**Taller IV**
**Pronósticos en Sección Cruzada – Tendencia, Estacionalidad y Ciclos.**
**2022 - I**

**Instrucciones para la entrega:**

1. El taller debe ser entregado en grupos de máximo 5 personas y mínimo de 3 personas. No se aceptarán trabajos que no sean realizados en grupo, con el número indicado de personas.

2. Los talleres deben ser entregados de **manera ordenada** en un solo archivo PDF. La solución de cada punto debe estar organizada numeral por numeral de menor a mayor. Si lo hacen así nombren el archivo de PDF -Solución Taller-

   Ahora, también pueden entregar el taller punto por punto en un archivo PDF distinto. La solución de cada punto debe estar organizada numeral por numeral de menor a mayor. Debe haber tantos PDF como puntos del taller y el nombre de cada archivo debe ser -Solucion X- donde X es el número del punto (p.ej. Solución 1, es la solución al punto 1 del taller).

3. Cada pregunta empírica debe estar acompañada por el M-File o Do-File y la base de datos relacionada. En caso que sea necesario revisar la programación de las respuestas dadas en el PDF, **los códigos deben correr** y así se corroborará la valides de sus respuestas.

4. Los M-FILE o Do-File deben tener las secciones y comentarios respectivos donde se describe el paso a paso de lo que realizan. En caso de usar archivos de excel para realizar gráficas o estimaciones secundarias descríbanlo en el M-FILE, Do-File y en el archivo de PDF donde responden formalmente el taller. Todos los archivos deben ser adjuntados.

5. Todos los PDF y M-File - Do-File - enviados deben estar marcados al inicio con el nombre y el código de cada uno de los integrantes del grupo.

6. Todos los archivos usados deben ser enviados al correo electrónico szapata@uniandes.edu.co en un archivo comprimido cuyo nombre será el primer apellido de cada uno de los integrantes del grupo a más tardar el 4 de marzo de 2022.

7. **No seguir las instrucciones y/o no entregar la solución del taller de manera ordenada y comprensible causará que el taller sea calificado sobre 3.**

## 1. Pronósticos en Sección Cruzada

Ustedes tienen una nueva posición en el equipo de mercadeo de la mayor empresa de cine en Colombia y después de dos años muy difíciles la empresa los contrata como el grupo de expertos que le ayudará a mejorar los pronósticos sobre las películas que más les gustarán a las personas. Sus estimaciones le servirán al equipo de inversiones de la empresa para informarse sobre los estrenos en los que deberían invertir, de igual forma el equipo de logística tendrá en cuenta sus estimaciones para definir el número de salas necesarias para satisfacer la demanda de la película que adquiera la empresa.

Para empezar su análisis ustedes toman una información filtrada y organizada que encontraron en internet. La información cuenta con la puntuación dada por IMDb, una empresa especializada en recomendar películas y series a nivel mundial[1]. La base de datos cuenta con 24 variables necesarias para más de 3.000 películas y carteles, que abarcan casi 100 años y 66 países. Hay más de 1.500 nombres de directores únicos y miles de actores/actrices. Las 25 variables con las que cuenta la base de datos son[2]:

**Tabla 1. Descripción de Variables exógenas de la estimación a realizar**

| **Variables** | **Descripción** |
|---|---|
| duration | Duración en minutos |
| director_name | Nombre del director de la película |
| director_facebook_likes | Número de likes del director en su página de facebook |
| actor_1_name | Actor principal de la pelicula |
| actor_1_facebook_likes | Número de likes del actor principal en su página de facebook |
| actor_2_name | Actor secundario que protagoniza la película |
| actor_2_facebook_likes | Número de likes del actor secundario en su página de facebook |
| actor_3_name | Actor terciario que protagoniza la película |
| actor_3_facebook_likes | Número de likes del actor terciario en su página de facebook |
| num_user_for_reviews | Número de usuarios que hicieron un review de la película |
| num_critic_for_reviews | Número de reviews con criticas en imdb |
| num_voted_users | Número de personas que votaron por la película |
| cast_total_facebook_likes | Número total de likes en facebook de todos los actores de la película |
| movie_facebook_likes | Número de likes en el facebook de la película |
| facenumber_in_poster | Número de actores de la película que aparecen en el poster de la película |

[1] https://www.imdb.com/

[2] Los datos necesarios se encuentran en el archivo de Excel denominado RATING que se encuentra adjunto en la carpeta de esta tarea.

| | |
|---|---|
| Color | Es una película en blanco y negro o a color |
| Genres | Categoría de la pelicula: 'Animation', 'Comedy', 'Romance', 'Horror', 'Sci-Fi', ' |
| title_year | Año en el que la película se estreno (1916:2016) |
| language | Idioma: English, Arabic, Chinese, French, German, Danish, Italian, Japanese |
| country | País en donde la película se produjo |
| content_rating | Rating del contenido de la película |
| aspect_ratio | Tamaño de grabación en el que la película se hizo |
| Gross | Ganancias de la película en dólares |
| budget | Presupuesto de la película en dólares |

Todos los integrantes del equipo de pronósticos de la empresa estudiaron economía y tristemente no conocen una teoría que los guie sobre cuáles son los determinantes para definir el rating de las películas. Sin embargo, saben que un modelo de proyección lineal puede ayudarles a dar una muy buena respuesta sobre los factores que definen que una película tenga un buen rating. Por esa razón inician su proceso de estimación con el siguiente modelo en mente:

$$Y_i = \alpha_0 + \sum_{i=1}^{24} \beta_i x_i + \varepsilon_i$$

Donde $Y_i$ es la variable (imdb_score) que mide el rating de cada película i en la muestra, $x_i$ representa cada una de las variables exógenas de la muestra para cada película i (tabla 1), $\alpha_0$ es la constante de la estimación, $\beta_i$ el parámetro que define cómo cada variable $x_i$ afecta la proyección de tener un buen o mal rating y $\varepsilon_i$ es el error de pronóstico de su estimación.

Para decidir cuál es el mejor modelo de proyección lineal siguen los siguientes pasos, entre los cuales se encuentra recordar algunos aspectos teóricos básicos:

## 1.1. Análisis Preliminares (15 Puntos)

1. ¿Cuál es el pronóstico óptimo para el rating de cada película i en la muestra? Discutan los supuestos necesarios para que el pronóstico propuesto sea, de hecho, óptimo.

2. Estimen las ganancias de cada película i (Gross) y el presupuesto de cada película i (nudget) como porcentaje del PIB de USA conforme al lanzamiento de la película (title-year). Para eso utilice los datos que se encuentran en la hoja de cálculo (USA GDP)[3].

3. Convierta todas las variables que muestran características cualitativas de la base de datos en variables categóricas (i.e. color, language, country, content_rating, director_name, actor_1_name, actor_2_name, actor_3_name, movie_title, genres).

4. Usando las observaciones de la muestra (las que reportan rating), haga un correlograma de sólo las variables numéricas de la base de datos y responda la siguiente pregunta[4]:

   a. ¿Cuáles son las variables que tienen una correlación mayor o igual a $|0{,}15|$ con el rating IDB?.

   b. Reporte el corrplot de las variables que tienen una correlación superior a $|0{,}15|$ con IDB y el IDB.

5. Usando las observaciones de la muestra (las que reportan rating), haga diagramas de dispersión en los que en el eje Y se encuentre el puntaje imdb (imdb_score) y en el eje X se encuentre cada una de las variables para los que la correlación es superior a $|0{,}15|$. Sin reportar los diagramas en la respuesta del taller, pero estos si deben correr en el código, respondiendo la siguiente pregunta:

   a. ¿Se observa evidencia de alguna relación no lineal (i.e. cuadrática, cubica, etc) entre alguna de las variables X y la variable Y, si es así para cuál o cuáles?

6. Para cada una de las variables cualitativas gráfique un boxplot. Sin reportar las gráficas, pero estas si deben correr en el código, defina qué variables cualitativas podrían ser informativas en la estimación y por qué.

7. Estime los siguientes modelos por MCO:

$$Y = \alpha_0 + \sum_{i=1}^{\max 8} \beta_i x_i + \varepsilon_i \qquad [1]$$

Donde:

$Y_i$: El rating imdb para cada película i

$x_i$: Las variables con mayor correlación seleccionadas en el punto 4. **NOTA:** <u>No deben ser más de 8 variables.</u>

$$Y = \alpha_0 + \sum_{i=1}^{\max 8} \beta_i x_i + \sum_{j=1}^{16} \gamma_j g_j + \varepsilon_i \qquad [2]$$

[3] Esto lo pueden hacer en Matlab con el comando outejoin o en Excel con el comando buscarv, como ustedes lo prefieran.

[4] Utilizando el comando removevars debe eliminar todas las variables categóricas y si aún las tiene en su muestra las variables del PIB de USA, la variable de ganancia en dolares (gross) y el presupuesto (budget) y el año de la película (title year).

Donde:

$Y_i$: El rating imdb para cada película i

$x_i$: Las variables con mayor correlación seleccionadas en el punto 4. **NOTA:** No deben ser más de 8 variables.

$g_i$: Es la variables categóricas que indica el género de la película (i.e. terror, comedia entre otros).

Una vez los hayan estimado, reporten el $R^2$ y el $R^2$ ajustado, e interprételos desde una perspectiva de pronósticos. ¿Cuál modelo es mejor, el modelo 1 o el modelo 2?.¿Incluir la variable categórica de género mejoraría la proyección o lo empeoraría?¿Dada la interpretación del $R^2$ de los modelos (1) y (2) que ustedes seleccionaron, creen que la selección de variables a partir de correlaciones para hacer pronósticos es suficiente?[5]

8. A partir de las estimaciones del numeral 7, usen el modelo [1] y pronostiquen la calificación Imdb para las películas que no lo reportan. Presente sus resultados en forma de gráfico de dispersión con la duración de las películas en el eje horizontal y el rating Imdb en el eje vertical. ¿Son los resultados esperados? ¿Cree que tiene un buen pronóstico del rating?

## 1.2. Estimación de Intervalos y Densidades (30 Puntos)

Asumamos que después de la inspección gráfica realizada previamente ustedes deciden que el mejor modelo es el [2] y deciden realizar unos primeros pronósticos del **rating IMDB**, seleccionando las siguientes variables:

Y: El rating imdb para cada película i

$x_1$: num_critic_for_reviews

$x_2$: duration

$x_3$: director_facebook_likes

$x_4$: num_voted_users

$x_5$: num_user_for_reviews

$x_6$: movie_facebook_likes

$x_7$: Ganancias como % del PIB

$g_i$:genres

Estas variables no son necesariamente las seleccionadas por ustedes, pero el director de pronósticos de la empresa esta convencido que podría dar buenos resultados. Tomando solo los datos que contaban inicialmente con el rating Imdb y asumiendo un valor igual a la **mediana** para cada una de las variables exógenas cuantitativas (i.e. num_critic_for_reviews, duration, director_facebook_likes, num_voted_users, num_user_for_reviews, movie_facebook_likes y

[5] Recuerden que, para el caso de los salarios, visto en la complementaria 2 el $R^2$ más elevado llegaba hasta 0,26.

Ganancias como % del PIB), además suponiendo que el género de la película es animada (i.e. Animation), estiman el pronóstico punto del rating Imdb para una película con estas características y realizan los siguientes ejercicios:

1. Asuman que $\varepsilon_i \sim N(0,\sigma^2)$. Sin hacer simulaciones, calculen un intervalo de confianza de 95% de confianza para su pronóstico. Reporten pronóstico puntual e intervalo inferior y superior de pronóstico.
2. Sigan asumiendo que $\varepsilon_i \sim N(0,\sigma^2)$. Ahora construyan un intervalo de confianza de 95% de confianza para su pronóstico usando $R = 50$, $R = 500$, $R = 10{,}000$. Comparen los intervalos con los del intervalo teórico asumido en 1.
3. Grafiquen las simulaciones del punto 2, incluyendo punto, intervalo y densidad (usen el comando subplot de MATLAB para incluir varios gráficos en uno).
4. Ahora supongan que no conoce la distribución de $\varepsilon_i$, excepto que estas tienen media cero. construyan un intervalo de confianza de 95% para su pronóstico asumiendo R=10.000.
5. Grafiquen las simulaciones del punto 4, incluyendo punto, intervalo y densidad.
6. Ahora, a la estimación realizada en el punto 4 agréguenle incertidumbre en los parámetros[6]. Construyan un intervalo de 95% de confianza para su pronóstico asumiendo R=10.000 tanto para las repeticiones de épsilon como para las simulaciones de los parámetros.
7. Grafiquen las simulaciones del punto 6, incluyendo punto, intervalo y densidad.
8. Finalmente, asuman que $\varepsilon_i \sim N(0,\sigma_i{}^7)$ -errores heteroscedasticos-, pero asuma **total certidumbre** en los parámetros. Construyan un intervalo de 95% de confianza para su pronóstico asumiendo R=10.000 tanto para las simulaciones de épsilon como para las simulaciones de los parámetros.
9. Grafiquen las simulaciones del punto 8, incluyendo punto, intervalo y densidad. ¿Se generó algún cambio con respecto a las estimaciones previas?

## 2. Tendencia, Estacionalidad y Ciclos

### 2.1. Análisis de propiedades modelos AR, MA y ARMA (30 Puntos)

1. Consideren el proceso $x_t = \phi x_{t-1} + \varepsilon_t$, donde $\varepsilon_t$ es ruido blanco y $|\phi| < 1$ . A partir de esta información, demuestren formalmente y expliquen la importancia de tener un numero de datos que tienda a infinito. Comprueben las implicaciones de esto en términos de su media no condicional, su varianza no condicional y su auto covarianza no condicional.

[6] Asuma normalidad de los parámetros
[7] Asuma normalidad de los parámetros

2. Se tiene un proceso estacionario AR (2) igual a:

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \varepsilon_t$$

Adicionalmente se conoce que $\varepsilon_t$ es ruido blanco. A partir de esta información estime la varianza no condicional y las autocorrelaciones hasta el rezago 3 del proceso AR(2) definido previamente.

3. Se tiene un proceso estacionario ARMA (1,1) igual a:

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \varepsilon_t + \beta_1 \varepsilon_{t-1}$$

Donde $\varepsilon_t$ es ruido blanco. Sabemos que la solución de este proceso se puede escribir de la siguiente manera:

$$y_t = \sum_{i=0}^{\infty} c_i \varepsilon_{t-1}$$

a. Demuestren que:

1. $c_0 = 1$
2. $c_1 = \alpha_1 + \beta_1$
3. $c_i = \alpha_1 c_{i-1} + \alpha_2 c_{i-2} \qquad para\ todo\ i \geq 2$

b. Estimen la media, la varianza no condicional y la covarianza no condicional del proceso ARMA(1,1) e indiquen si el proceso es estacionario.
Pista: el punto b desarróllenlo a partir de la solución $MA(\infty)$

## 2.2. Estimación AR, MA y ARMA (25 Puntos)

1. En el archivo de excel denominado ISE se encuentra la serie mensual del índice de seguimiento a la economía (ISE) de Colombia desde enero de 2005 hasta noviembre de 2021. Por el momento, suponga que esta variable sigue un proceso AR(1): $y_t = c + \vartheta y_{t-1} + \varepsilon_t$ donde $\varepsilon_t$ es un proceso ruido blanco.

1.1. Para la serie del ISE realicen los siguientes procedimientos:

a) Conviertan la serie del ISE a logaritmo natural, estime su primera diferencia y describa sus principales características (asegúrense de multiplicar por 100 este valor para interpretar todos los resultados como variación porcentual mensual)[8].

b) Estimen y grafiquen las 20 primeras autocorrelaciones de la primera diferencia logarítmica del ISE e Interpreten los resultados.

c) Estimen y grafiquen las 20 primeras autocorrelaciones parciales de la primera diferencia logarítmica del ISE e interpreten los resultados.

[8] Recuerde que ante variaciones muy pequeñas de una serie la diferencia del logaritmo natural es aproximadamente igual a la variación porcentual.

1.2. Debido a los cierres económicos sectoriales llevados a cabo en 2020, como consecuencia de la pandemia, las series del ISE presentan datos muy inusuales. Esto podría tener repercusiones no deseadas en las proyecciones del ISE que desean hacer, por lo que ustedes llevan a cabo los siguientes procedimientos:

a. Detectan y remplazan los outliers de la primera diferencia del logaritmo natural del ISE.

**Para este punto les recomiendo seguir este procedimiento:**

Utilicen el siguiente comando de MATLAB para hacerlo:

[B,TF,lower,upper,center] = filloutliers(A,fillmethod,findmethod)[9]

**Donde filloutlier es la función de detección y sustitución. Para el caso puntual de este taller deben elegir las siguientes opciones de este comando:**

A, debe ser la primera diferencia del logaritmo natural del ISE; fillmethod, es el método que usaran para remplazar el outlier detectado, en este caso deben seleccionar **'CLIP'** que remplaza los valores de los outliers (positivos y negativos) por los umbrales definidos con el método de detección seleccionado; y findmethod es el método de búsqueda de los outliers, que para este caso sería **'median'** el cual utiliza la desviación absoluta de la mediana como método de detección.

A partir de esta función obtendrán los resultados que se presentan al lado izquierdo de la igualdad del comando, que son:

B: Las series de la primera diferencia del logaritmo del ISE con los datos de los outliers remplazados.

TF: Una variable dicótoma que es igual a 1 si el dato es un outlier y cero de lo contrario

lower: El umbral mínimo detectado.

Upper: El umbral máximo detectado.

Center: Para este caso el valor de la mediana.

d) Para la serie de la primera diferencia del logaritmo del ISE, con los datos de los outliers remplazados, estimen y grafiquen las 20 primeras autocorrelaciones. ¿Qué diferencia tiene esta serie ajustada por outliers con la serie original?

e) Para la serie de la primera diferencia del logaritmo del ISE, con los datos de los outliers remplazados, estimen y grafiquen las 20 primeras autocorrelaciones parciales. ¿Qué diferencia tiene esta serie ajustada por outliers con la serie original?

f) Estimen los parámetros $c$ y $\vartheta$ por MCO, tomando como variable endógena la primera diferencia del logaritmo natural del ISE. Se deben estimar los datos tanto para la variable

[9] Los detalles de este comando se pueden ver acá: https://la.mathworks.com/help/matlab/ref/filloutliers.html

original como para la variable ajustada por outliers. ¿qué diferencias observan?¿incluir los datos de la pandemia afecta la estimación?

g) Ahora suponga que el proceso sigue un proceso AR(2). Estimen los parámetros por medio de MCO (la constante y los dos coeficientes AR). Igual que con el punto anterior se debe estimar tanto para la variable original como para la variable ajustada por outliers. ¿qué diferencias observa? ¿incluir los datos de la pandemia afecta la estimación?