

Taller I
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 validez 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 mundial¹. 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 son²:

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 película
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 críticas 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 película: '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 guíe 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), α_0 es la constante de la estimación, β_i el parámetro que define cómo cada variable x_i afecta la proyección de tener un buen o mal rating y ε_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.

El pronóstico óptimo es aquel que genera la menor pérdida condicional. Para el caso de la ecuación que se mostró previamente, si se estima los parámetros con mínimos cuadrados ordinarios y estos cumplen con las condiciones ideales (FIC):

1. El modelo representa el proceso generador de datos.

- a. El modelo tiene una relación lineal entre la variable endógena y las exógenas.
 - b. Los coeficientes son fijos a través de la muestra.
 - c. Los errores no están correlacionados entre si.
2. X es fijo en muestras repetidas.
 3. X no tiene problemas de multicolinealidad.
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).

1.1.3. Convirtiendo los datos en categóricos:

% Para la tabla con rating

```
Cr.color = categorical(Cr.color);
Cr.language = categorical(Cr.language);
Cr.country = categorical(Cr.country);
Cr.content_rating=categorical(Cr.content_rating);
Cr.director_name=categorical(Cr.director_name);
Cr.actor_1_name=categorical(Cr.actor_1_name);
Cr.actor_2_name=categorical(Cr.actor_2_name);
Cr.actor_3_name=categorical(Cr.actor_3_name);
Cr.movie_title=categorical(Cr.movie_title);
Cr.genres=categorical(Cr.genres);
```

% Para la tabla sin rating

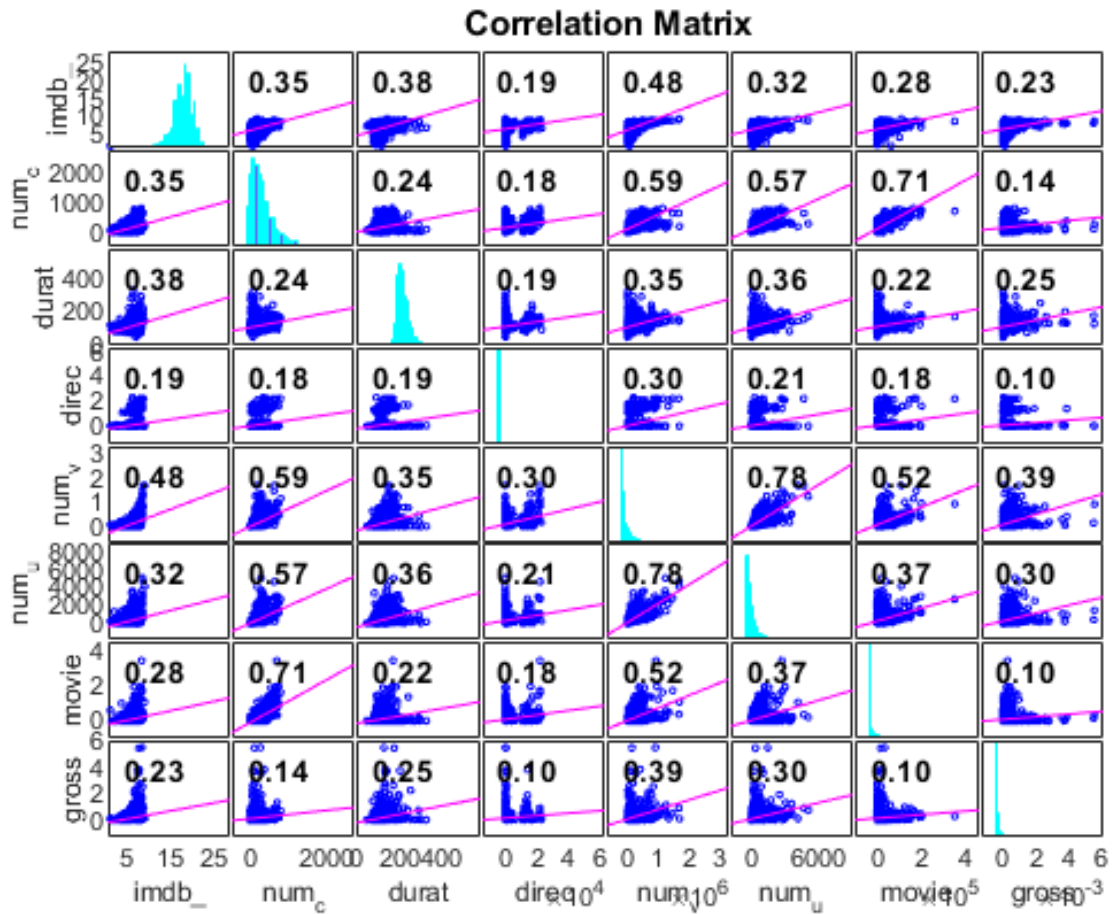
```
Sr.color = categorical(Sr.color);
Sr.language = categorical(Sr.language);
Sr.country = categorical(Sr.country);
Sr.color = categorical(Sr.color);
Sr.content_rating=categorical(Sr.content_rating);
Sr.director_name=categorical(Sr.director_name);
Sr.actor_1_name=categorical(Sr.actor_1_name);
Sr.actor_2_name=categorical(Sr.actor_2_name);
Sr.actor_3_name=categorical(Sr.actor_3_name);
Sr.movie_title=categorical(Sr.movie_title);
Sr.genres=categorical(Sr.genres);
```

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

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 pregunta⁴:
- ¿Cuáles son las variables que tienen una correlación mayor o igual a $|0,15|$ con el rating IDB?
 - Reporte el corrplot de las variables que tienen una correlación superior a $|0,15|$ con IDB y el IDB.

```
5. Cr1=removevars(Cr,{ 'color', 'language', 'country', 'content_rating',...  
    'director_name', 'actor_1_name', 'actor_2_name', 'actor_3_name',...  
    'movie_title', 'genres', 'GDP', 'gross', 'budget', 'title_year'});  
  
%1.1.4 a  
% LAS variables  
  
%1.1.4 b Las variables que tienen una correlación superior a 0,15 en valor  
%absoluta con el Imdb son: número de criticas, duración, likes del director  
%en Facebook, número de votos de los usuarios, número de usuarios que  
%hicieron reviews, facebook likes de la película, ganancias como % del PIB  
  
Cr2=removevars(Cr1,{ 'actor_3_facebook_likes', 'actor_1_facebook_likes'...  
    'cast_total_facebook_likes', 'facenumber_in_poster', 'actor_2_facebook_likes'...  
    , 'aspect_ratio', 'budget2', 'grossOverBudget'});  
  
% el correlograma sería el siguiente:  
  
figure(1)  
corrplot(Cr2)
```

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).



6. 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?

```
% En general todas podrían tener una relación concava con el idb score. Sin
% embargo, la duración muestra una clara relación concava con el idb score
% que se debería tener en cuenta. Despues de los 250 minutos de duración se
% ve una clara disminucion en la calificación del apelicula. Esto no es
% claro para el resto de variables, al menos a primera vista.
```

7. 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.

8. Estime los siguientes modelos por MCO:

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

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.

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

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 interprétenlos 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

1.1.7. Estimación modelos de regresión y R2:

```
Cr.d2=Cr.duration.^2;
Modelo1=fitlm(Cr,'imdb_score~num_critic_for_reviews+duration+director_facebook_likes+num_vo
ted_users+num_user_for_reviews+movie_facebook_likes+gross2')

Modelo2=fitlm(Cr,'imdb_score~num_critic_for_reviews+duration+director_facebook_likes+num_vo
ted_users+num_user_for_reviews+movie_facebook_likes+gross2+genres')

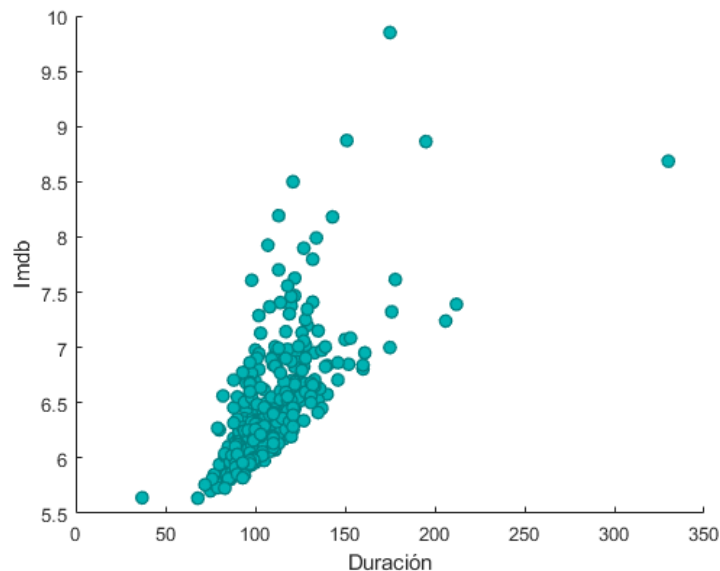
% el modelo con la variable DUMMY y con el intercepto tendría un mejor
% desempeño, su R2 muestra que es capaz de explicar el 38% de la
% variabilidad del rating, mientras que el modelo sin DUMMY cambia por
% alcanzaría a explicar tan solo el 30%.
```

9. 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?

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

Estimación ratings no presentados:

```
betahat=Modelo1.Coefficients.Estimate;  
  
Sr.imdb_score=betahat(1,1)+Sr.num_critic_for_reviews.*betahat(2,1)+Sr.duration.*betahat(3,1)  
(+Sr.director_facebook_likes.*betahat(4,1)...  
  
+Sr.num_voted_users.*betahat(5,1)+Sr.num_user_for_reviews.*betahat(6,1)+Sr.movie_facebook_l  
ikes.*betahat(7,1)+Sr.gross2.*betahat(8,1);  
  
figure(13)  
scatter(Sr.duration,Sr.imdb_score,40,'MarkerEdgeColor',[0 .5 .5],...  
        'MarkerFaceColor',[0 .7 .7],...  
        'Linewidth',1)  
xlabel('Duración')  
ylabel('Imdb')
```



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 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 $\epsilon_i \sim N(0, \sigma^2)$. 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.

Intervalo de confianza sin simulaciones.

%estimemos la desviación y media de los errores de pronóstico:

```
S=std(Modelo2.Residuals.Raw);
mu=mean(Modelo2.Residuals.Raw);
```

% Mediana de las variables exogenas

```
MNCR=median(Cr.num_critic_for_reviews);
Mduration=median(Cr.duration);
MDFL=median(Cr.director_facebook_likes);
MNVU=median(Cr.num_voted_users);
MNUFR=median(Cr.num_user_for_reviews);
MMFL=median(Cr.movie_facebook_likes);
MG2=median(Cr.gross2);
```

% Estimación del rating medio de la ecuacion [2] del taller y características establecidas:

```
betahat=Modelo2.Coefficients.Estimate;
```

```
MuR=betahat(1,1)+MNCR*betahat(2,1)+Mduration*betahat(3,1)+MDFL*betahat(4,1)...
+MNVU*betahat(5,1)+MNUFR*betahat(6,1)+MMFL*betahat(7,1)+betahat(9,1)+MG2*betahat(24,1);
```

%intervalo 95%:

```
MuRU=MuR+norminv(0.975)*S;
MuRD=MuR-norminv(0.975)*S;
```

```
% reportando los datos en una tabla:
format long g
MODELLOS = {'Pronóstico Rating Promedio'};

Pronostico=MUR;
Intervalo_Superior=MURU;
Intervalo_Inferior=MURD;
Intervalo = table(Pronostico,Intervalo_Inferior,Intervalo_Superior, 'RowNames',MODELLOS)
```

	Pronostico	Intervalo_Inferior	Intervalo_Superior
Pronóstico Salario medio	6.67	5.06	8.29

2. Sigamos 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. % R=50:

RNormal50=normrnd(0,s,[1,50]);
MuR50_Normal=(RNormal50'+MuR);

% calculamos los intervalos de predicción a 95% de confianza:
format long g
pw50normal = prctile(MuR50_Normal, [2.5 97.5]);

% R=500:

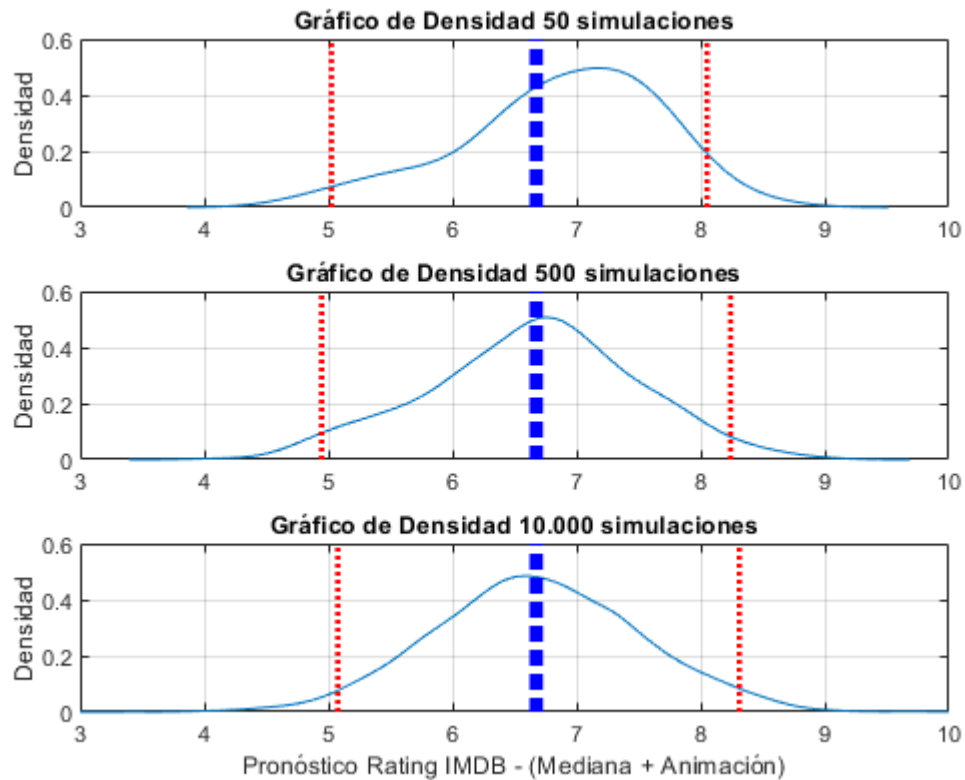
RNormal500=normrnd(0,s,[1,500]);
MuR500_Normal=(RNormal500'+MuR);

% calculamos los intervalos de predicción a 95% de confianza:
format long g
pw500normal = prctile(MuR500_Normal, [2.5 97.5]);

% R=10000:

RNormal10000=normrnd(0,s,[1,10000]);
MuR10000_Normal=(RNormal10000'+MuR);

% calculamos los intervalos de predicción a 95% de confianza:
format long g
pw10000normal = prctile(MuR10000_Normal, [2.5 97.5])
```



4. Ahora supongan que no conoce la distribución de ε_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.

1.3. 1.2.4 y 1.2.5 simulaciones no paramétrica:

```
% a. generamos 10.000 numeros aleatorios que se obtienen directamente de los errores
estimados:

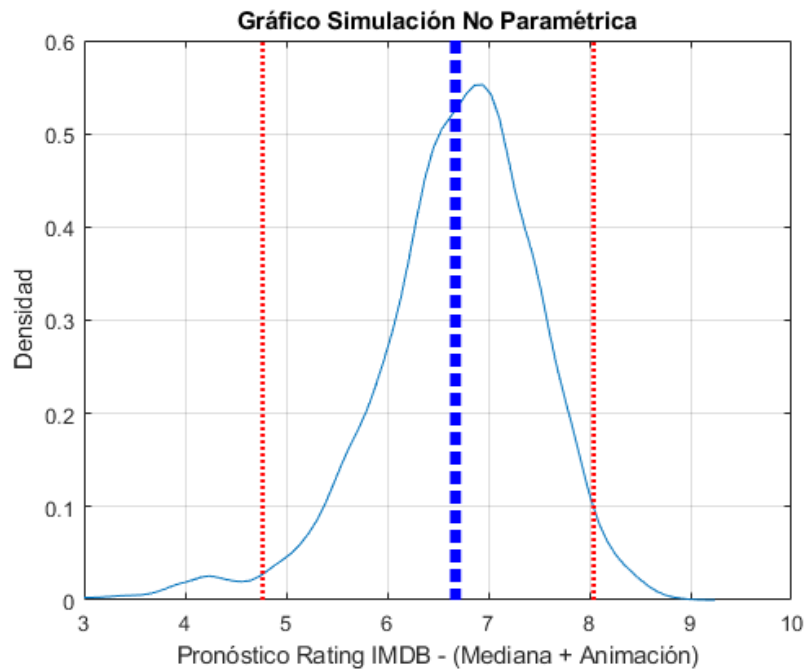
RNoP = datasample(Modelo2.Residuals.Raw,10000);

%b. De la estimación del punto con distribución normal tomamos la estimación y generamos
10000 pronosticos con los
%errores no parametricos:

MuR_NoP=(RNoP'+MuR);

% c. calculamos los intervalos de predicción a 95% de confianza:
```

```
pw1NoP = prctile(MuR_NoP, [2.5 97.5])
```



6. Ahora, a la estimación realizada en el punto 4 agréguele incertidumbre en los parámetros⁶. Construyan un intervalo de 95% de confianza para su pronóstico asumiendo $R=10.000$ tanto para las repeticiones de ϵ como para las simulaciones de los parámetros.

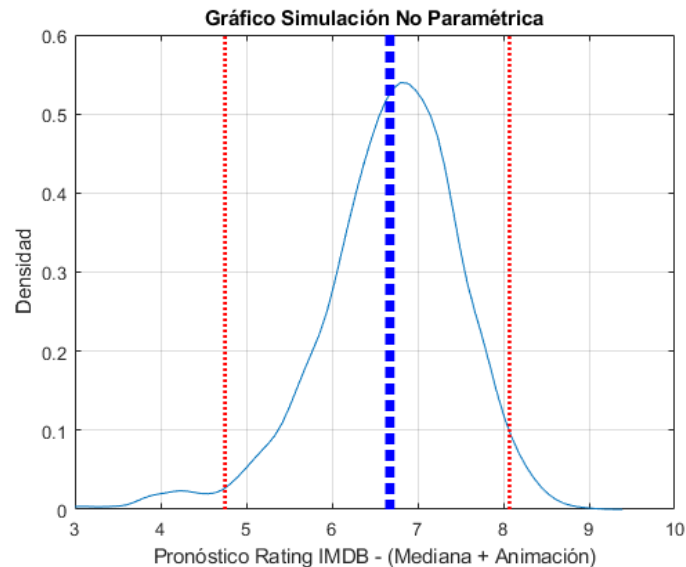
7. Grafiquen las simulaciones del punto 6, incluyendo punto, intervalo y densidad.

simulaciones no paramétrica en errores y asumiendo normalidad en los parametros.

```
% Generando los parametros;
betasigma=Modelo2.CoefficientCovariance;
RBetas = mvnrnd(betahat,betasigma,10000);
% simulaciones:
Salario_Betas=[];
for i=1:10000
    MuR_Betas(i)=RBetas(i,1)+MNCR*RBetas(i,2)+Mduration*RBetas(i,3)+MDFL*RBetas(i,4)...
    +MNVU*RBetas(i,5)+MNUFR*RBetas(i,6)+MMFL*RBetas(i,7)+RBetas(i,9)+MG2*RBetas(i,24)+RNoP(i,1)
;
end
% d. calculamos los intervalos de predicción a 95% de confianza:
```

⁶ Asuma normalidad de los parámetros

```
pw1betas = prctile(MuR_Betas, [2.5 97.5])
```



8. Finalmente, asuman que $\varepsilon_i \sim N(0, \sigma_i)$ -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 ε 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?

simulaciones asumiendo heteroscedasticidad.

```
% ya tenemos la estimación por MCO. hagamos la estimacion del cuadrado de
% los errores de estimacion.

Cr.Residuales2=Modelo2.Residuals.Raw.^2;

% estimando la regresion contra los errores al cuadrado;

Modelo3=fitlm(Cr,'Residuales2~num_critic_for_reviews+duration+director_facebook_likes+num_v
oted_users+num_user_for_reviews+movie_facebook_likes+gross2+genres', 'RobustOpts','on')

% organizando las matrices para la estimación por minimos cuadrados
% generalizados:

Sigma2=diag(Modelo3.Fitted);
SIGMA=inv(Sigma2);
```

```

Y=Cr.imdb_score;
unos=ones(length(Y),1);

% para genero es mejor crear una variable categorica en un vector:

GenresD=dummyvar(Cr.genres);
GenresD=GenresD(:,2:17);

X=[unos Cr.num_critic_for_reviews Cr.duration Cr.director_facebook_likes
Cr.num_voted_users...
    Cr.num_user_for_reviews Cr.movie_facebook_likes GenresD Cr.gross2];
HBetas=inv((X'*SIGMA*X))*X'*SIGMA*Y;

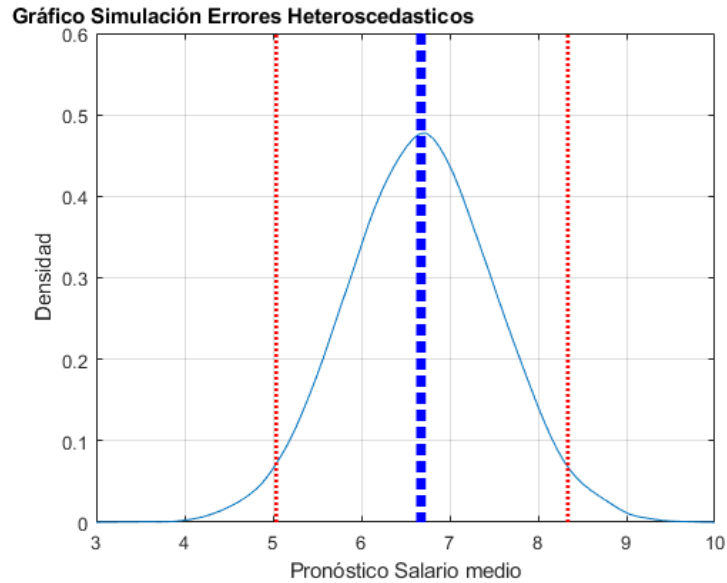
% %10000 simulaciones de los errores
Y_HAT=HBetas'*X';
Error=Y-Y_HAT';
Ss=std(Error);
Ms=mean(Error);
RNormalH=normrnd(0,Ss,[1,10000]);

MuR_H=(RNormalH'+MuR);

% d. calculamos los intervalos de prediccion a 95% de confianza:

pw1betasH = prctile(MuR_H, [2.5 97.5])

```



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 ε_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.

Supongamos que el proceso inicia en el periodo cero, por lo tanto x_0 es una condición inicial determinística. Por tanto, utilizando sustitución recursiva tenemos que:

$$x_t = \phi(\phi x_{t-2} + \varepsilon_{t-1}) + \varepsilon_t$$

$$x_t = \phi^2 x_{t-2} + \phi \varepsilon_{t-1} + \varepsilon_t$$

$$x_t = \phi^2(\phi x_{t-3} + \varepsilon_{t-2}) + \phi \varepsilon_{t-1} + \varepsilon_t$$

$$x_t = \phi^3 x_{t-3} + \phi^2 \varepsilon_{t-2} + \phi \varepsilon_{t-1} + \varepsilon_t$$

⋮

$$x_t = \phi^t x_0 + \sum_{i=0}^{t-1} \phi^i \varepsilon_{t-i} \quad [1]$$

Por tanto, el valor esperado de [1] sería:

$$E(x_t) = \phi^t x_0 \quad [2]$$

Si, se envía s periodos hacia adelante la ecuación [2]:

$$E(x_{t+s}) = \phi^{t+s} x_0 \quad [3]$$

Tendríamos que el valor esperado de la ecuación [2] y [3] no es igual, pues este depende del tiempo que transcurre. Entonces, estas medias tendrán a ser iguales en tanto $t \rightarrow \infty$ y $|\phi| < 1$. En este caso el valor esperado descrito en [2] sería:

$$E(x_t) = 0 = \mu \quad [4]$$

Con respecto a la varianza se tiene que:

$$E(x_t - \mu)^2 = E \left[\left(\sum_{i=0}^{t-1} \phi^i \varepsilon_{t-i} \right) \left(\sum_{i=0}^{t-1} \phi^i \varepsilon_{t-i} \right) \right]^2$$

$$E(x_t - \mu)^2 = E[(\varepsilon_t + \phi \varepsilon_{t-1} + \phi^2 \varepsilon_{t-2} + \dots)(\varepsilon_t + \phi \varepsilon_{t-1} + \phi^2 \varepsilon_{t-2} + \dots)]^2$$

$$E(x_t - \mu)^2 = \sigma^2(1 + (\phi)^2 + (\phi)^4 + \dots)$$

$$E(x_t - \mu)^2 = \frac{\sigma^2}{1-\phi^2} \quad [5]$$

Con respecto a la auto covarianza se tiene que:

$$E(x_t - \mu)(x_{t-s} - \mu) = E[(\varepsilon_t + \phi \varepsilon_{t-1} + \phi^2 \varepsilon_{t-2} + \dots + \phi^s \varepsilon_{t-s} + \phi^{s+1} \varepsilon_{t-s-1} + \dots)(\varepsilon_{t-s} + \phi \varepsilon_{t-s-1} + \phi^2 \varepsilon_{t-s-2} + \dots)]$$

$$E(x_t - \mu)(x_{t-s} - \mu) = \sigma^2(\phi)^s(1 + (\phi)^2 + (\phi)^4 + \dots)$$

$$E(x_t - \mu)(x_{t-s} - \mu) = \frac{\sigma^2(\phi)^s}{1-\phi^2} \quad [6]$$

En resumidas cuentas, si $t \rightarrow \infty$ y $|\phi| < 1$ en la ecuación [1], la secuencia de x_t será estacionaria.

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 ε_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.

Para responder a esta pregunta usemos la metodología de Yule-Walker definida en el libro de Diebold.

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

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

$$\vdots$$

$$E(y_t y_{t+s}) = \alpha_1 E(y_{t-1} y_{t+s}) + \alpha_2 E(y_{t-2} y_{t+s}) + E(\varepsilon_t y_{t+s}) \quad [7]$$

Por definición las auto covarianzas de un proceso estacionario son tales que $E(y_t y_{t+s}) = E(y_{t+s} y_t) = \gamma_s$. También sabemos que $E(\varepsilon_t y_t) = \sigma^2$ y que $E(\varepsilon_t y_{t-s}) = 0$. Entonces podemos usar la ecuación [7] y obtenemos que:

$$\gamma_0 = \alpha_1 \gamma_1 + \alpha_2 \gamma_2 + \sigma^2 \quad [8]$$

$$\gamma_1 = \alpha_1 \gamma_0 + \alpha_2 \gamma_1 \quad [9]$$

$$\gamma_s = \alpha_1 \gamma_{s-1} + \alpha_2 \gamma_{s-2} \quad [10]$$

Dividiendo [9] y [10] por [8] se obtiene:

$$\rho_1 = \alpha_1 \rho_0 + \alpha_2 \rho_1 \quad [11]$$

$$\rho_s = \alpha_1 \rho_{s-1} + \alpha_2 \rho_{s-2} \quad [12]$$

Se sabe que $\rho_0 = 1$ entonces las autocorrelaciones hasta el tercer rezago serían:

$$\rho_1 = \frac{\alpha_1}{1-\alpha_2} \quad [13]$$

$$\rho_2 = \frac{\alpha_1^2}{1-\alpha_2} + \alpha_2 \quad [14]$$

$$\rho_3 = \alpha_1 \left[\frac{\alpha_1^2}{1-\alpha_2} \right] + \alpha_2 \left[\frac{\alpha_1}{1-\alpha_2} \right] \quad [15]$$

Para estimar la varianza, de la ecuación [8] sabemos que:

$$\gamma_0 [1 - \alpha_1 \rho_1 - \alpha_2 \rho_2] = \sigma^2$$

Sustituyendo [13] y [14] en la anterior expresión tenemos que la varianza del proceso es:

$$\begin{aligned} \gamma_0 \left[1 - \alpha_1 \left(\frac{\alpha_1}{1-\alpha_2} \right) - \alpha_2 \left(\frac{\alpha_1^2}{1-\alpha_2} + \alpha_2 \right) \right] &= \sigma^2 \\ \gamma_0 \left[\frac{1 - \alpha_2 - \alpha_1^2 - \alpha_2 \alpha_1^2 - \alpha_2^2 (1 - \alpha_2)}{1 - \alpha_2} \right] &= \sigma^2 \\ \gamma_0 &= \sigma^2 \left[\frac{1 - \alpha_2}{(1 - \alpha_2)(1 - \alpha_2^2) - \alpha_1^2 (1 + \alpha_2)} \right] \end{aligned} \quad [16]$$

Se tiene un proceso estacionario ARMA (2,1) igual a:

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

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

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

a. Demuestren que:

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

Dada la información del enunciado, para que la solución de:

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

Exista. Se debe satisfacer que:

$$\begin{aligned} c_0 \varepsilon_t + c_1 \varepsilon_{t-1} + c_2 \varepsilon_{t-2} + c_3 \varepsilon_{t-3} + \dots &= \\ \alpha_1 (c_0 \varepsilon_{t-1} + c_1 \varepsilon_{t-2} + c_2 \varepsilon_{t-3} + c_3 \varepsilon_{t-4} + \dots) &+ \\ \alpha_2 (c_0 \varepsilon_{t-2} + c_1 \varepsilon_{t-3} + c_2 \varepsilon_{t-4} + c_3 \varepsilon_{t-5} + \dots) &+ \\ \varepsilon_t + \beta_1 \varepsilon_{t-1} & \end{aligned}$$

Para que los coeficientes del lado izquierdo y el lado derecho asociados a los términos $\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \dots$ sean iguales es necesario:

1. $c_0 \varepsilon_t = \varepsilon_t \Rightarrow c_0 = 1$
2. $c_1 \varepsilon_{t-1} = \alpha_1 c_0 \varepsilon_{t-1} + \beta_1 \varepsilon_{t-1} \Rightarrow c_1 = \alpha_1 + \beta_1$
3. $c_2 \varepsilon_{t-2} = \alpha_1 c_1 \varepsilon_{t-2} + \alpha_2 c_0 \varepsilon_{t-2} \Rightarrow c_2 = \alpha_1 c_1 + \alpha_2 c_0$
4. $c_3 \varepsilon_{t-3} = \alpha_1 c_2 \varepsilon_{t-3} + \alpha_2 c_1 \varepsilon_{t-3} \Rightarrow c_3 = \alpha_1 c_2 + \alpha_2 c_1$

Por tanto, para todo $i > 2$:

$$c_i = \alpha_1 c_{i-1} + \alpha_2 c_{i-2}$$

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 $MA(\infty)$

Media no condicional:

$$E(y_t) = E\left(\sum_{i=0}^{\infty} c_i \varepsilon_{t-i}\right) = 0$$

Varianza no condicional:

$$\begin{aligned} var(y_t) &= E[(c_0 \varepsilon_t + c_1 \varepsilon_{t-1} + c_2 \varepsilon_{t-2} + c_3 \varepsilon_{t-3} + \dots)^2] \\ var(y_t) &= \sigma^2 \sum_{i=0}^{\infty} c_i^2 \end{aligned}$$

Covarianza no condicional:

Sabemos que para este proceso $c_0 = 1$; entonces:

$$cov(y_t, y_{t-1}) = E[(\varepsilon_t + c_1 \varepsilon_{t-1} + c_2 \varepsilon_{t-2} + \dots)(\varepsilon_{t-1} + c_1 \varepsilon_{t-2} + c_2 \varepsilon_{t-3} + c_3 \varepsilon_{t-4} + \dots)]$$

$$cov(y_t, y_{t-1}) = \sigma^2(c_1 + c_2 c_1 + c_3 c_2 + \dots)$$

$$cov(y_t, y_{t-2}) = E[(\varepsilon_t + c_1 \varepsilon_{t-1} + c_2 \varepsilon_{t-2} + \dots)(\varepsilon_{t-2} + c_1 \varepsilon_{t-3} + c_2 \varepsilon_{t-4} + c_3 \varepsilon_{t-5} + \dots)]$$

$$cov(y_t, y_{t-2}) = \sigma^2(c_2 + c_3 c_1 + c_4 c_2 + \dots)$$

Por lo tanto:

$$cov(y_t, y_{t-s}) = \sigma^2(c_s + c_{s+1} c_1 + c_{s+2} c_2 + \dots)$$

De esta manera tenemos una media no condicional constante y las varianzas y covarianzas no dependen del tiempo, el proceso es estacionario en media y en covarianza.

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 ε_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)⁷.

Importando los datos

```
% Importando los datos como un conjunto de datos:
```

```
ds = readtable("ISE.xlsx", "UseExcel", false);  
ds=table2timetable(ds);
```

```
% generando la variación mensual del ISE:
```

```
ds.LISE=log(ds.ISE);  
DISE=diff(ds.LISE)*100;  
ds.DISE=[NaN;DISE];  
clear DISE
```

- 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.

2.3. Dibuje las series, estime y grafique ACF y PACF

```
summary (ds)
```

```
figure (1)  
subplot(2,1,1)  
plot(ds.Fecha, ds.DISE)  
title('Crecimiento mensual del ISE')  
xlabel('Fecha')  
ylabel('%')
```

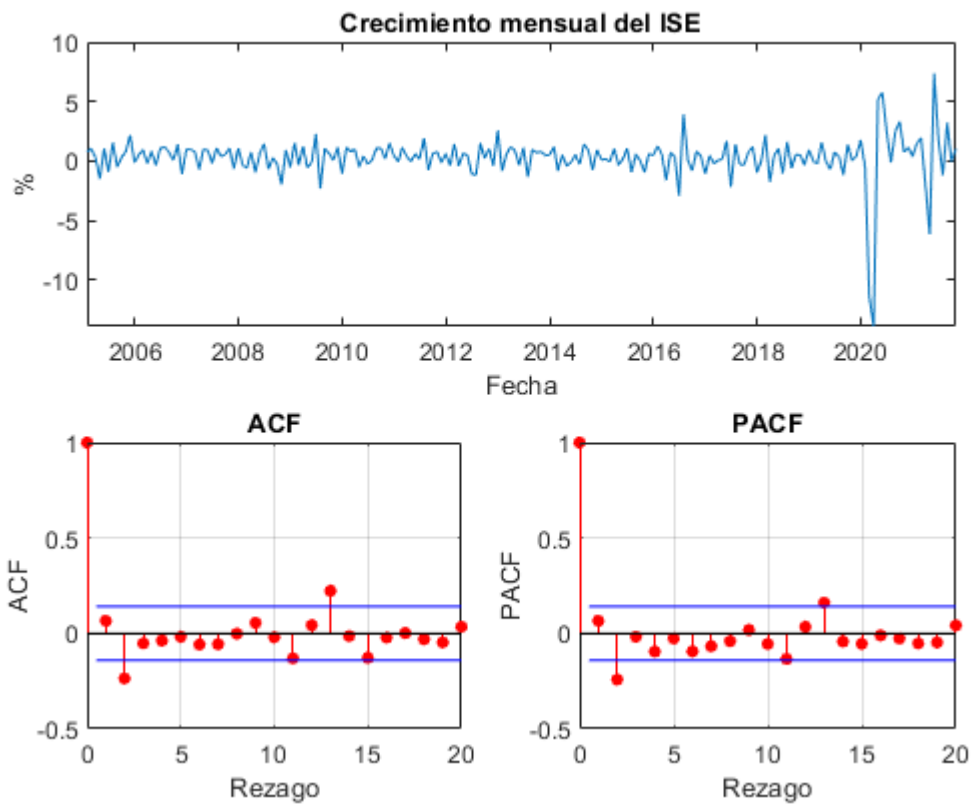
⁷ Recuerde que ante variaciones muy pequeñas de una serie la diferencia del logaritmo natural es aproximadamente igual a la variación porcentual.

```

subplot(2,2,3)
autocorr(ds.DISE)
ylabel('ACF')
xlabel('Rezago')
title('ACF')

subplot(2,2,4)
parcorr(ds.DISE)
ylabel('PACF')
xlabel('Rezago')
title('PACF')

```



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)8
```

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 reemplaza 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?

2.4. elimine los outliers y observe de nuevo el autocorrelograma

```
[ds.DISEO,ds.TF,lower,upper,center] = filloutliers(ds.DISE,'clip','median');
```

```
% 1.a, 1.b y 1.c. Dibuje las series, estime y grafique ACF y PACF
```

```
figure (3)
subplot(2,1,1)
plot(ds.Fecha, ds.DISEO)
title('Crecimiento Mensual del ISE sin Outliers')
xlabel('Fecha')
```

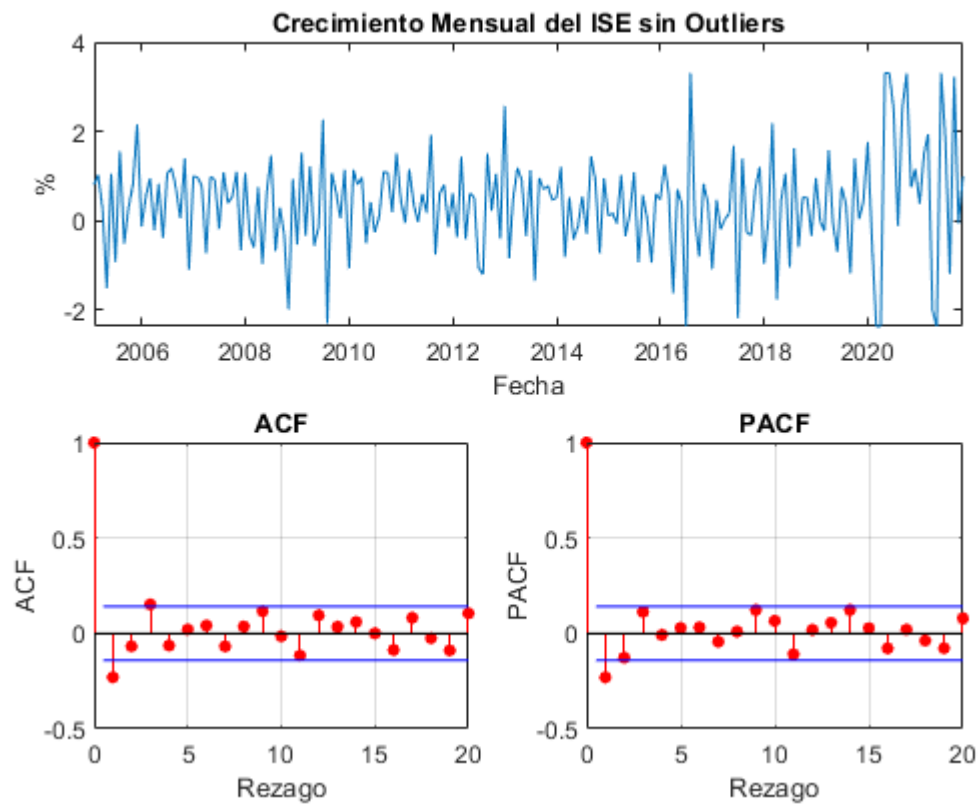
⁸ Los detalles de este comando se pueden ver acá: <https://la.mathworks.com/help/matlab/ref/filloutliers.html>

```

ylabel('%')
subplot(2,2,3)
autocorr(ds.DISEO)
ylabel('ACF')
xlabel('Rezago')
title('ACF')

subplot(2,2,4)
parcorr(ds.DISEO)
ylabel('PACF')
xlabel('Rezago')
title('PACF')

```



- f) Estimen los parámetros c y θ por MCO, tomando como variable endógena la primera diferencia del logaritmo natural del ISE. Se deben estimar los datos tanto para la variable original como para la variable ajustada por outliers. ¿qué diferencias observan? ¿incluir los datos de la pandemia afecta la estimación?

Estime los parámetros AR(1) por MCO

```
% variable original:

% llamamos una función de MCO construida:

[AR10,stAR10]=OLS(ds.DISE(2:end-13),1)

% obtenemos que la constante es 0.2785 y el parametro AR(1) es 0.0707, sin
% ser este estadísticamente significativo..

% variable ajustada por outliers:
% llamamos una función de MCO construida:

[AR1,stAR1]=OLS(ds.DISEO(2:end-13),1)

% obtenemos que la constante es 0.4632 y el parametro AR(1) es -0.2289, los
% dos estadísticamente significativos.
```

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?

2.5. Estime los parametros AR(2) por MCO

```
% variable original

% llamamos una función de MCO construida:

[AR20,stAR20]=OLS(ds.DISE(2:end-13),2)

% obtenemos que los parametros son: constante 0.34, AR(1) 0.0812 y
% AR(2) -0.2401.

% variable ajustada por Outliers:

% llamamos una función de MCO construida:

[AR2,stAR2]=OLS(ds.DISEO(2:end-13),2)
```

```
% obtenemos que los parametros son: constante 0.5183, AR(1) -0.2589 y  
% AR(2) -0.1259.
```