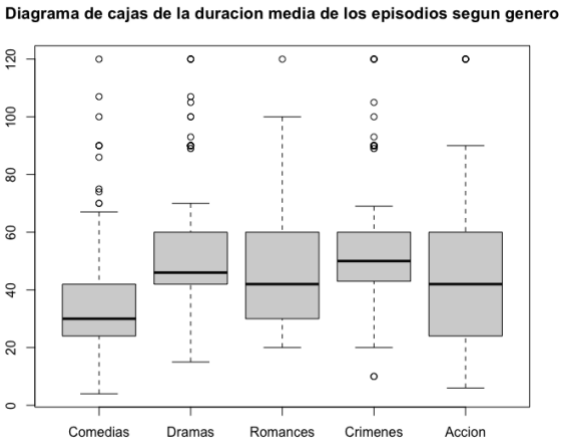



**Figura 8 :** *Diagrama de cajas de la duración media de los episodios según género*