Code

APLICACIONES DEL MODELO DE REGRESIÓN LINEAL CLÁSICO EN R: UN ANÁLISIS TEÓRICO Y APLICADO

ISADORE NABI

09/10/2021

A. APLICACIONES EN BIOESTADÍSTICA

Se realizó un estudio para analizar la velocidad de nado de las personas mayores de 18 años que son miembros regulares de un equipo de natación, y se tomaron en cuenta algunas variables que pueden estar relacionadas con esta velocidad. Se hizo una prueba a los participantes y se tomó el tiempo que duraban en nadar 50m. Entonces como medida de la velocidad de nado se tiene el tiempo (en segundos) el cual se puede transformar a la velocidad dividiendo la distancia entre el tiempo. Esta variable se llama **veloc**. Como variables predictoras se tienen las siguientes:

- **edad**: la edad en años cumplidos.
- **sexo**: el sexo codificado como 0 (mujeres) y 1 (hombres).
- **imc**: el índice de masa corporal se calcula dividiendo el peso en kilogramos entre la altura al cuadrado (en metros), lo cual da una medida en kg/m^2.
- **pierna**: la longitud promedio de ambas piernas (en centímetros).
- **brazo**: la longitud promedio de ambos brazos (en centímetros).

a. Se debe comenzar por leer el archivo (nativo de R) de datos mediante la siguiente sintaxis:

Hide

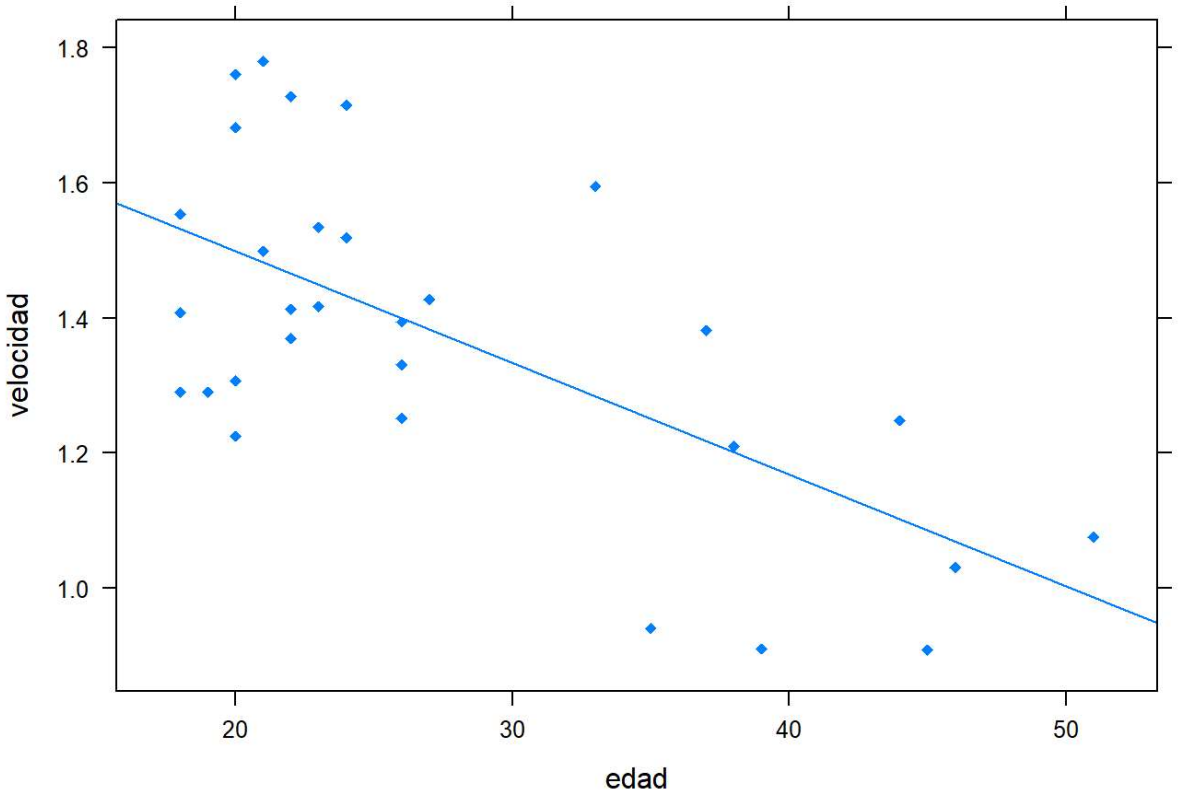
```
setwd("C:/Users/User/Desktop/Carpeta de Estudio/Mis Códigos en R")
load("velo.Rdata")
attach(base)
```

Puede analizarse gráficamente la relación entre la respuesta promedio (la esperanza matemática de la variable dependiente) y cada uno de los predictores numéricos (los valores de las variables independientes).

Puede usarse el paquete “lattice”:

Hide

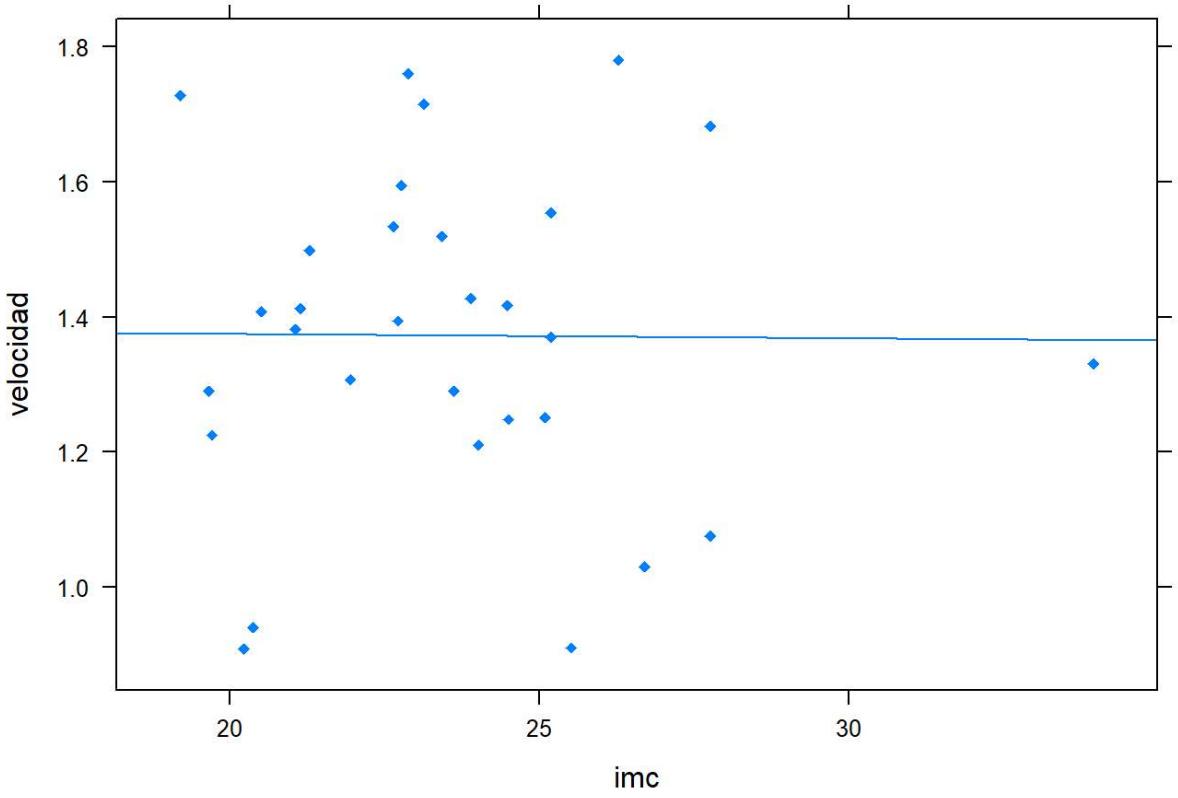
```
library(lattice)
xyplot(veloc~edad,type=c("p","r"),pch=18,ylab="velocidad")
```



"pch" es la forma que tendrán los puntos de datos a graficar. Se escogió 18 por cuestiones de gusto, pero pudo ser, por ejemplo, 5 (que los grafica como cuadrados), 35 (que graficar numerales) u otro.

Hide

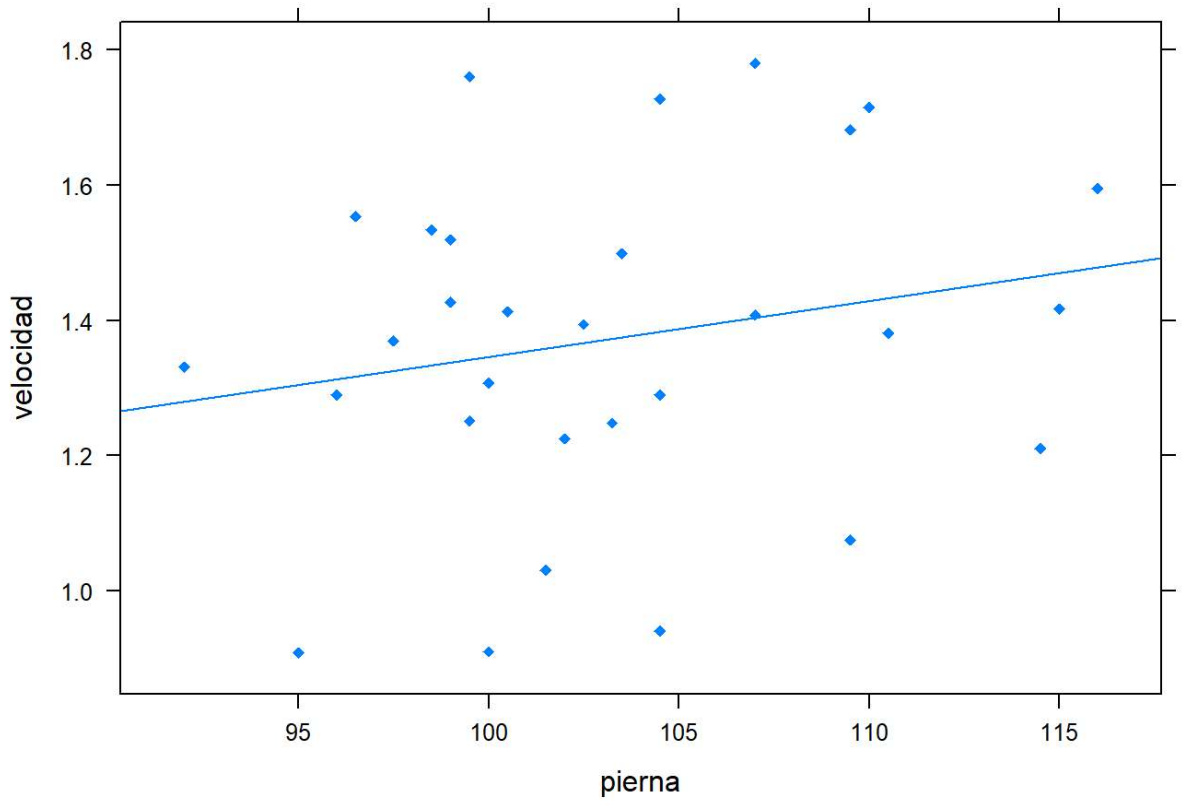
```
xyplot(veloc~imc,type=c("p","r"),pch=18,ylab="velocidad")
```



Se escribe “type” porque esta función de R sirve para determinar el tipo o modo de almacenamiento (a nivel de la estructura interna de R) de un objeto. Si desea una recta de regresión junto con su diagrama de dispersión, puede usarse la sintaxis “type”. La lista dada por la sintaxis c (“p”, “r”) indica a la sintaxis principal xyplot () que se requiere graficar tanto los puntos (“p”) como una recta de regresión (“r”); véase https://homerhanumat.github.io/tigerstats/xyplot.html (https://homerhanumat.github.io/tigerstats/xyplot.html).

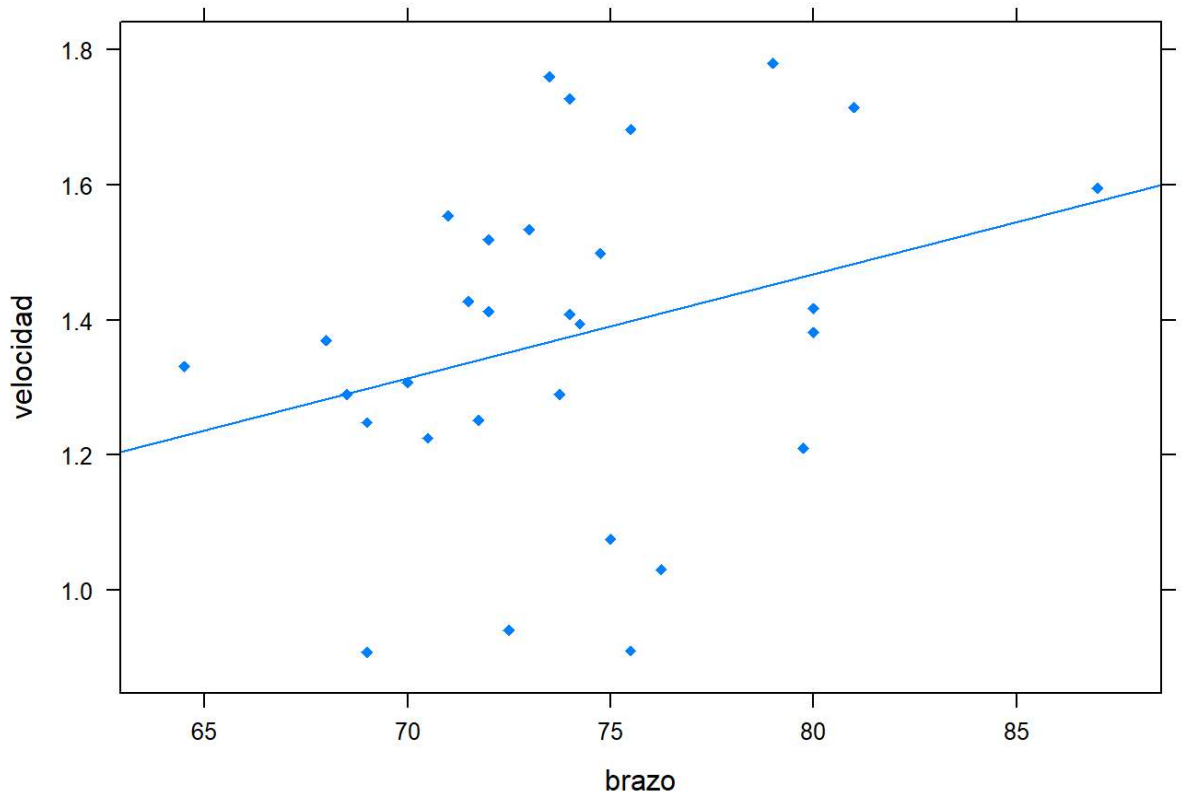
Hide

```
xyplot(veloc~pierna,type=c("p","r"),pch=18,ylab="velocidad")
```



Hide

```
xyplot(veloc~brazo,type=c("p","r"),pch=18,ylab="velocidad")
```



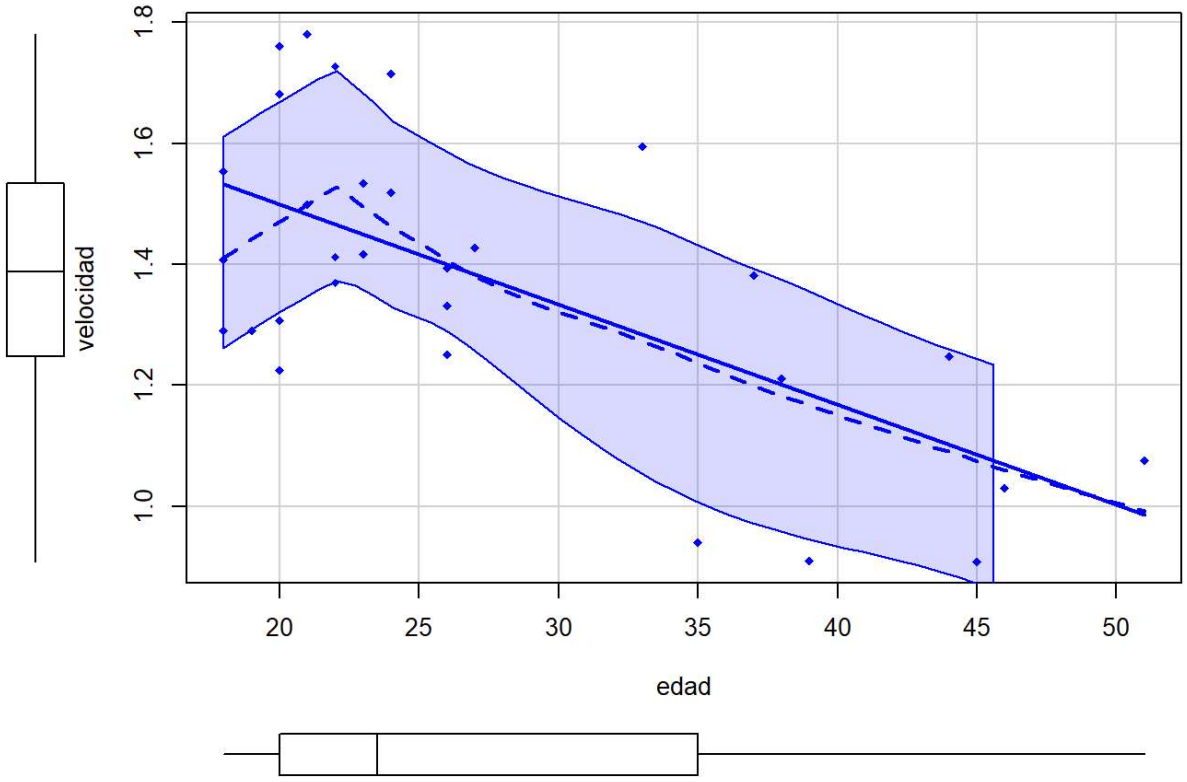
O puede usarse el paquete “car”:

La curva sólida azul es la recta ideal (generada por una relación lineal perfecta, en este caso con pendiente negativa) que es contrastada contra la curva punteada de color azul, que es una aproximación gráfica a la curva que modela las observaciones proporcionadas.

La región color azul claro permite identificar la zona de mayor concentración de valores que no se pudieron ajustar a la curva punteada azul (y, por consiguiente, permite explorar gráficamente -en relación a la curva punteada generada para ajustar los datos a un patrón geométrico de comportamiento- la dispersión en términos promedio y de su variabilidad).

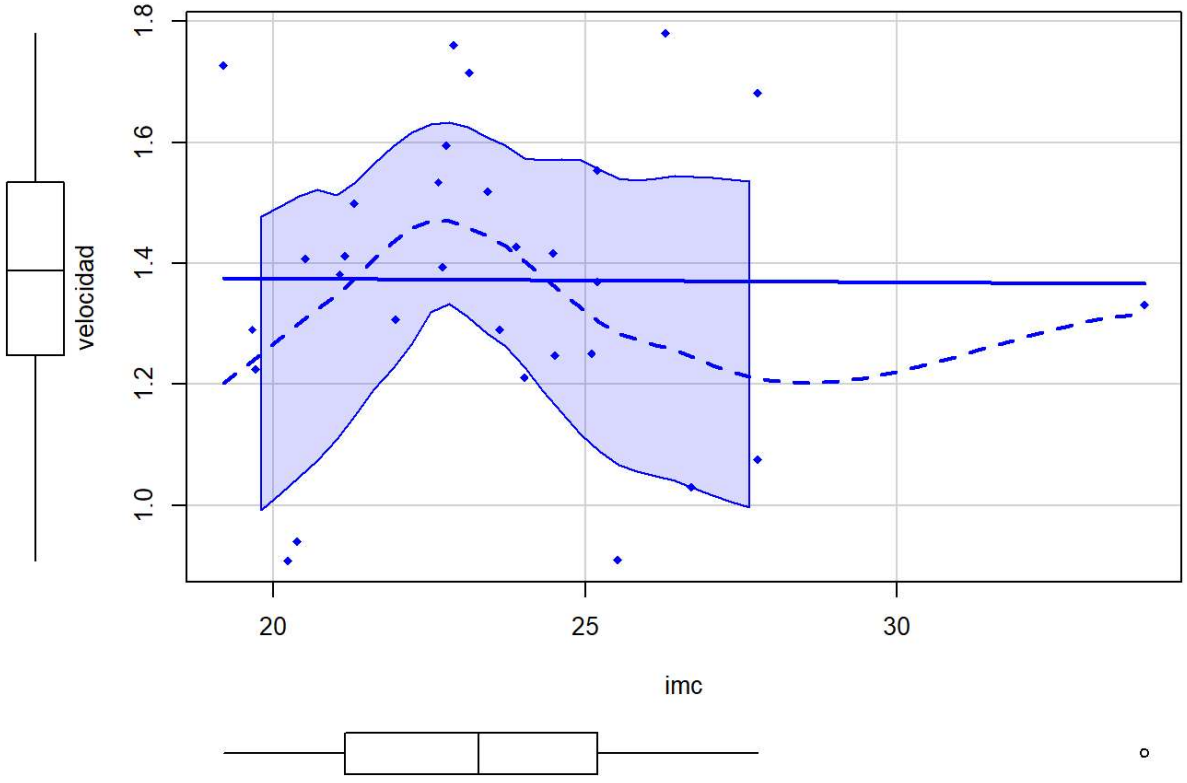
Hide

```
library(car)
scatterplot(edad,veloc,pch=18,ylab="velocidad")
```



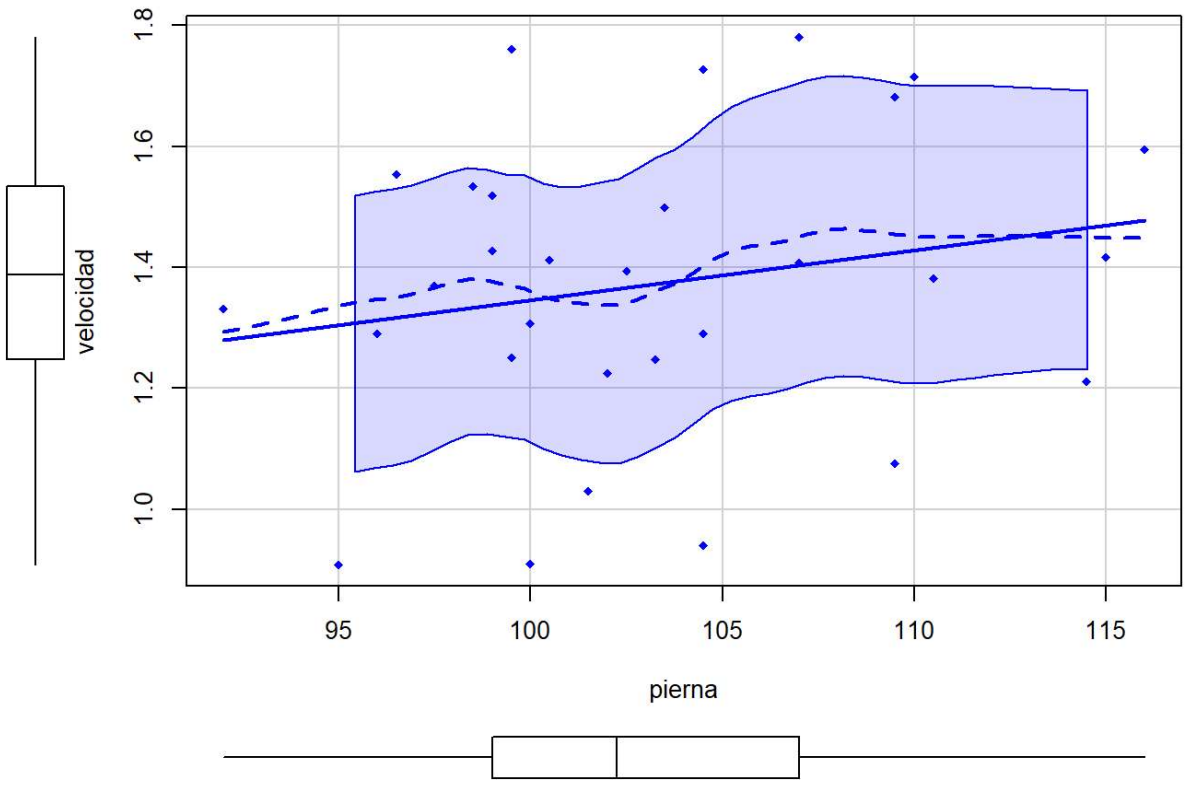
Hide

```
scatterplot(imc,veloc,pch=18,ylab="velocidad")
```



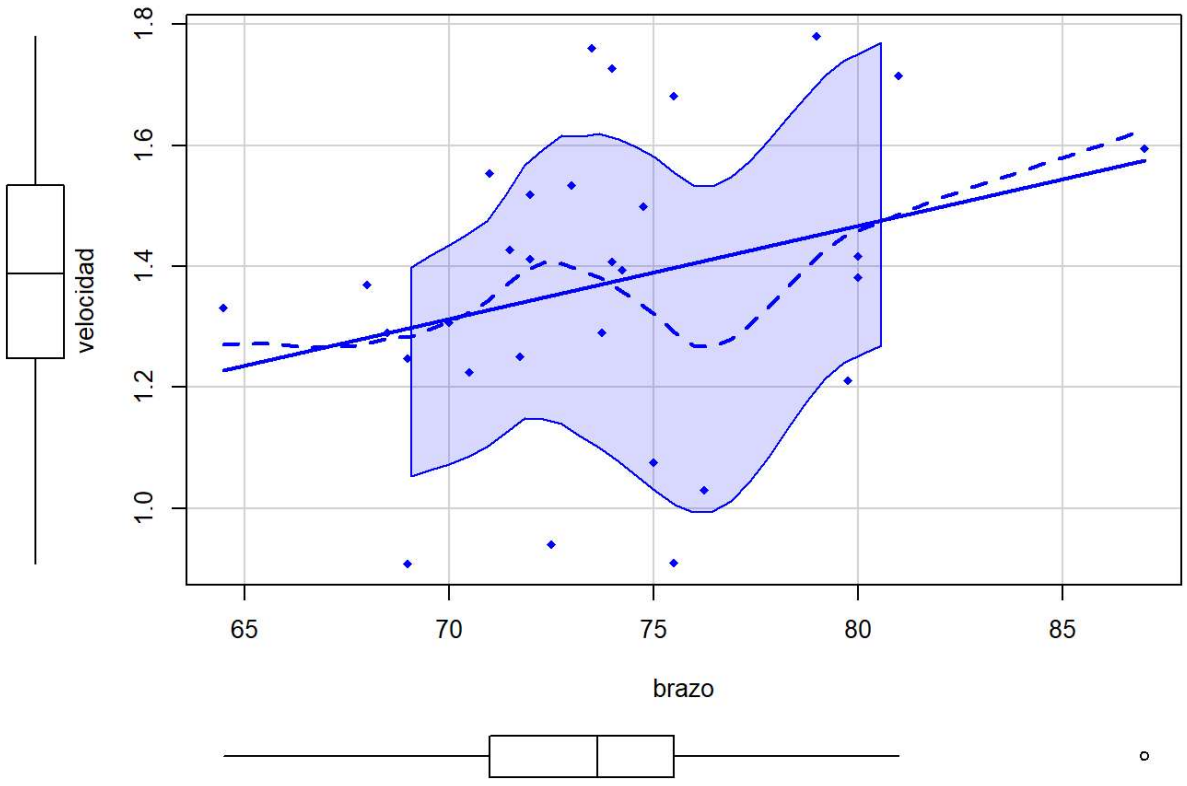
Hide

```
scatterplot(pierna,veloc,pch=18,ylab="velocidad")
```



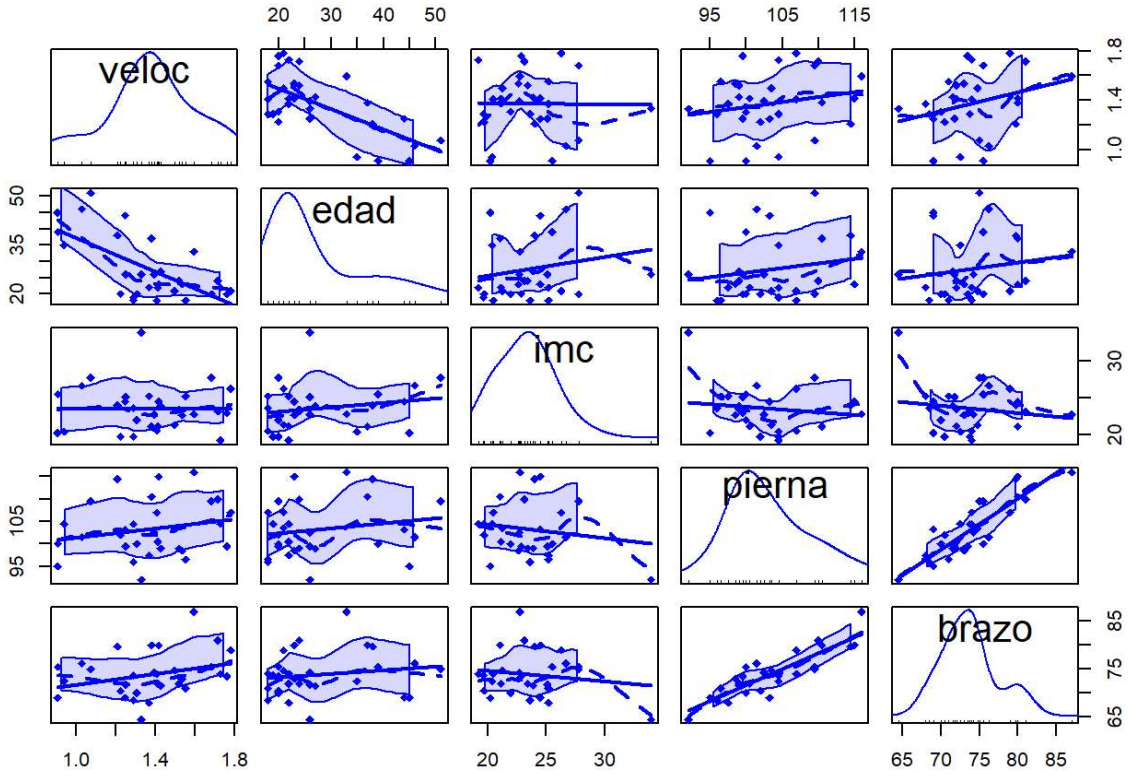
Hide

```
scatterplot(brazo,veloc,pch=18,ylab="velocidad")
```



Hide

```
scatterplotMatrix(base[,-3],pch=18)
```



Como se verifica en https://www.rdocumentation.org/packages/lattice/versions/0.10-10/topics/xyplot (https://www.rdocumentation.org/packages/lattice/versions/0.10-10/topics/xyplot), estas son las funciones Trellis (funciones que producen red algebraica, denominadas en inglés y en el contexto de la teoría de grupos como “Lattice”) de alto nivel más comúnmente utilizadas para trazar pares de variables. Con mucho, el más común es xyplot, diseñado principalmente para dos variables continuas (aunque también se pueden proporcionar factores, en cuyo caso simplemente se convertirán en numéricos), que produce gráficos de dispersión condicionales (como se señala en https://www.itl.nist.gov/div898/handbook/eda/section3/condplot.htm (https://www.itl.nist.gov/div898/handbook/eda/section3/condplot.htm), una gráfica condicional, también conocida como gráfica de coplot o subconjunto, es una gráfica de dos variables condicionado (que se hace en función de) sobre el valor de una tercera variable (llamada variable condicionante). La variable condicionante puede ser una variable que toma solo unos pocos valores discretos o una variable continua que se divide en un número limitado de subconjuntos; para este caso se trata de la esperanza matemática de la respuesta). Como se verifica en https://rpubs.com/mjs3pf/carpackage (https://rpubs.com/mjs3pf/carpackage), “car”, el nombre del paquete, es un acrónimo de Companion to Applied Regression. Este paquete no se utiliza para realizar técnicas de regresión aplicada, sino que complementa estas técnicas al proporcionar numerosas funciones que realizan pruebas, crean visualizaciones y transforman datos. Para comprobar la validez de numerosas técnicas de regresión, se necesitan realizar numerosas pruebas respecto a los resultados obtenidos. Este paquete proporciona las herramientas necesarias para hacerlo. Historia del paquete “car”: Este paquete se remonta a 2001. En 2002, el paquete se describió como “principalmente funciones para regresión aplicada, modelos lineales y modelos lineales generalizados, con énfasis en el diagnóstico de regresión, en

particular los métodos de diagnóstico gráfico". En 2010, se lanzó la versión 2.0 y el paquete pasó a depender de los paquetes R (> = 2.1.1), estadísticas, gráficos, MASS, nnet, saltos y supervivencia. En la versión 3.0, el paquete pasó a depender únicamente de R (> = 3.2.0) y carData (> = 3.0-0). carData importa una gran cantidad de conjuntos de datos que pueden ser utilizados por las funciones del paquete del automóvil. Esta dependencia se analizará en la siguiente subsección. Al 23 de septiembre de 2020, el paquete estaba en la versión 3.0-10. Para que las inferencias estadísticas (no confundir con la estadística inferencial) derivadas del uso del paquete "car" sean válidas (las relativas a regresión lineal múltiple) se deben cumplir los siguientes supuestos: 1. Los errores tienen un valor medio de 0, 2. Homoscedasticidad: los errores tienen una varianza constante en todos los niveles de las variables independientes, 3. Los errores se distribuyen normalmente, 4. Los errores son independientes entre sí (no hay autocorrelación). Como se verifica en https://cran.r-project.org/web/packages/car/car.pdf (https://cran.r-project.org/web/packages/car/car.pdf), la función "scatterplot" utiliza gráficos de R básicos para dibujar un diagrama de dispersión bidimensional, con opciones para permitir mejoras en el diagrama que a menudo son útiles con problemas de regresión. Las mejoras incluyen la adición de diagramas de caja marginales, media estimada (la media estimada -como primer momento de probabilidad-, esperanza matemática o valor esperado no debe confundirse con la media artimética o media simple) y funciones de varianza utilizando métodos paramétricos o no paramétricos, identificación de puntos, fluctuación, configuración de características de puntos y líneas como color, tamaño y símbolo, puntos de marcado y líneas de ajuste condicionadas a una variable de agrupación, y otras mejoras. "sp" es una abreviatura de "scatterplot". Esta función proporciona una interfaz conveniente para que la función de pares (una función que genera matrices con gráficos de dispersión, como se verifica en https://www.rdocumentation.org/packages/graphics/versions/3.6.2/topics/pairs (https://www.rdocumentation.org/packages/graphics/versions/3.6.2/topics/pairs)) produzca matrices de diagramas de dispersión mejorados (como se verifica en https://rdrr.io/cran/car/man/scatterplot.html (https://rdrr.io/cran/car/man/scatterplot.html), los diagramas dispersión mejorados introduce características adicionales a los gráficos de dispersión en la parte superior del gráfico de dispersión existente, en este caso para proporcionar más información sobre aspectos relacionados a la regresión), incluidas visualizaciones univariadas en la diagonal y una variedad de rectas ajustadas ajustadas (regresiones), suavizadores, funciones de varianza y elipsoides de concentración. "spm" es una abreviatura de "scatterplotMatrix". Cada paquete ofrece herramientas diferentes. El paquete "lattice" está orientado a graficar datos, mientras que el paquete "car" está orientado a realizar gráficas en el contexto del análisis de regresión. "lattice" se vale de una red algebraica para representar al conjunto de datos, mientras que "car" usa para graficar procesos (técnicas, metodologías, etc.) relativos al análisis de regresión.

CONCLUSIÓN: Asumiendo que se cumplen los supuestos mencionados en relación a las rectas de regresión generadas por las sintaxis que usan regresión lineal múltiple, la variable que posee una mejor relación lineal con la respuesta promedio es la edad. La otras variables tienen comportamientos que no parecen ser lineales; sin embargo, no se puede apreciar ninguna tendencia no-lineal clara. En el caso de imc la tendencia se aproxima a una recta horizontal, es decir, no hay ni aumento ni disminución de la velocidad promedio al aumentar el imc. En el caso de brazo habría que tener más datos en el rango de 75 a 80 para que la tendencia se dibujara de forma más clara.

b. Procédase a estimar los coeficientes de correlación lineal de Pearson entre los pares de variables que sea posible generar del conjunto disponible, en donde cada par siempre lleva incluida a la velocidad como variable de respuesta.

Lo primero que puede hacerse es construir un vector de variables X y definir Y como el vector de respuesta que contendrá los datos de la velocidad (la variable a explicar o dependiente), para luego estimar los coeficientes requeridos.

"cbind" sirve para meter todas esas variables dentro de una misma matriz, en donde las variables se localizan en las columnas. Almacena y nombra "R" una matriz que contiene en su primera columna el tipo de variable independiente (edad, imc, pierna y brazo) y en su segunda columna el valor del coeficiente de correlación de Pearson en relación con la variable de respuesta Y (velocidad).

Hide

```
X=cbind(edad,imc,pierna,brazo)
print(X)
```

```
##       edad      imc pierna brazo
##  [1,]   22 21.14769 100.50 72.00
##  [2,]   20 21.95529 100.00 70.00
##  [3,]   18 23.62755  96.00 68.50
##  [4,]   20 19.72104 102.00 70.50
##  [5,]   22 25.19531  97.50 68.00
##  [6,]   21 21.29529 103.50 74.75
##  [7,]   23 24.48565 115.00 80.00
##  [8,]   23 22.64738  98.50 73.00
##  [9,]   22 19.20415 104.50 74.00
## [10,]   24 23.43374  99.00 72.00
## [11,]   20 22.88750  99.50 73.50
## [12,]   38 24.02128 114.50 79.75
## [13,]   33 22.77614 116.00 87.00
## [14,]   39 25.52055 100.00 75.50
## [15,]   37 21.06676 110.50 80.00
## [16,]   35 20.38157 104.50 72.50
## [17,]   45 20.22913  95.00 69.00
## [18,]   46 26.70362 101.50 76.25
## [19,]   18 25.19531  96.50 71.00
## [20,]   24 23.14150 110.00 81.00
## [21,]   20 27.76980 109.50 75.50
## [22,]   18 20.51509 107.00 74.00
## [23,]   21 26.28571 107.00 79.00
## [24,]   19 19.66412 104.50 73.75
## [25,]   26 33.96130  92.00 64.50
## [26,]   26 22.72044 102.50 74.25
## [27,]   51 27.76980 109.50 75.00
## [28,]   26 25.09950  99.50 71.75
## [29,]   44 24.50895 103.25 69.00
## [30,]   27 23.90003  99.00 71.50
```

Hide

```
Y=veloc
R=cor(X,Y)
round(R,2) #Rodendeo de cifras
```

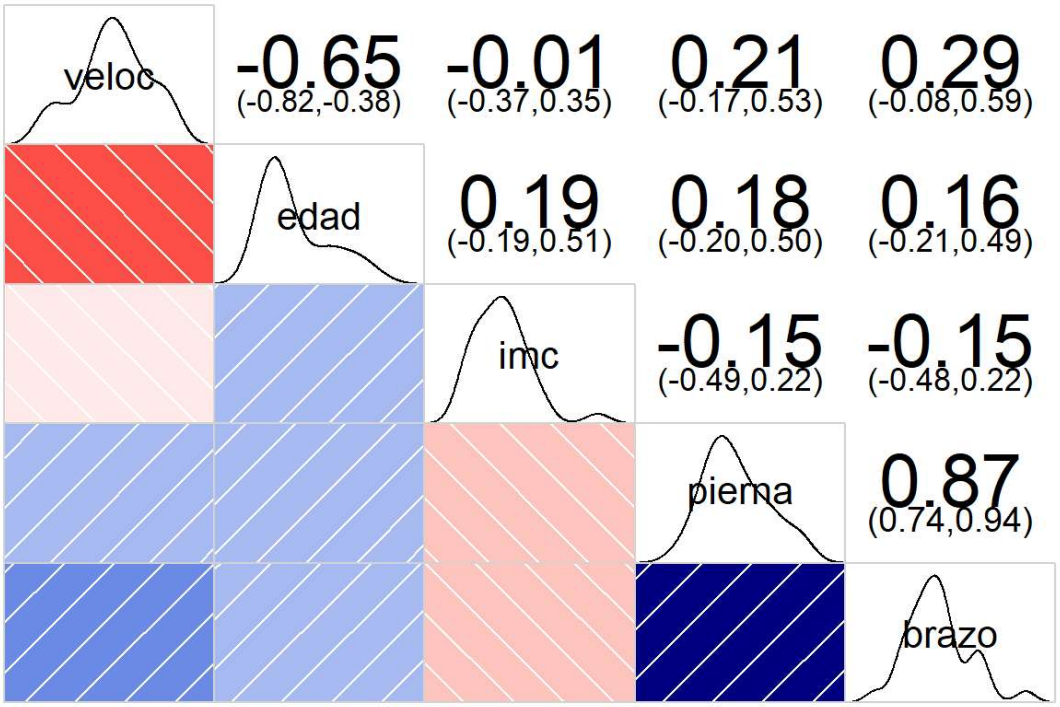
```
##         [,1]
## edad   -0.65
## imc    -0.01
## pierna  0.21
## brazo   0.29
```

CONCLUSIÓN: La única variable que muestra una asociación importante con la respuesta es la edad.

c. Con la finalidad de, entre otras cuestiones, realizar una primera aproximación inferencial a la verificación (o no) del supuesto del modelo clásico de regresión lineal de no-multicolinealidad (ninguna de las variables explicativas puede explicarse como una adición y/o multiplicación por escalar del resto de variables explicativas), es posible hacer un gráfico que contenga todas las correlaciones de Pearson entre variables numéricas. Para ello, puede usarse la función "corrgram" de la librería "corrgram". Simplemente debe escribirse dentro de la función el objeto que contiene las variables a analizar: "corrgram(base[,-3])". Los colores más intensos representan asociaciones más fuertes, donde rojo es negativo y azul es positivo. Esta función permite combinaciones de gráficos, por ejemplo, se podría presentar en la diagonal una densidad univariada mediante el parámetro "diag.panel=panel.density" y en la parte superior los valores de las correlaciones con sus intervalos de confianza usando "upper.panel=panel.conf".

Hide

```
library(corrgram)
corrgram(base[,-3], upper.panel=panel.conf,
         diag.panel=panel.density)
```



CONCLUSIONES: Se confirma que la variable que tiene una asociación lineal más fuerte con la velocidad es la edad, siendo esta asociación inversa. Por otro lado, se ve que el imc tiene una asociación prácticamente nula con la velocidad. Se nota que brazo y pierna están altamente correlacionadas y siendo estos dos predictores (presuntamente lineales) se estaría incumpliendo de no-multicolinealidad.

d. Es posible construir un modelo de regresión que tenga como predictores solamente "edad" y "largo promedio del brazo". Se plantea un modelo en tales términos debido a muy baja conrrelación de "imc" con la respuesta, y se excluyen "largo promedio de las piernas" y "largo promedio del brazo" por su elevada correlación.

e. A continuación, puede construirse la matriz de variables regresoras X_.

Hide

```
X_=cbind(1,edad,brazo)
Y=veloc
```

f. Puede procederse a realizar las estimaciones para los coeficientes del modelo que minimizan la suma de cuadrados de los residuales, es decir, la expresión derivada por el método de mínimos cuadrados: $\hat{\beta} = (X^T X)^{-1}(X^T Y)$. Recordar que la multiplicación de matrices en R se realiza con el operador `%*%`, la transpuesta de una matriz con la función `t` y la inversión de matrices con la función `solve`.

Hide

```
beta=solve(t(X_)%*%X_)%*%t(X_)%*%Y
round(beta,3)
```

```
##         [,1]
##        0.292
## edad  -0.018
## brazo  0.021
```

Los resultados arrojan el siguiente modelo de regresión: $\hat{y} = 0.29 - 0.02X_1 + 0.02X_2$

g. Estimando las desviaciones estándar para las regresoras estadísticamente relevantes.

Hide

```
sd_edad=sd(edad)
sd_brazo=sd(brazo)
```

h. Puede realizarse una interpretación de los coeficientes considerando que cada variable predictora (explicativa, independiente) aumenta alrededor de una desviación estándar.

el "2" en la segunda posición de la sintaxis "round" denota el número de infinitesimales deseado en el redondeo "[]" denota que de la matriz columna que contiene los coeficientes de regresión se extraerá el elemento de la posición 3 (en el caso anterior el de la posición 2) y se multiplicará por el nuevo nivel de la variable (que se desvía en una unidad de su magnitud inicial)

Hide

```
round(beta[2]*10,2)
```

```
## [1] -0.18
```

Hide

```
round(beta[3]*5,2)
```

```
## [1] 0.11
```

Al comparar dos grupos de personas cuyas edades difieren en 10 años y manteniendo fijo el largo promedio de los brazos, la velocidad promedio disminuye 0.18m/s en el grupo de mayor edad con respecto al grupo de menor edad. Al comparar dos grupos de personas cuyo largo promedio de los brazos difiere en 5cm y manteniendo fija la edad, la velocidad promedio aumenta 0.11m/s en el grupo de mayor largo de brazos con respecto al grupo con menor largo de brazos.

i. Con el modelo de regresión obtenido anteriormente, es posible calcular el valor asociado a una edad de 24 años y un largo de brazos promedio de 71 cm.

Hide

```
Y_est_=t(beta)%*%c(1,24,71)
```

Con la sintaxis “c” se le indica a R que debe extraer de la matriz de regresoras X_ un “1” (que se requiere por una cuestión de estimación computacional dada la estructura de la sintaxis, no matemáticas), un 24 (que le indica a R que extraiga todos los elementos del vector edad con valores iguales a 24) y un 71 (que le indica a R que se extraiga del vector de longitudes de brazo todos aquellos elementos cuya magnitud sea igual a 71).

Hide

```
beta[1]+beta[2]*24+beta[3]*71
```

```
## [1] 1.376928
```

La velocidad promedio de todas las personas de 24 años que tengan un largo de brazos promedio de 71 cm., se estima en 1.38 m/s.

j. El investigador podría desear conocer cuánto cambiaría este valor si se consideran ahora personas de 34 años y con el mismo largo promedio de brazos

Hide

```
beta[1]+beta[2]*34+beta[3]*71
```

```
## [1] 1.195234
```

La velocidad promedio disminuiría 0.18 m/s, es decir, la velocidad promedio sería 1.20 m/s para todas las personas de 34 años con un largo de brazos promedio de 71 cm.

k. Los coeficientes de la ecuación de regresión también pueden estimarse de forma automatizada mediante la sintaxis "lm" (guardando el resultado en una estructura de datos adecuada bajo algún nombre, por ejemplo, "mod") y extrayendo de dicha estructura los coeficientes redondeados mediante la sintaxis "round(mod$coef,2)".

Hide

```
mod=lm(veloc~edad+brazo)
round(mod$coef,2)
```

```
## (Intercept)        edad       brazo
##        0.29       -0.02        0.02
```

l. Adicionalmente, también se puede realizar de forma automatizada la estimación de la media condicional de la velocidad, por ejemplo, para personas con edad de 24 y un largo de brazos promedio de 71cm. Para esto puede usarse la función "predict" construyendo un `data.frame` para almacenar los valores de los predictores. El proceso descrito antes se realiza de la forma que se muestra a continuación.

Hide

```
predict(mod,data.frame(edad=24,brazo=71))
```

```
##        1
## 1.376928
```

m. También es posible realizar una estimación con todas las variables explicativas del modelo:

Hide

```
mod1=lm(veloc~edad+imc+pierna+brazo)
round(mod$coef,3)
```

```
## (Intercept)        edad       brazo
##       0.292      -0.018       0.021
```

Así, la ecuación de regresión adopta la siguiente forma:
$\hat{y} = -0.135 - 0.019X_1 + 0.016X_2 - 0.003X_3 + 0.026X_4 +$

n. Puede proseguirse a estimar los coeficientes estandarizados (se les llama así porque provienen de variables estandarizadas y su uso permite comparar los coeficientes de regresión entre variables con diferentes escalas de medida) del modelo de regresión lineal siguiendo los pasos que se exponen a continuación.

ASPECTOS TEÓRICOS SOBRE LOS COEFICIENTES ESTANDARIZADOS

Como se señala en https://www.wikiwand.com/en/Standardized_coefficient (https://www.wikiwand.com/en/Standardized_coefficient), la estandarización del coeficiente generalmente se hace para responder a la pregunta de cuál de las variables independientes tiene un mayor efecto sobre la variable dependiente en un análisis de regresión múltiple donde las variables se miden en diferentes unidades de medida (por ejemplo, ingresos medidos en dólares

y el tamaño de la familia medido en número de individuos). También puede considerarse una medida general del tamaño del efecto, cuantificando la "magnitud" del efecto de una variable sobre otra. Para la regresión lineal simple con predictores ortogonales, el coeficiente de regresión estandarizado es igual a la correlación entre las variables independientes y la dependiente. Los defensores de los coeficientes estandarizados señalan que los coeficientes son independientes de las unidades de medida de las variables involucradas (es decir, los coeficientes estandarizados no tienen unidades), lo que facilita las comparaciones. Los críticos expresan su preocupación de que tal estandarización puede ser muy engañosa. Debido al cambio de escala basado en las desviaciones estándar de la muestra, cualquier efecto aparente en el coeficiente estandarizado puede deberse a la confusión con las particularidades (especialmente: la variabilidad) de las muestras de datos involucradas. Además, la interpretación o el significado de un "cambio de una desviación estándar" en el regresor X puede variar marcadamente entre distribuciones no-normales (por ejemplo, cuando está sesgada, asimétrica o multimodal).

Sobre la relación de ortogonalidad (i.e., independencia lineal) entre los predictores, señala https://www.statisticshowto.com/orthogonality/ (https://www.statisticshowto.com/orthogonality/) que el término "ortogonal" generalmente solo se aplica al ANOVA clásico. Un ANOVA ortogonal tiene todas las variables independientes categóricas y cada celda en una tabla de dos factores tiene el mismo número de observaciones (llamado diseño balanceado). En contraste, los modelos lineales generales nunca son ortogonales, ya que al menos una variable independiente no es categórica (los GLM tienen una variable continua).

Finalmente, puede consultarse la importancia de la estandarización de unidades de medición en https://byjus.com/physics/dimensional-analysis/ (https://byjus.com/physics/dimensional-analysis/), que explica con cierto nivel de detalle la relevancia del análisis dimensional en contextos como la Física y las Ingenierías, por ejemplo.

n.1. Obtener las correlaciones entre la respuesta y cada uno de los predictores. Para ello, constrúyase un objeto llamado "X" y deposítese en los predictores, con el fin de construir la estructura de datos "rxy=cor(X,Y)" con el uso de la sintaxis "cor()":

Hide

```
X=cbind(edad,imc,pierna,brazo)
Y=veloc
rxy=cor(X,Y)
```

n.2. Obtener la matriz de correlaciones entre los predictores construyendo el objeto "rxx=cor(X)":

Hide

```
rxx=cor(X)
```

n.3. Obtener los coeficientes estandarizados a través de la expresión derivada por mínimos cuadrados: $\hat{\beta}^{S} = (r_{XX})^{-1}(r_{XY})$, donde r_{XX}. En la expresión anterior, r_{XX} es la matriz anterior con las correlaciones entre los predictores y r_{XY} es el vector de correlaciones entre los predictores y la respuesta:

Hide

```
beta.s=solve(rxx) %*% rxy
round(beta.s,2)
```

```
##         [,1]
## edad   -0.76
## imc     0.20
## pierna -0.07
## brazo   0.50
```

n.4. La ecuación de regresión resultante toma la siguiente forma:
$\hat{y}^S = -0.76X_1^S + 0.2X_2^S - 0.07X_3^S + 0.5X_4^S$

n.5. También es posible obtener los coeficientes estandarizados usando directamente las variables estandarizadas en la sintaxis "lm". Para estandarizar todas las variables se puede recurrir a la función "scale" dentro de la función "sapply" de la siguiente forma "sapply(base[,-3],scale)".

Hide

```
#Forma Manual
base2=data.frame(sapply(base[,-3],scale))
mod_base2=lm(veloc~-1+edad+imc+pierna+brazo,data=base2)
round(mod_base2$coef,2)
```

```
##   edad    imc pierna  brazo
##  -0.76   0.20  -0.07   0.50
```

Hide

```
#Forma Automatizada
library(QuantPsyc)
mod=lm(veloc ~ ., data=base)
coef_lmbeta=lm.beta(mod)
print(coef_lmbeta)
```

```
##       edad       sexo        imc     pierna      brazo
## -0.7986598  0.2079902  0.1772575 -0.1056223  0.4106855
```

n.6. Pueden obtenerse las desviaciones estándar de cada variable (respuesta y predictores) con la sintaxis "apply". En este caso se colocan las variables en un objeto, luego se indica la dirección: fila 1, columna 2, y luego la acción a realizar a realizar (en este caso, que se calcule la desviación estándar mediante "sd"). Ya que se desean resultados por columna, se escribe un 2 en la sintaxis de la siguiente forma: "apply(base[,-3],2,FUN="sd")".

Hide

```
desv=apply(base[,-3],2,FUN = "sd")
```

n.7. A partir de los coeficientes no-estandarizados pueden obtenerse los coeficientes estandarizados usando las desviaciones estándar anteriores. Este procedimiento se modela mediante la siguiente expresión:

$$\hat{\beta}_j^S = \hat{\beta}_j \times \frac{s_{X_j}}{s_Y}$$

Hide

```
round(mod1$coef[-1]*desv[-1]/desv[1],2)
```

```
##   edad    imc pierna  brazo
##  -0.76   0.20  -0.07   0.50
```

CONCLUSIONES Y FORMA DE INTERPRETACIÓN DE LOS RESULTADOS ESTANDARIZADOS: La variable que tiene mayor impacto es “edad”, pues disminuye la respuesta promedio en 0.76 deviaciones estándar por cada aumento de una desviación estándar de la edad. En segundo lugar, se coloca la variable relativa al largo promedio de los brazos, dado que aumenta la respuesta promedio en 0.5 desviaciones estándar por cada aumento de una desviación estándar de la variable brazo. Hay que notar que el “imc” no tiene un coeficiente tan bajo como alguien podría esperar en la etapa inicial de la investigación, así como también que el largo de las piernas parece tener una incidencia muy leve, a juzgar por su parámetro de regresión; esto es extraño por cuanto las variables “brazo” y “piernas” están muy correlacionadas. Finalmente, debe observarse también que en el modelo no-estandarizar la variable con el coeficiente más alto era “brazo”, mientras que al estandarizar las variables la de mayor coeficiente es “edad”.

o. Es posible también construir los intervalos de confianza para los coeficientes de regresión (estandarizados o no-estandarizados, aunque aquí se usarán los segundos); esto se hace en el literal o.12. Para ello, se construirán tres modelos con las variables disponibles: 1. Un modelo con los cuatro predictores originales, 2. uno con tres predictores (se eliminará “edad”) y 3. uno con los cuatro predictores originales, pero transformando el largo promedio de brazo de centímetros a milímetros. Posteriormente, se debe proceder a ajustar los tres modelos y comparar los coeficientes obtenidos.

o.1. Modelo vel=f(edad,imc,pierna,brazo)

Hide

```
mod1=lm(veloc~edad+imc+pierna+brazo)
```

o.2. Modelo vel=f(imc,pierna,brazo)

Hide

```
mod2=lm(veloc~imc+pierna+brazo)
```

o.3. Modelo vel=f(edad,imc,pierna,brazomm)

Hide

```
brazomm=brazo*10
mod3=lm(veloc~edad+imc+pierna+brazomm)
```

o.4. Construyendo la comparación entre coeficientes:

Hide

```
print(beta)
```

```
##              [,1]
##        0.29220247
## edad  -0.01816942
## brazo  0.02141960
```

Hide

```
beta1=mod1$coef
beta2=mod2$coef
beta3=mod3$coef
round(beta1,3)
```

```
## (Intercept)        edad         imc      pierna       brazo
##      -0.135      -0.019       0.016      -0.003       0.026
```

Hide

```
round(beta2,3)
```

```
## (Intercept)         imc      pierna       brazo
##       0.306       0.002      -0.008       0.025
```

Hide

```
round(beta3,3)
```

```
## (Intercept)        edad         imc      pierna     brazomm
##      -0.135      -0.019       0.016      -0.003       0.003
```

Entre el modelo 1 y 3 no se aprecian diferencias relevantes en los coeficientes, excepto en “brazomm” que fue la variable que se transformó. En este caso, como “brazo” se multiplicó por 10, el coeficiente se divide entre 10. Cuando se elimina “edad” sí cambian los coeficientes de “imc” y “pierna”. El coeficiente de “imc” en el modelo 1 representa el aumento en la velocidad promedio al aumentar “imc” en una unidad manteniendo “edad”, “pierna” y “brazo” constantes. Por otro lado, aunque en el modelo 2 represente lo mismo, cabe mencionar, a diferencia del anterior, sí se toma en consideración la variable “edad”.

0.5. También pueden obtenerse los valores ajustados (los valores de las variables independientes, dados los valores de los coeficientes -generados a partir del conjunto de datos-, que solucionan el sistema de ecuaciones para Y -la variable de respuesta-, en donde tal valor expresa la esperanza condicional de Y) para la variable de respuesta, por ejemnplo, el modelo 1, usando notación matricial.

Hide

```
X=cbind(1,edad,imc,pierna,brazo)
est=X%*%beta1

round(X[1,],2)
```

```
##          edad    imc pierna  brazo
##   1.00  22.00  21.15 100.50  72.00
```

Hide

```
round(est[1],2)
```

```
## [1] 1.4
```

La esperanza condicional de la velocidad de todas las personas de 22 años con un imc de 21.15, largo de pierna de 100.5 cm. y de brazo de 72 cm. se estima en 1.4m/s.

o.6. Es posible obtener los valores ajustados de las variables explicativas de tal forma que sea posible obtener las medias condicionales de la velocidad de cada uno de los elementos de la muestra (los individuos sujetos de estudio) para cada modelo y comparar los resultados. Para ello, puede usarse la función “predict” simplemente indicando como argumento el modelo, por ejemplo “predict(mod)”; “predict” es una función genérica para predicciones a partir de los resultados de varias funciones de ajuste de modelos. La función invoca métodos particulares que dependen de la clase del primer argumento (es decir, del tipo de modelo con el que se le solicite a R hacer las predicciones -en función de ello R seleccionará algún método numérico-).

Hide

```
y1=predict(mod1)
y2=predict(mod2)
y3=predict(mod3)
round(cbind(y1,y2,y3),2)
```

```
##      y1   y2   y3
## 1  1.40 1.34 1.40
## 2  1.40 1.30 1.40
## 3  1.44 1.30 1.44
## 4  1.37 1.29 1.37
## 5  1.37 1.28 1.37
## 6  1.49 1.39 1.49
## 7  1.61 1.43 1.61
## 8  1.44 1.39 1.44
## 9  1.41 1.36 1.41
## 10 1.40 1.36 1.40
## 11 1.51 1.39 1.51
## 12 1.31 1.43 1.31
## 13 1.57 1.59 1.57
## 14 1.24 1.44 1.24
## 15 1.30 1.46 1.30
## 16 1.14 1.32 1.14
## 17 0.88 1.31 0.88
## 18 1.14 1.45 1.14
## 19 1.52 1.36 1.52
## 20 1.61 1.49 1.61
## 21 1.61 1.37 1.61
## 22 1.50 1.34 1.50
## 23 1.67 1.48 1.67
## 24 1.47 1.35 1.47
## 25 1.35 1.26 1.35
## 26 1.40 1.39 1.40
## 27 1.00 1.36 1.00
## 28 1.38 1.35 1.38
## 29 0.95 1.26 0.95
## 30 1.34 1.35 1.34
```

Sobre la convergencia y divergencia (según el caso) entre los resultados de los tres modelos, debe decirse que los resultados del modelo 1 y 3 coinciden porque se están usando las mismas variables; esto se dice por cuanto el cambio de escala hecho a “brazo” (deviniendo en “brazomm”) se compensa con el cambio en su coeficiente. En contraste, el modelo 2 no tiene “edad”, por lo que la media condicional de la velocidad no tiene por qué ser igual tanto si no se toma en consideración la edad como si se toma.

0.7.1. SOBRE EL ERROR CUADRÁTICO MEDIO (MS) Y EL ERROR CUADRÁTICO MEDIO DE REGRESIÓN (MSE) 0.7.1.1. Como se señala en https://support.minitab.com/es-mx/minitab/18/help-and-how-to/modeling-statistics/anova/supporting-topics/anova-statistics/understanding-mean-squares/ (https://support.minitab.com/es-mx/minitab/18/help-and-how-to/modeling-statistics/anova/supporting-topics/anova-statistics/understanding-mean-squares/), el cuadrado medio de un factor (tratamiento, en el contexto usual de la Bioestadística) se obtiene dividiendo la suma de los cuadrados del tratamiento (del factor o variable explicativa en cuestión) entre los grados de libertad n-p (es un error cuadrático medio específico a cada factor), donde n es el tamaño de la muestra y p es la cantidad de variables independientes; el error cuadrático medio de estimación del parámetro correspondiente a una variable explicativa (MS) representa la variación promedio del error de predicción correspondiente a dicha variable explicativa. Por otro lado, el cuadrado medio del

error (MSE) se obtiene dividiendo la suma de los cuadrados del error residual (que se genera como error de predicción global del modelo) entre los grados de libertad. El MSE representa la variación dentro de las muestras. Por ejemplo, si se realiza un experimento para probar la efectividad de tres detergentes para ropa y se recolectan 20 prendas sobre las que se ha aplicado cada detergente (60 observaciones en total), la variación entre las medias de “Detergente 1”, “Detergente 2” y “Detergente 3” (es decir, la variación del efecto promedio que tienen los detergentes en la ropa) es representada por el cuadrado medio de los errores de estimación considerando individualmente a alguna de estas variables y a su coeficiente de regresión (junto con el intercepto, si fuese el caso). De manera complementaria, la variación dentro de las muestras es representada por el cuadrado medio del error de estimación MSE. En suma, el error cuadrático medio de estimación (MSE o ECM por su nombre en español) de un estimador mide el promedio de los errores de estimación al cuadrado, es decir, la diferencias al cuadrado entre el estimador (o conjunto de estimadores, en el caso de ser un modelo multivariado) y lo que se estima. El MSE o ECM es una función de riesgo, correspondiente al valor esperado de la pérdida del error al cuadrado o pérdida cuadrática. La diferencia se produce debido a la aleatoriedad o porque el estimador no tiene en cuenta la información que podría producir una estimación más precisa. Adicionalmente, se señala en https://support.minitab.com/es-mx/minitab/18/help-and-how-to/modeling-statistics/anova/supporting-topics/anova-statistics/understanding-mean-squares/ (https://support.minitab.com/es-mx/minitab/18/help-and-how-to/modeling-statistics/anova/supporting-topics/anova-statistics/understanding-mean-squares/) que la suma ajustada de cuadrados (MS) no depende del orden en que se ingresan los factores en el modelo y es la porción única del error cuadrático medio de la regresión explicada por un factor (variable independiente), asumiendo como constantes todos los demás factores en el modelo. Cuando las variables explicativas se consideren fijas el denominador de los estadísticos F será el MSE. Sin embargo, para los modelos que incluyen términos aleatorios, el MSE no siempre es el término de error correcto a utilizar en la prueba F. Por ello, pueden examinarse los cuadrados medios esperados (E[MSE]) para determinar el término de error que se utilizó en la prueba F; esta técnica no se examina aquí. Lo anterior es importante puesto que, como se señala en https://blog.minitab.com/es/comprension-del-analisis-de-varianza-anova-y-la-prueba-f (https://blog.minitab.com/es/comprension-del-analisis-de-varianza-anova-y-la-prueba-f), a pesar de ser una relación de varianzas, la prueba F se puede utilizar en una amplia variedad de situaciones y, por ello, la comprensión del MSE (y las estructuras algebraicas afines al MSE) es de fundamental importancia aplicada. Finalmente, debe decirse que, como es sabido, la prueba F puede ser utilizada en la prueba de hipótesis de igualdad de las varianzas; sin embargo, al cambiar las varianzas que se incluyen en la relación, la prueba F se convierte en una prueba muy flexible. Por ejemplo, las estadísticas F y las pruebas F se pueden utilizar para evaluar la significancia general de un modelo de regresión, para comparar el ajuste de diferentes modelos, para probar términos de regresión específicos y para evaluar la igualdad de las medias. 0.7.1.2. Como se señala en
https://web.archive.org/web/20170510161951/http://support.minitab.com/en-us/minitab/17/topic-library/basic-statistics-and-graphs/introductory-concepts/basic-(https://web.archive.org/web/20170510161951/http://support.minitab.com/en-us/minitab/17/topic-library/basic-statistics-and-graphs/introductory-concepts/basic-) conceptos / df /, los grados de libertad (gl) son la cantidad de información que brindan sus datos y que puede “gastar” para estimar los valores de parámetros de población desconocidos y calcular la variabilidad de estas estimaciones. Este valor está determinado por el número de observaciones en su muestra y el número de parámetros en su modelo. Aumentar el tamaño de la muestra proporciona más información sobre la población y, por lo tanto, aumenta los grados de libertad en sus datos. Agregar parámetros a su modelo (al aumentar el número de términos

en una ecuación de regresión, por ejemplo) “gasta” información de sus datos y reduce los grados de libertad disponibles para estimar la variabilidad de las estimaciones de los parámetros. Los grados de libertad también se utilizan para caracterizar una distribución específica. Muchas familias de distribuciones, como t, F y chi-cuadrado, usan grados de libertad para especificar qué distribución específica t, F o chi-cuadrado es apropiada para diferentes tamaños de muestra y diferentes números de parámetros del modelo. Por ejemplo, la siguiente figura muestra las diferencias entre distribuciones de chi-cuadrado con diferentes grados de libertad. Por ejemplo, la prueba t de 1 muestra estima solo un parámetro: la media de la población. El tamaño de muestra de n constituye n piezas de información para estimar la media poblacional y su variabilidad. Se gasta un grado de libertad en estimar la media y los n-1 grados de libertad restantes estiman la variabilidad. Por lo tanto, una prueba t de 1 muestra utiliza una distribución t con n-1 grados de libertad. Por el contrario, la regresión lineal múltiple debe estimar un parámetro para cada término que elija incluir en el modelo, y cada uno consume un grado de libertad. Por lo tanto, incluir términos excesivos en un modelo de regresión lineal múltiple reduce los grados de libertad disponibles para estimar la variabilidad de los parámetros y puede hacerlo menos confiable. 0.7.1.3. Como se señala en https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/mean-squared-error/ (https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/mean-squared-error/), cuanto menor sea el error cuadrático medio de regresión “SCreg”, más cerca estará de encontrar la línea de mejor ajuste. Dependiendo del conjunto de datos disponible, puede ser imposible obtener un valor muy pequeño para el error cuadrático medio. El “SCreg” es particularmente útil cuando existen varias ecuaciones de regresión, puesto que la que produzca el menor “SCreg” es la ecuación de regresión que genera la recta de mejor ajuste. 0.7.1.4. También pueden emplear las sumas de cuadrados totales residuales “SCTR” con el fin de compararlas y explicar sus diferencias. 0.7.1.5. Recapitulando, no deben confundirse conceptualmente la suma de cuadrados medios de la regresión “SCreg” (véase https://blog.minitab.com/en/marilyn-wheatleys-blog/what-the-heck-are-sums-of-squares-in-regression (https://blog.minitab.com/en/marilyn-wheatleys-blog/what-the-heck-are-sums-of-squares-in-regression) y https://economipedia.com/definiciones/suma-de-cuadrados-de-la-regresion-scr.html (https://economipedia.com/definiciones/suma-de-cuadrados-de-la-regresion-scr.html)) con la suma de cuadrados medios residuales “SCres” (que es una estimación del error cometido en la predicción, véase https://economipedia.com/definiciones/suma-de-cuadrados-de-los-residuos-sce.html (https://economipedia.com/definiciones/suma-de-cuadrados-de-los-residuos-sce.html)), así como tampoco con la suma de cuadrados totales residuales “SCT” o con la suma de cuadrados totales de Y “SCTY” (que es invariante y se conoce a priori, véase https://economipedia.com/definiciones/suma-total-de-cuadrados-stc.html (https://economipedia.com/definiciones/suma-total-de-cuadrados-stc.html)), así como tampoco alguna de estas entre sí. De lo anteriormente expuesto se desprende que la suma de cuadrados total de la respuesta (que es a priori y se denotará aquí como “SCTY”) es igual a la suma de cuadrados medios de regresión “SCreg” más la suma de cuadrados totales residuales “SCTR”, es decir, SCreg1+SCTR1, SCreg2+SCTR2 y SCreg3+SCTR4 son iguales a la suma de cuadrados totales a priori de Y (que es siempre la misma con independencia de las variables en consideración, porque es el promedio de las distancias de los valores a priori de la respuesta respecto a su media). En suma, “SCreg= sum((fit-m)^2” se construye con las distancia entre el promedio a priori y el promedio estimado (donde fit= predict(mod)), “SCTY= sum(anova(mod)[,2]” se construye con las distancias del promedio a priori y las observaciones i-ésimas que componen a Y, “SCTR=anova(mod1)[5,2]= t(r1)%*%r1” se construye transponiendo la matriz de errores y multiplicándola por la matriz de errores (donde r es la matriz de errores) y “SCres=

anova(mod1)[5,3]" se construye con las distancias entre la media estimada y las observaciones i-ésimas que componen Y. 0.7.2. Se puede proceder a calcular los valores residuales (r_i) relativos al SSE_i correspondiente a cada modelo. Esto se puede hacer de dos formas: 1. conociendo la relación $r = y - \hat{y}$ (que equivaldría a un cálculo manual consistente en multiplicar la transpuesta de la matriz de errores por la matriz de errores) o 2. directamente con "mod$res" (utilizando directamente los residuos generados por algún modelo de estudio). A continuación, se presentan los cálculos pertinentes utilizando la opción 2.

Hide

```
mod4=lm(veloc~edad)
r1=mod1$res; SCTR1=t(r1)%*%r1
r2=mod2$res; SCTR2=t(r2)%*%r2
r4=mod4$res; SCTR4=t(r4)%*%r4
round(c(SCTR1,SCTR2,SCTR4),4)
```

```
## [1] 0.6452 1.5634 0.9902
```

En los modelos 1 y 3 se obtiene la misma SSE, puesto que se obtuvieron las mismas estimaciones. En el modelo 2 se obtuvo una SSE mayor, puesto que la variable edad se excluyó del análisis y la distribución condicional de la respuesta va a tener mayor variabilidad. Puesto que error cuadrático medio (MSE, por sus siglas en inglés) es una estimación de la variancia condicional de la respuesta, es de esperarse que sea mayor cuando se deja alguna variable que es estadísticamente significativa en términos de su capacidad explicativa, pues esto hace que al no estar explícita en el modelo, su ausencia sea capturada en el residuo para explicar esa variabilidad en la respuesta condicional generada por la ausencia de una variable explicativa relevante en la explicación del comportamiento de la variable a priori con la que se cuenta. Al contrario, cuando se incluye dicha variable relevante se logra reducir la variabilidad de la respuesta condicional. Por tanto, se concluye que el incluir o no la variable edad va a reflejarse en tener (o no) una mayor variabilidad condicional y, por consiguiente, los estimadores tendrán mayor o menor precisión.

0.9. El MSE es el segundo momento (sobre el origen) del error, y por lo tanto incorpora tanto la varianza del estimador, así como su sesgo (véase la estructura matemática del operador esperanza para el segundo momento centrado en el origen -por supuesto, sería interesante explicar filosóficamente por qué al estar centrado en el origen incluye el sesgo-). Por ello es evidente que en el caso de un estimador insesgado, el MSE "SCMres" equivale a la varianza del estimador "var_error". Al igual que la varianza, el MSE tiene las mismas unidades de medida que el cuadrado de la cantidad que se estima. En una analogía con la desviación estándar, tomando la raíz cuadrada del MSE produce el error de la raíz cuadrada de la media o la desviación de la raíz cuadrada media (RMSD), que tiene las mismas unidades que la cantidad que se estima; para un estimador insesgado, el RMSE es la raíz cuadrada de la varianza, conocida como la desviación estándar. Véase https://www.wikiwand.com/es/Error_cuadr%C3%A1tico_medio (https://www.wikiwand.com/es/Error_cuadr%C3%A1tico_medio). La varianza del error "var_error" se obtiene como cociente de la suma de cuadrados totales residual SCTR sobre la diferencia entre el tamaño de muestra n y el número de parámetros estimados p; es la varianza residual. Es recomendable estimar los valores "n" y "k_i" mediante las sintaxis de "n=nrow(base)" y "p=length(beta)", respectivamente; en este caso, k_i se expresará como k_1 porque es el k que pertenece al modelo 1).

Hide

```
n=nrow(base); k1=length(beta1)
var_error1=SCTR1/(n-k1)
SCMres1=SCTR1/(n-k1)
round(var_error1,3)
```

```
##       [,1]
## [1,] 0.026
```

n es el tamaño de la muestra y k_i (en este caso k_1) el número de parámetros o coeficientes de regresión a estimar (incluyendo el intercepto). #El valor 0.026 es la estimación de la variancia de la distribución de la variable "velocidad" para cada valor de "edad", "imc", "brazo" y "pierna", así como también valores fijos de los parámetros (calculados a partir del conjunto de datos). Sin importar cuáles sean estos valores la varianza será la misma, siempre que se cumpla el supuesto de que la varianza es constante a lo largo de las observaciones (para ver las diferencias entre homogeneidad de varianza y heterocedasticidad, consúltese https://marxianstatistics.com/2021/10/04/aspectos-conceptuales-generales-del-diseno-experimental-por-bloques/ (https://marxianstatistics.com/2021/10/04/aspectos-conceptuales-generales-del-diseno-experimental-por-bloques/)).

0.10. Puede obtenerse la matriz de varianza-covarianza de los coeficientes de regresión y explique el contenido de la diagonal de dicha matriz. Esto puede realizarse de dos formas: 1. usando la relación $V(\hat{\beta}) = V(Y|X) \times (X^T X)^{-1}$ o 2. con la sintaxis "vcov(mod)".
###PRIMERA FORMA DE CÁLCULO (MANUAL)

Hide

```
varb1=c(var_error1)*solve(t(X)%*%X)
round(varb1,5)
```

```
##                    edad      imc   pierna    brazo
##         0.36244  0.00024 -0.00320 -0.00203 -0.00112
## edad    0.00024  0.00001 -0.00001  0.00000  0.00000
## imc    -0.00320 -0.00001  0.00010  0.00001  0.00001
## pierna -0.00203  0.00000  0.00001  0.00010 -0.00011
## brazo  -0.00112  0.00000  0.00001 -0.00011  0.00017
```

Hide

```
###SEGUNDA FORMA DE CÁLCULO (AUTOMATIZADA)
varb2=vcov(mod1)
round(varb2,5)
```

```
##             (Intercept)     edad      imc   pierna    brazo
## (Intercept)     0.36244  0.00024 -0.00320 -0.00203 -0.00112
## edad            0.00024  0.00001 -0.00001  0.00000  0.00000
## imc            -0.00320 -0.00001  0.00010  0.00001  0.00001
## pierna         -0.00203  0.00000  0.00001  0.00010 -0.00011
## brazo          -0.00112  0.00000  0.00001 -0.00011  0.00017
```

0.11. Es posible obtener los errores estándar de los coeficientes de regresión; debe recordarse que el "error estándar" se define como "la desviación estándar de la distribución de un estimador. Para ello basta con tomar la diagonal de la matriz de varianzas-covarianzas y obtener su raíz cuadrada. Otra forma de obtenerlos es usando la sintaxis "summary" sobre el modelo, de la forma"summary(mod)$coef"; esta última forma además proporciona información adicional sobre el modelo.

Estimación manual

Hide

```
ee=sqrt(diag(varb1))
round(ee,3)
```

```
##          edad    imc pierna  brazo
##  0.602  0.003  0.010  0.010  0.013
```

Hide

```
###Estimación Automatizada
summary(mod1)$coef
```

```
##                 Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept) -0.135421508 0.602032005 -0.2249407 0.8238538974
## edad        -0.019171956 0.003214148 -5.9648635 0.0000031526
## imc          0.015531614 0.010022984  1.5495997 0.1338054038
## pierna      -0.002598893 0.009907203 -0.2623236 0.7952215100
## brazo        0.026261747 0.012982263  2.0228944 0.0539028268
```

Hide

```
round(summary(mod1)$coef[,2],3)
```

```
## (Intercept)        edad         imc      pierna       brazo
##       0.602       0.003       0.010       0.010       0.013
```

0.12. Finalmente, es posible obtener los intervalos de confianza para los coeficientes individualmente considerados. Esto se puede realizar de dos formas: 1. con el método clásico que proviene de la distribución asintótica de los coeficientes: $IC(\beta_j) = \hat{\beta}_j + t_{1-\alpha/2,n-p} \times ee_{\hat{\beta}_j}$, donde el valor de la distribución t se calcula con "qt(0.975,n-p)", o 2. directamente con la función "confint(mod)".

Estimación Manual

Hide

```
t=qt(.975,n-k1)
ic=cbind(beta1-t*ee,beta1+t*ee)
round(ic,4)
```

```
##                [,1]    [,2]
## (Intercept) -1.3753  1.1045
## edad        -0.0258 -0.0126
## imc         -0.0051  0.0362
## pierna      -0.0230  0.0178
## brazo       -0.0005  0.0530
```

Hide

```
###Estimación Automatizada
round(confint(mod1),4)
```

```
##                2.5 %  97.5 %
## (Intercept) -1.3753  1.1045
## edad        -0.0258 -0.0126
## imc         -0.0051  0.0362
## pierna      -0.0230  0.0178
## brazo       -0.0005  0.0530
```

En relación a la interpretación de los intervalos, debe decirse en una primera mirada parece que solamente el intervalo para el coeficiente de la variable “edad” no presenta gran amplitud (en relación a los demás) y esto se refleja no solamente en los valores infinitesimales bajos sino también en que sus dos extremos tienen el mismo signo. Los otros intervalos parecerían reflejar únicamente que en este modelo (el modelo 1) no se ha demostrado una relación importante entre las otras variables y la respuesta promedio, a causa de su amplitud. Si se desea realizar una interpretación del coeficiente de la variable “edad” en términos de que la variable de “edad” aumenta en cinco años, se deben multiplicar los extremos del intervalo de dicha variable por 5. En tal escenario, se concluiría que al aumentar la edad en 5 años y manteniendo constantes “imc”, “brazo” y “piernas”, la velocidad promedio disminuye entre 0.06 y 0.13 m/s.

Hide

```
round(ic[2,]*5,2)
```

```
## [1] -0.13 -0.06
```

0.13. Si se desease conocer con cuál modelo se debe esperar la mejor estimación de la media condicional y en cuál caso la peor estimación, así como también justificar el porqué de la ocurrencia de ello, debe decirse que el modelo que tiene la mayor variabilidad condicional es el que no considera la variable “edad”, ya que el SSR es igual a 0.06. Esto se explica por el hecho de que la edad es una variable muy importante y en cada distribución condicional se tienen unidades muestrales (las personas) con idénticos valores para “imc”, “pierna” y/o “brazo”, pero con diferentes valores para la variable “edad”. Estas diferencias de edad se traducen en diferencias de velocidad. Además, debe decirse preliminarmente (a falta de pruebas de hipótesis más específicas) que los otros dos modelos son muy similares estadísticamente hablando, a pesar de ello el que reduce en mayor medida la variancia condicional es el que incluye las cuatro variables; sin embargo, debe destacarse que el modelo que sólo contempla “edad” tiene una variancia condicional muy parecida a la del modelo de cuatro variables, lo cual es cierta fuerza de evidencia del poder explicativo de la variable “edad”, puesto que cuando las otras variables no

son consideradas el modelo no parece sufrir cambios esenciales en su significancia estadística. Puede tomarse, por ejemplo, el caso de unidades muestrales (personas) con los valores "edad"=22 años, "sexo"=hombre, "imc"=$20kg/m^2$, "pierna"=$100cm$, "brazo"=$70cm$.

Hide

```
y1=predict(mod1,data.frame(edad=22,imc=20,pierna=100,brazo=70))
y2=predict(mod2,data.frame(imc=20,pierna=100,brazo=70))
y4=predict(mod4,data.frame(edad=22))
round(c(y1,y2,y4),2)
```

```
##    1    1    1
## 1.33 1.29 1.47
```

El primer valor indica que la media de velocidad estimada es 1.33 m/s para personas de 22 años, "imc" de 20 kg/m^2 , longitud promedio de la piernas de 100 cm y de brazos de 70 cm, para ambos sexos. El segundo valor indica que la media de velocidad estimada es 1.29 para personas con imc de 20 , longitud promedio de la piernas de 100 y de brazos de 70 , de ambos sexos y todas las edades (mayores de 18). El tercer valor indica que la media de velocidad estimada es 1.47 m/s para personas de 22 años de ambos sexos y las otras medidas variando.

p. Puede obtenerse la variancia del estimador de la media en cada modelo. Debe recordarse que $\hat{V}(\hat{Y}_h) = X_h^T \hat{V}(\hat{\beta}) X_h$ y $\hat{V}(\hat{\beta})$ se pueden estimar con la sintaxis "vcov".

Hide

```
varb1=vcov(mod1)
varb2=vcov(mod2)
varb4=vcov(mod4)
xh1=c(1,22,20,100,70)
xh2=c(1,20,100,70)
xh4=c(1,22)
vy1=t(xh1)%*%varb1%*%xh1
vy2=t(xh2)%*%varb2%*%xh2
vy4=t(xh4)%*%varb4%*%xh4
round(c(vy1,vy2,vy4),3)
```

```
## [1] 0.003 0.007 0.002
```

La mayor variancia es la del modelo que no tiene "edad", lo que significa que la probabilidad de que el valor estimado esté cerca del verdadero promedio es más baja en ese caso que en los otros.

No necesariamente se puede afirmar que en el modelo que no tiene "edad" el valor estimado está más lejos de la verdadera media que en los otros casos, sino que existe mayor probabilidad de que esté más lejos en dicho escenario (en el que no incluye "edad"). Esto es así porque correlación no implica causalidad, que es un aspecto filosófico-científico y no meramente estadístico-matemático.

q. Adicionalmente pueden elaborarse intervalos de confianza del 0.95 para la media de la velocidad en cada uno de los tres modelos antes construidos.

Hide

```
n=nrow(base)
k1=length(mod1$coef); k2=length(mod2$coef); k3=length(mod3$coef)
t1=qt(0.975,n-k1); t2=qt(0.975,n-k2); t3=qt(0.975,n-k3)
k_i=c(k1,k2,k3); t=c(t1,t2,t3); y=c(y1,y2,y3)
ee=sqrt(c(vy1,vy2,vy4))
ic=cbind(y-t*ee,y+t*ee)
round(ic,2)
```

```
##    [,1] [,2]
## 1  1.22 1.45
## 1  1.12 1.47
## 1  1.32 1.48
## 2  1.29 1.52
## 3  1.26 1.61
## 4  1.29 1.46
## 5  1.25 1.48
## 6  1.31 1.66
## 7  1.52 1.69
## 8  1.32 1.55
## 9  1.24 1.59
## 10 1.32 1.48
## 11 1.39 1.62
## 12 1.13 1.48
## 13 1.49 1.65
## 14 1.12 1.35
## 15 1.12 1.47
## 16 1.06 1.22
## 17 0.77 1.00
## 18 0.96 1.31
## 19 1.44 1.61
## 20 1.49 1.72
## 21 1.44 1.78
## 22 1.42 1.59
## 23 1.55 1.78
## 24 1.30 1.64
## 25 1.27 1.43
## 26 1.29 1.52
## 27 0.83 1.18
## 28 1.30 1.46
## 29 0.83 1.06
## 30 1.16 1.51
```

Con 0.95 de nivel de confianza se puede esperar que la media de la velocidad de las personas de 22 años, imc de 20 kg/m^2, longitud promedio de las piernas de 100 cm (para ambos sexos) y longitud promedio de brazos de 70 cm (para ambos sexos) se encuentre entre 1.22 y 1.45 m/s. Las otras esperanzas condicionales se interpretan de forma similar.

r. Si se necesitase estudiar la distribución de la velocidad, por ejemplo, para una persona con las características mencionadas, se debe construir un intervalo que incluya el 95% de las observaciones de la distribución para cada

caso (i.e., un intervalo del tipo mencionado para cada modelo); a estos intervalos se les conoce como intervalos de predicción.

SOBRE LOS TIPOS DE INTERVALOS ESTADÍSTICOS

Además de los intervalos de confianza, también existen otros dos tipos de intervalo. Como se señala en https://statisticsbyjim.com/hypothesis-testing/confidence-prediction-tolerance-intervals/ (https://statisticsbyjim.com/hypothesis-testing/confidence-prediction-tolerance-intervals/), los intervalos (en general) son métodos de estimación en estadísticas que utilizan datos de muestra para producir rangos de valores que probablemente contengan el valor de la población de interés. Por el contrario, las estimaciones puntuales son estimaciones de valor único de un valor de población. De los diferentes tipos de intervalos estadísticos, los intervalos de confianza son los más conocidos. Sin embargo, ciertos tipos de análisis y situaciones requieren otros tipos de rangos que brinden información diferente. Por ejemplo, supóngase que se toma una muestra aleatoria de un producto con la intención de medir su resistencia y el intervalo de confianza al 0.95 va desde 100 hasta 120 unidades. Se puede tener un 0.95 de nivel de confianza en que la resistencia media de toda la población se encuentra dentro de este rango o, dicho de otra forma, se puede tener un 95% de confianza de que el hecho de que la cantidad de unidades que poseen la resistencia media estimada se encuentra entre 100 y 120 no es un resultado producto del azar (de que existen indicios estadísticos de causalidad). Sin embargo, nótese que el nivel de confianza de 0.95 no indica que el 95% de las observaciones se encuentren dentro de este rango. Para sacar ese tipo de conclusión, necesitamos usar un tipo diferente de intervalo. Así, además de los intervalos de confianza e intervalos de tolerancia, existen los denominados intervalos de confianza de la predicción o simplemente “intervalos de predicción”, que son un rango que probablemente (a un nivel de confianza especificado) contiene el valor medio de la variable dependiente dados valores específicos de las variables independientes. Al igual que los intervalos de confianza regulares, estos intervalos proporcionan un rango para el promedio de la población. En este caso, es una población particular definida por los valores de sus variables independientes. De manera similar, estos rangos no dicen nada al investigador sobre la distribución de las observaciones (individualmente consideradas) alrededor de la media de la población. Como se señala en (Statistical Tolerance Regions: Theory, Applications, and Computation, Kalimuthu Krishnamoorthy & Thomas Mathew, 2009, p. 1), “Los intervalos estadísticos calculados en base a una muestra aleatoria tienen una amplia aplicabilidad, con el propósito de cuantificar la incertidumbre acerca de una cantidad escalar asociada con una población muestreada. El tipo de intervalo a calcular depende obviamente del problema subyacente y la aplicación. Se utiliza un intervalo de confianza basado en una muestra aleatoria para proporcionar límites para un parámetro poblacional escalar desconocido, como la media de la población, la desviación estándar, el percentil, la probabilidad de cola, etc. Un intervalo de predicción basado en una muestra aleatoria se utiliza para proporcionar límites para uno. o más observaciones futuras de una población muestreada univariada. Para poblaciones multivariadas, tenemos regiones de confianza y regiones de predicción correspondientes. El tema de este libro es un tercer tipo de intervalo y región, a saber, intervalos de tolerancia y regiones de tolerancia. Para una población univariante, un intervalo de tolerancia es un intervalo, basado en una muestra aleatoria, que se espera que contenga una proporción específica o más de la población muestreada. Una región de tolerancia se define de manera similar para una población multivariante.” Adicionalmente, como se señala en la página 4 de la misma fuente, existen dos tipos de intervalos de tolerancia de dos colas o lados. Uno se construye de modo que contenga al menos una proporción p de la población con una confianza 1 - α, y se denomina simplemente intervalo de tolerancia. Un segundo tipo de intervalo de tolerancia se construye de modo que contenga al menos una proporción p del centro de la población con una confianza 1 - α, y generalmente se denomina intervalo de tolerancia de colas iguales. Así, como puede verificarse,

el intervalo de tolerancia deseado en este caso de aplicación es un intervalo de tolerancia de colas iguales. Puede encontrarse el fundamento técnico-matemático de la metodología de cálculo los intervalos de tolerancia de distribuciones normales en el paquete 'tolerance' en el documento: https://journal.r-project.org/archive/2016/RJ-2016-041/RJ-2016-041.pdf (https://journal.r-project.org/archive/2016/RJ-2016-041/RJ-2016-041.pdf)

Hide

```
SCMres1=anova(mod1)[5,3]
SCMres2=anova(mod2)[4,3]
SCMres4=anova(mod4)[2,3]
vyind1=SCMres1+vy1
vyind2=SCMres2+vy2
vyind4=SCMres4+vy4
ee=sqrt(c(vyind1,vyind2,vyind4))
ic=cbind(y-t*ee,y+t*ee)
round(ic,2)
```

```
##    [,1] [,2]
## 1  0.98 1.68
## 1  0.76 1.83
## 1  1.01 1.80
## 2  1.05 1.75
## 3  0.90 1.97
## 4  0.98 1.77
## 5  1.02 1.72
## 6  0.95 2.02
## 7  1.21 2.00
## 8  1.09 1.79
## 9  0.88 1.95
## 10 1.01 1.80
## 11 1.16 1.86
## 12 0.77 1.84
## 13 1.17 1.96
## 14 0.89 1.59
## 15 0.76 1.83
## 16 0.75 1.54
## 17 0.53 1.23
## 18 0.60 1.67
## 19 1.13 1.92
## 20 1.26 1.96
## 21 1.08 2.14
## 22 1.11 1.90
## 23 1.32 2.02
## 24 0.94 2.00
## 25 0.95 1.74
## 26 1.05 1.75
## 27 0.47 1.54
## 28 0.99 1.78
## 29 0.60 1.30
## 30 0.81 1.87
```

Se espera que la velocidad del 95% de las personas de 22 años, imc de 20 kg/m^2, longitud promedio de la piernas de 100 cm y longitud promedio de brazos de 70 cm (ambas longitudes para ambos sexos) se encuentre entre 0.98 y 1.68 m/s . Si una persona tiene las características mencionadas (edad=22 años, sexo=hombre, imc=20kg/m^2, pierna=100cm, brazo=70cm), y quiere saber cuál es su velocidad, ¿qué información se puede proporcionar a partir de estos resultados? Sobre ello debe decirse que, puesto que se han elaborado tres modelos, hay tres respuestas posibles a la pregunta de la persona interesada; sin embargo, la respuesta más precisa se daría con base en el modelo más preciso, que es aquel que incluye “edad” y excluye “sexo”.

Dos de los diferentes tipos de intervalos expuestos teóricamente antes, pueden obtenerse de forma automática con la ayuda de la sintaxis “predict”. Para ello, debe agregarse “interval=”confidence"” para obtener el intervalo de confianza para la velocidad media e “interval=”prediction"” para el intervalo de confianza de la velocidad individual.

Hide

```
round(predict(mod1,data.frame(edad=22,imc=20,pierna=100,brazo=70),interval="con
fidence"),2)
```

```
##    fit  lwr  upr
## 1 1.33 1.22 1.45
```

Hide

```
round(predict(mod1,data.frame(edad=22,imc=20,pierna=100,brazo=70),interval="pre
diction"),2)
```

```
##    fit  lwr  upr
## 1 1.33 0.98 1.68
```

Los intervalos de tolerancia se estiman en R con el paquete ‘tolerance’ y existen diferentes maneras de calcular los intervalos de tolerancia. Esto es así debido a que existen dos tipos de intervalos de tolerancia (como se vio antes) y al hecho de que el ajuste empírico se realiza según las características fundamentales de la muestra (si se considera como información a priori o a posteriori, del tipo de distribución que se siga). Puesto que aquí se tratará a la variable de respuesta Y (velocidad) como información a priori (puesto que a priori se asumió que se distribuía normalmente), debe señalarse que la investigación empírica demuestra, en el contexto de las prior, que la mayor precisión estadística se alcanza al trabajar con prior construidas desde la filosofía bayesiana objetiva (véase https://www.jstor.org/stable/2291752 (https://www.jstor.org/stable/2291752) y https://www.scielo.br/j/aabc/a/VgvwMxdcKGggLFMt5p4KbHq/?format=pdf&lang=en (https://www.scielo.br/j/aabc/a/VgvwMxdcKGggLFMt5p4KbHq/?format=pdf&lang=en)). Se realiza un ajuste de distribución y, puesto que se ha asumido normalidad en la respuesta, la variable estocástica Y se distribuye a priori normalmente. Los hiperparámetros m0 y n0 (que deben ser mayores que cero) requeridos para configurar la sintaxis “bayesnormtol.int”, que en el contexto de los prior son los parámetros que rigen la distribución de los parámetros de la distribución de la variable Y=“velocidad”, no son cantidades de tamaño de muestra anterior, sino que son cantidades ajustables para reflejar la precisión previa relativa al tamaño de muestra https://journal.r-project.org/archive/2016/RJ-2016-041/RJ-2016- (https://journal.r-

project.org/archive/2016/RJ-2016-041/RJ-2016-) 041.pdf, pág. 204. Si se trabaja con distribuciones a priori, pueden realizar contrastes de normalidad para justificar asumirlas a priori como normales (aunque evidentemente en todos los literales anteriores se ha asumido implícitamente eso). Para ello, se realizarán dos pruebas de normalidad, la de Kolmogórov-Smirnov y la de Shapiro-Wilk.

Hide

```
library(tolerance)
library(fitdistrplus)
attach(base)
```

¿Se comporta normalmente la variable de respuesta? Existen diferentes formas de comprobar esto.

1. Realizar un ajuste de distribución asumiendo normalidad y contrastarlo contra uno o más ajuste de distribución (que no asuma o asuman normalidad) y seleccionar el mejor ajuste de distribución, siguiendo los criterios de máxima verosimilitud, el criterio bayesiano de información (BIC) elaborado por Schwarz y el criterio de información de Akaike (AIC).

Hide

```
fit_normal<-fitdist(base$veloc, "norm")
fit_lnormal<-fitdist(base$veloc, "lnorm")
summary(fit_normal)
```

```
## Fitting of the distribution ' norm ' by maximum likelihood
## Parameters :
##       estimate Std. Error
## mean 1.3732782 0.04384321
## sd   0.2401391 0.03099941
## Loglikelihood:  0.2279484   AIC:  3.544103   BIC:  6.346498
## Correlation matrix:
##              mean           sd
## mean 1.000000e+00 3.772298e-14
## sd   3.772298e-14 1.000000e+00
```

Hide

```
summary(fit_lnormal)
```

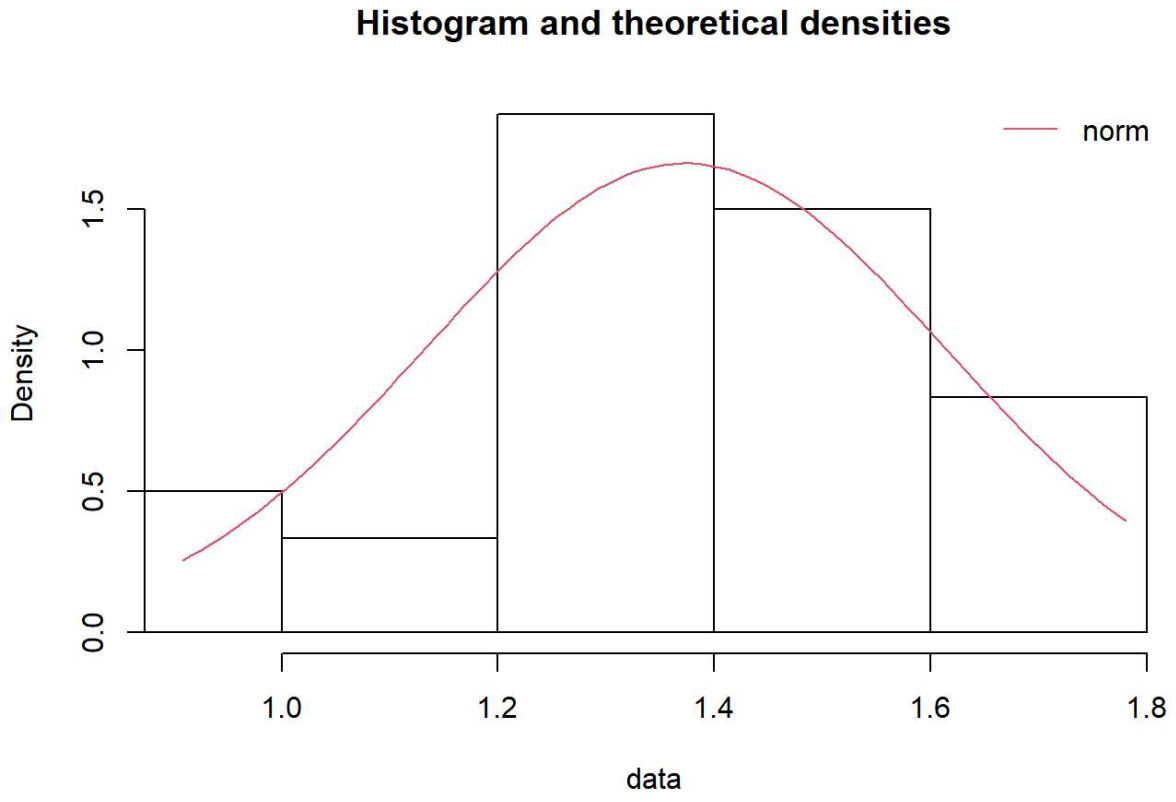
```
## Fitting of the distribution ' lnorm ' by maximum likelihood
## Parameters :
##          estimate Std. Error
## meanlog 0.3008487 0.03366571
## sdlog   0.1843947 0.02380210
## Loglikelihood:  -0.8733177   AIC:  5.746635   BIC:  8.54903
## Correlation matrix:
##               meanlog         sdlog
## meanlog  1.000000e+00 -2.891324e-13
## sdlog   -2.891324e-13  1.000000e+00
```

Sobre la interpretación de los resultados de "fitdist()" debe decirse que la máxima verosimilitud se interpreta en términos de cuál escalar o magnitud producida por la sintaxis en cuestión es mayor, lo cual es evidente por el hecho de que lo que se desea es maximizar la verosimilitud, expresada en el coeficiente o estadístico de prueba obtenido. Sobre AIC y BIC debe decirse que si bien ambos buscan maximizar matemáticamente la verosimilitud (o siendo más precisos, buscan maximizar el logaritmo de la función de verosimilitud, véase https://marxianstatistics.files.wordpress.com/2020/12/sobre-los-estimadores-de-bayes-el-analisis-de-grupos-y-las-mixturas-gaussianas-isadore-nabi.pdf (https://marxianstatistics.files.wordpress.com/2020/12/sobre-los-estimadores-de-bayes-el-analisis-de-grupos-y-las-mixturas-gaussianas-isadore-nabi.pdf)), estos criterios introducen el coeficiente del estadístico de prueba (el escalar real que expresa logaritmo de la máxima verosimilitud del conjunto de datos) en fórmulas (diferentes, y por ello difieren los criterios -los argumentos dados para ello por sus creadores son puramente filosóficos, véase https://es.wikipedia.org/wiki/Criterio_de_informaci%C3%B3n_bayesiano- (https://es.wikipedia.org/wiki/Criterio_de_informaci%C3%B3n_bayesiano-), aunque tanto su estimación como su interpretación sea en el mismo sentido) cuyo resultado se interpreta en sentido inverso: el coeficiente AIC o BIC con menor valor es el deseado. A pesar de ello, existe una cantidad de casos no trivial en la que los coeficientes difieren y, según motivos que en última instancia son más filosóficos que técnicos, puede escogerse uno u otro.

2. Se realiza un análisis gráfico (evidentemente de carácter exploratorio) sobre el ajuste de la respuesta Y Comparando el histograma de distribución de frecuencias con la densidad normal teórica mediante la sintaxis "denscomp()"

Hide

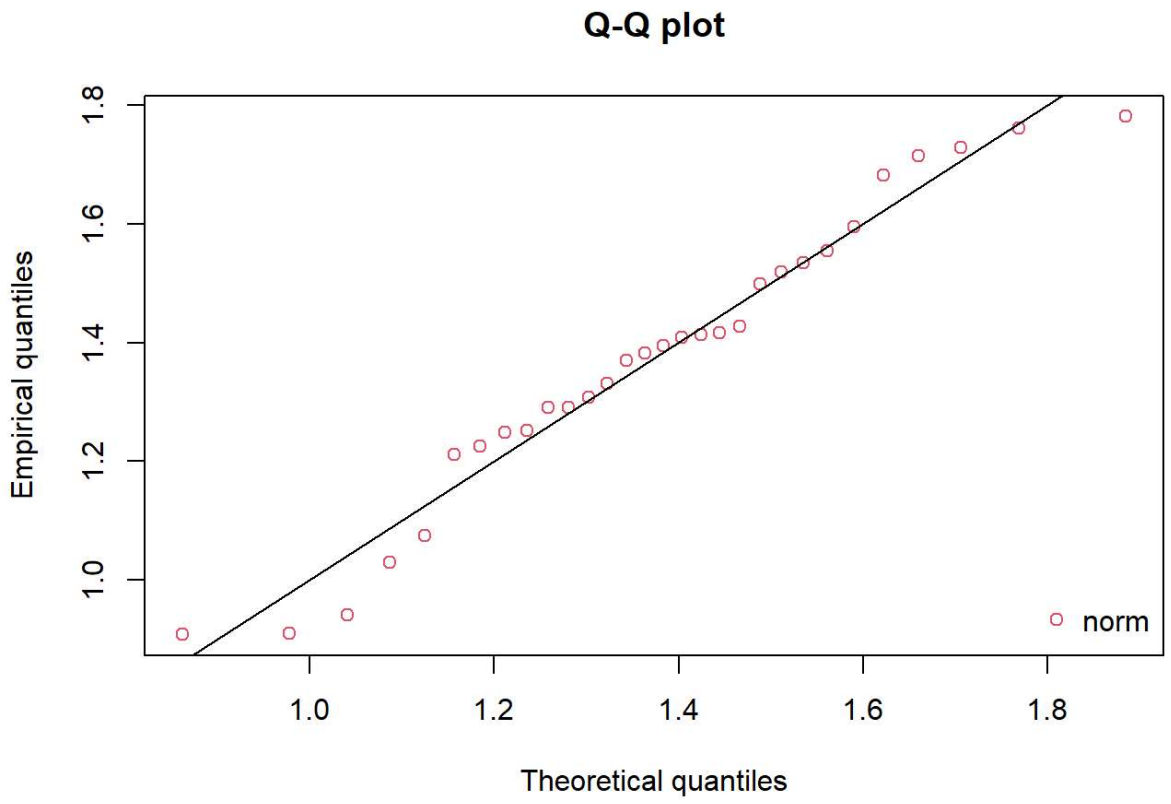
```
plot.legend <- c("Distribución Normal Teórica vs Distribución Empírica de la Re
spuesta") #Si se quisiese añadir una leyenda a la gráfica
denscomp(list(fit_normal))
```



2.1. Gráfico Q-Q: Como se señala en https://www.wikiwand.com/en/Q%E2%80%93Q_plot (https://www.wikiwand.com/en/Q%E2%80%93Q_plot), una gráfica Q – Q (cuantiles-cuantiles) es una gráfica de probabilidad, que es un método gráfico para comparar dos distribuciones de probabilidad trazando sus cuantiles entre sí. Es un método gráfico para el diagnóstico de diferencias entre la distribución de probabilidad de una población de la que se ha extraído una muestra aleatoria y una distribución usada para la comparación mediante el contraste de sus cuantiles.

Hide

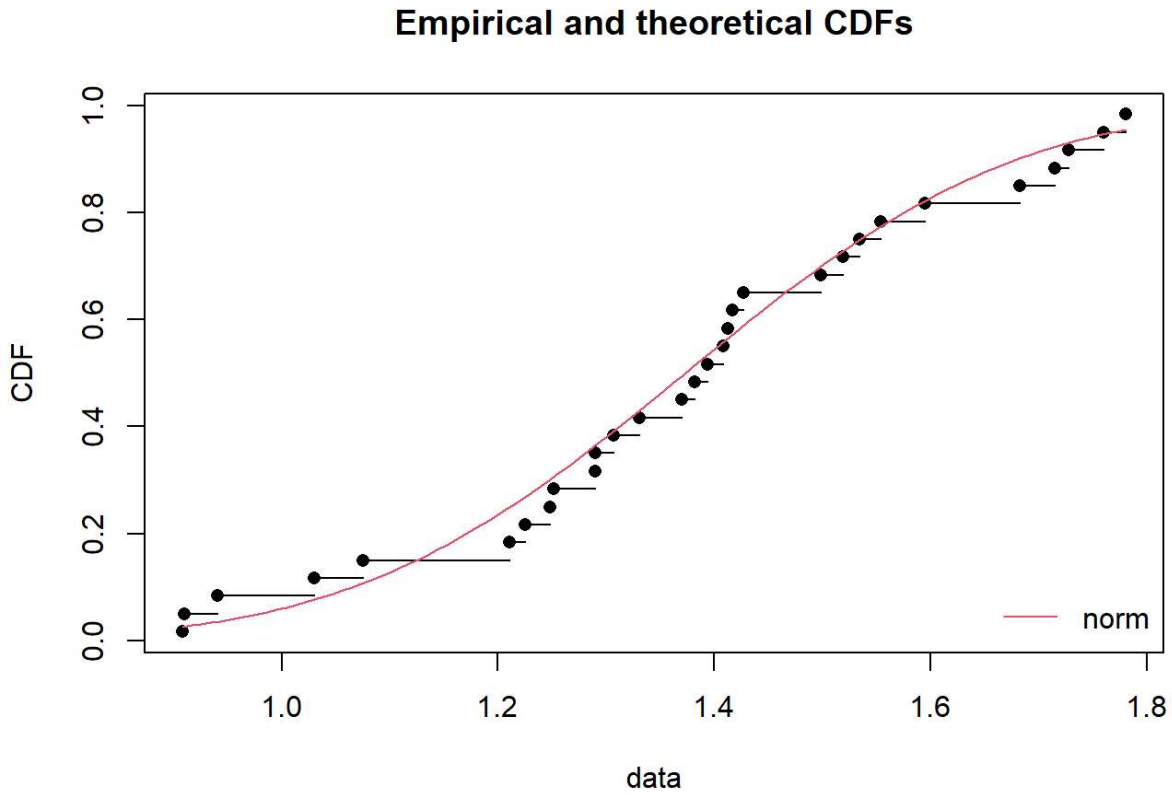
```
qqcomp(list(fit_normal))
```



2.2. Comparando el comportamiento gráfico de la distribución de probabilidad acumulada empírica versus la distribución de probabilidad acumulada normal teórica

Hide

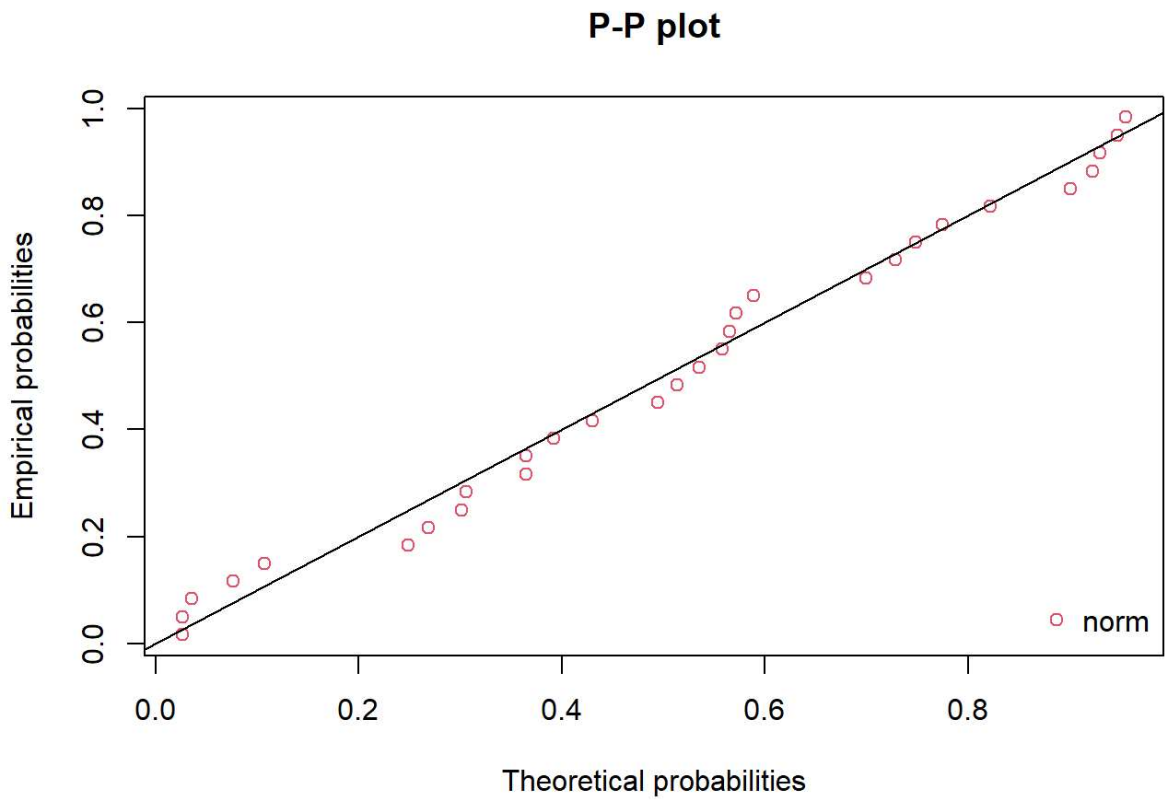
```
cdfcomp(list(fit_normal))
```



2.3. Gráfico P-P: Como se verifica en https://www.ibm.com/docs/es/spss-statistics/SaaS?topic=SSLVMB_sub/statistics_mainhelp_ddita/spss/base/idh_gnpp.html (https://www.ibm.com/docs/es/spss-statistics/SaaS?topic=SSLVMB_sub/statistics_mainhelp_ddita/spss/base/idh_gnpp.html), el gráfico de probabilidad es una técnica gráfica, utilizada para contrastar la distribución de un conjunto de datos. Permite comparar la distribución empírica de una muestra de datos, con alguna distribución de referencia.

Hide

```
ppcomp(list(fit_normal))
```



3. Contrastes de Normalidad Kolmogórov-Smirnov y Shapiro-Wilk 3.1. Contraste de Shapiro-Wilk para verificar normalidad

Hide

```
shapiro.test(base$veloc)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  base$veloc
## W = 0.96187, p-value = 0.3454
```

Hide

```
velocidad<-rnorm(1.000000e+00,3.772298e-14) #
```

Generando una distribución normal con números aleatrorios (con la media y la desviación especificada en el ajuste por máxima verosimilitud antes realizado), debido a que será utilizada en el contraste de Kolmogórov-Smirnov a causa de que la metodología misma del contraste exige una comparación directa contra otra distribución que sea normal. #3.1. Contraste de Kolmogórov-Smirnov para verificar normalidad

Hide

```
ks.test(base$veloc,velocidad)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  base$veloc and velocidad
## D = 1, p-value = 0.2879
## alternative hypothesis: two-sided
```

4. Una vez verificada la normalidad de las variables se puede realizar la estimación de los intervalos de tolerancia, considerando a Y una distribución normal a priori.

Se realiza un ajuste de distribución y, puesto que se ha asumido normalidad en la respuesta, la variable estocástica Y se distribuye a priori normalmente. ¿Qué es una distribución a priori (de ahora en adelante, prior)? Como se señala en https://www.wikiwand.com/en/Prior_probability (https://www.wikiwand.com/en/Prior_probability), en la inferencia estadística bayesiana, una distribución de probabilidad previa, a menudo llamada simplemente anterior, de una cantidad incierta es la distribución de probabilidad que expresaría las creencias de uno sobre esta cantidad antes de que se tenga en cuenta alguna evidencia. El teorema de Bayes calcula el producto puntual renormalizado de la función previa y de verosimilitud, para producir la distribución de probabilidad posterior, que es la distribución condicional de la cantidad incierta dados los datos. De manera similar, la probabilidad previa de ocurrencia de un evento aleatorio o de una proposición incierta es la probabilidad incondicional que se asigna a dicho evento antes de que se tenga en cuenta cualquier evidencia relevante sobre su comportamiento. Un prior se puede determinar a partir de información pasada, como experimentos previos. Un prior puede obtenerse de la evaluación puramente subjetiva de un experto experimentado (lo que evidentemente relativiza mucho lo subjetivo de su opinión). Se puede crear un prior no informativo para reflejar un equilibrio entre los resultados cuando no hay información disponible. Los priores también se pueden elegir de acuerdo con algún principio, como la simetría o la maximización de la entropía dadas las restricciones; ejemplos son el anterior de Jeffreys o el anterior de referencia de Bernardo. Cuando existe una familia de prior conjugados, elegir un prior de esa familia simplifica el cálculo de la distribución posterior. Un prior se puede determinar a partir de información pasada, como experimentos previos. Un prior puede obtenerse de la evaluación “puramente subjetiva” de un experto experimentado. Se puede crear un prior no informativo para reflejar un equilibrio entre los resultados cuando no hay información disponible. Los prior también se pueden elegir de acuerdo con algún principio, como la simetría o la maximización de la entropía dadas las restricciones; ejemplos son el anterior de Jeffreys o el prior de referencia de Bernardo. Cuando existe una familia de prior conjugados, elegir un prior de esa familia simplifica el cálculo de la distribución posterior.

Hide

```
bayesnormtol.int(x = veloc, alpha = 0.01, P = 0.95,
                 side = 1, method = "HE",
                 hyper.par = list(mu.0 = 1.000000e+00,
                                  sig2.0 = .772298e-14, n.0 = 1.000000e+00, m.0
= .772298e-14))
```

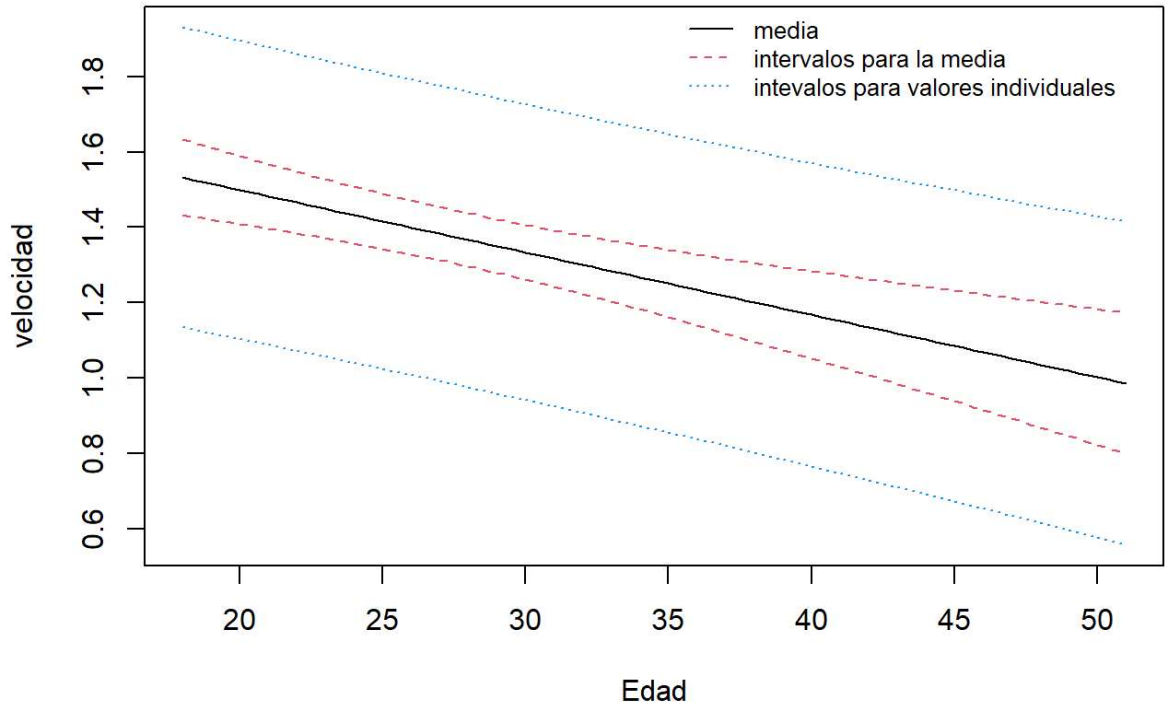
```
##   alpha    P 1-sided.lower 1-sided.upper
## 1  0.01 0.95     0.7245638       1.99791
```

Como se adelantó, una prior se puede determinar a partir de información pasada, como experimentos previos. Así, la validez de configurar los hiperparámetros en la sintaxis "bayesnormtol.int" se justifica por el hecho de que, como es conocido, teóricamente la igualdad de parámetros antes mencionada implica que los parámetros de la distribución subyacente son iguales a los de la distribución estudiada. En la sintaxis "bayesnormtol.int", mu.0 es la media de la distribución observable y n.0 es la media de la distribución subyacente, mientras que sig2.0 es la varianza de la distribución observable y m.0 es la varianza de la distribución subyacente. Aquí se asumirá que existe convergencia en el primer y el segundo momento de la distribución observable (prior) y la distribución subyacente (la distribución de los parámetros de la distribución prior), lo cual se justifica por el hecho que si las medias r-ésimas (los r-ésimos estadísticos de prueba) son únicas y existe convergencia en distribución entre las muestras en comparación distribución, estas tendrán también las mismas medias r-ésimas (véase https://marxianstatistics.com/2021/10/07/una-aproximacion-teorica-a-la-determinacion-de-la-igualdad-de-varianzas-de-dos-poblaciones/ (https://marxianstatistics.com/2021/10/07/una-aproximacion-teorica-a-la-determinacion-de-la-igualdad-de-varianzas-de-dos-poblaciones/)). Si se asume a la distribución subyacente como normal (lo cual no es en lo absoluto descabellado debido a que la respuesta estudiada se muestra normal en los distintos tipos de contrastes) y sabiendo que todos los momentos de la distribución normal existen (porque existe su función característica y esta sólo puede existir si todos sus momentos existen -y todos sus momentos existen porque sin importar su orden se pueden expresar como combinaciones lineales de momentos inferiores, véase http://srabbani.com/moments.pdf- (http://srabbani.com/moments.pdf-)), entonces sus momentos son idénticos. En relación a la diferencia entre una prueba de una cola y otra de dos colas, se señala en https://help.xlstat.com/s/article/cual-es-la-diferencia-entre-una-prueba-de-dos-colas-bilateral-y-de-una-cola-unilateral?language=es (https://help.xlstat.com/s/article/cual-es-la-diferencia-entre-una-prueba-de-dos-colas-bilateral-y-de-una-cola-unilateral?language=es), una prueba "de dos colas" se asocia a una hipótesis alternativa para la cual se desconoce el signo de la potencial diferencia, mientras que una prueba "de una cola" está asociada a una hipótesis alternativa para la cual se conoce el signo de la potencial diferencia antes de ejecutar el experimento y la prueba.

r. Para observar gráficamente el comportamiento de las estimaciones en términos de los intervalos de confianza, se puede construir un escenario de estimación en el que un determinado modelo predictivo estime valores generados de manera pseudoaleatoria (que desempeñarían la función que desempeñan los "datos de prueba" en el campo del Machine Learning -diferente de los datos de entrenamiento-) dentro de un intervalo generado con la edad mínima y la edad máxima de la muestra real que se ha venido estudiando. Posteriormente, graficar este escenario permitiría estudiar geométricamente el comportamiento de las predicciones realizadas, en términos de los intervalos de confianza. Se escogerá para ello un modelo que emplee únicamente una predictora, ello con la finalidad de que no sea necesario usar una gráfica de tres dimensiones. Así, el modelo "mod4" contendrá como único parámetro (además del intercepto) al correspondiente a la variable explicativa "edad", que fue la que se registró a lo largo de las pruebas que permitía un mejor ajuste. Una vez construido el algoritmo, pueden graficarse los intervalos de confianza con ayuda de la sintaxis "matplot". Es recomendable el uso de colores para diferenciar los límites de las medias y los límites de los valores individuales, así como también agregar una guía con "legend" para reconocer las líneas representadas. Esto con la finalidad de estudiar gráficamente el comportamiento de las predicciones en términos de sus intervalos de confianza.

Hide

```
mod4=lm(veloc~edad)
Edad=seq(18,51,length=100)
ic=matrix(nrow=100,ncol=5)
for(i in 1:100){
  ic[i,1:3]=predict(mod4,data.frame(edad=Edad[i]),interval="confidence")
  ic[i,4:5]=predict(mod4,data.frame(edad=Edad[i]),interval="prediction")[-1]
}
matplot(Edad,ic,type="l",lty=c(1,2,2,3,3),col=c(1,2,2,4,4),ylab="velocidad")
legend(35,2,c("media","intervalos para la media","intevalos para valores indivi
duales"),bty="n",
       lty=c(1,2,3),col=c(1,2,4),cex=0.8)
```



s. Lo que se dijo antes sobre MS y MSE es, como sus abreviaturas lo indican, relativo a los valores promedio (la "M" es de "Mean"). Sin embargo, también pueden realizarse estimaciones relativas al error cuadrático total y no al error cuadrático medio o promedio, lo cual se conoce como suma de cuadrados totales ("sqt" aquí para fines de cálculo, es decir, es la varianza residual total); suma de cuadrados ajustados en el caso de los relativos a la regresión y simplemente suma de cuadrados cuando se trate del valor a priori de la respuesta -del valor de la respuesta sin considerar información adicional-). Así, es posible obtener la suma de cuadrados total (que es el total de la suma de todos los errores cuadrados localizados en la segunda columna de la tabla ANOVA que genera R con la sintaxis "anova()"). A continuación, se presenta la suma de cuadrados total para la variable Y considerada a priori; en el literal t se abordará el caso a posteriori. Esta suma de cuadrados total a priori se abreviará aquí como "sqt", mientras que (como se adelantó en el literal o.7.1.6) la suma de cuadrados totales de Y (a priori) se denota aquí para fines de cálculo como "sqresp".

Hide

```
SCTY1=sum(anova(mod1)[,2])
SCTY2=sum(anova(mod2)[,2])
SCTY4=sum(anova(mod4)[,2])
c(SCTY1,SCTY2,SCTY4)
```

```
## [1] 1.730004 1.730004 1.730004
```

La razón por la que se multiplica por (n-1) obedece a que ya se utilizó 1 grado de libertad (observación) para estimar la varianza y, de forma general, se puede explicar de la siguiente manera. Usando el concepto de campo probabilístico (acuñado por Kolmogórov) es fácil comprenderlo (el campo está metido dentro de un espacio de Lebesgue). Por definición se describe el campo como tal en términos de parámetros de forma, localización y escala. Eso significa que cada uno de esos parámetros caracteriza al sistema de probabilidades de alguna manera y, por consiguiente, su localización dentro del campo (y si es localizable es porque existe, entonces el ser sólo es ser en cuanto ser localizado, dasein o existencia), es única (no puede tener dos medias una distribución normal, por ejemplo), por lo que si se utilizan las mismas coordenadas para localizarlo sería una contradicción lógica y, por consiguiente, se requiere hacer algún tipo de ajuste. ¿Qué tipo de ajuste? Pues resulta que el campo se generó a partir de un determinado número de elementos (observaciones, puntos de datos, etc.) y que cada combinación de estos elementos (en las que sí importa el orden) genera una coordenada dentro de dicho campo, como no existen más observaciones disponibles (n es fijo y finito -sino no existirían los momentos, en términos de análisis matemático porque significaría que el operador esperanza no converge en el infinito o muestra grande-) y, por consiguiente, inexorablemente para dar una coordenada diferente se requiere considerar (n-k) elementos, siguiendo el principio de mínima acción (la naturaleza y, por consiguiente, los sistemas en general, son económicos en sus movimientos) entonces se requiere únicamente considerar (para el caso de una segunda estimación relativa a un parámetro o hiperparámetro (n-1) elementos, para una tercera estimación n-2 elementos y así sucesivamente hasta llegar a (n-k) elementos, donde n>k.

t. El error cuadrático medio "SCreg" para cada modelo (el otro caso mencionado sobre la suma total al cuadrado que solamente se mencionó en el literal s) puede estimarse y compararse con la suma de cuadrados totales de Y para evaluar cuánto de la suma de cuadrados de la media no-condicional de Y (obtenida con la sintaxis "mean(veloc)") es explicada por las medias condicionales Y_i de cada uno de los modelos evaluados (que conceptualmente representan nueva información disponible sobre el evento, que es lo que establece el teorema de Bayes).

Hide

```
m=mean(veloc)
fit1=predict(mod1)
fit2=predict(mod2)
fit4=predict(mod4)
SCreg1=sum((fit1-m)^2) #suma total de errores al cuadrado de todas las diferenc
ias entre el valor observado y el estimado elevadas al cuadrado para el modelo
 1
SCreg2=sum((fit2-m)^2)
SCreg4=sum((fit4-m)^2)
round(c(SCreg1,SCreg2,SCreg4),2)
```

```
## [1] 1.08 0.17 0.74
```

Clasificándolos jerárquicamente desde el que más se acerca a la suma de cuadrados total (SCT) no-condicional hasta el que menos, el primer modelo tiene una SCT de 1.08, el cuarto modelo de 0.74 y el segundo de 0.17. El primer modelo tiene una SCT que se acerca más a la SCT no-condicional porque toma en consideración una mayor cantidad de variables (esto se justifica por la estructura matemática de la SCT, véase https://www.wikiwand.com/en/Residual_sum_of_squares (https://www.wikiwand.com/en/Residual_sum_of_squares)); el tercero, aunque considera únicamente una variable, al tratarse de "edad" (que se vio antes tiene un poder explicativo de la velocidad de los nadadores superior en relación a sus pares) explica más que el segundo, a pesar que este tiene tres variables.

u. Puede obtenerse la suma de cuadrados totales residuales "SCTR" de cada modelo modelo, así como también la suma de cuadrados totales de Y (a priori y, por consiguiente, es un valor constante si Y es fija) "SCTY" con finalidad de estimar manualmente con estos dos escalares el coeficiente de determinación de Pearson S_R=1-SCTR/SCTY, para luego mostrar complementariamente cómo se hace de forma automatizada en R.

Hide

```
anova(mod1)
```

```
## Analysis of Variance Table
##
## Response: veloc
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## edad       1 0.73979 0.73979 28.6666 1.495e-05 ***
## imc        1 0.02264 0.02264  0.8773  0.357892
## pierna     1 0.21680 0.21680  8.4008  0.007697 **
## brazo      1 0.10560 0.10560  4.0921  0.053903 .
## Residuals 25 0.64517 0.02581
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hide

```
anova(mod2)
```

```
## Analysis of Variance Table
##
## Response: veloc
##           Df  Sum Sq  Mean Sq F value Pr(>F)
## imc        1 0.00014 0.000135  0.0022 0.9625
## pierna     1 0.07413 0.074131  1.2329 0.2770
## brazo      1 0.09237 0.092374  1.5362 0.2262
## Residuals 26 1.56336 0.060129
```

Hide

```
anova(mod4)
```

```
## Analysis of Variance Table
##
## Response: veloc
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## edad       1 0.73979 0.73979  20.919 8.888e-05 ***
## Residuals 28 0.99021 0.03536
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hide

```
SCMres1=anova(mod1)[5,3]
SCMres2=anova(mod2)[4,3]
SCMres4=anova(mod4)[2,3]
round(c(SCMres1,SCMres2,SCMres4),2)
```

```
## [1] 0.03 0.06 0.04
```

Hide

```
SCTR1=anova(mod1)[5,2]
SCTR2=anova(mod2)[4,2]
SCTR4=anova(mod4)[2,2]
round(c(SCTR1,SCTR2,SCTR4),2)
```

```
## [1] 0.65 1.56 0.99
```

Hide

```
#Estimación de SCTY
SCTY1=sum(anova(mod1)[,2])
SCTY2=sum(anova(mod2)[,2])
SCTY4=sum(anova(mod4)[,2])
c(SCTY1,SCTY2,SCTY4)
```

```
## [1] 1.730004 1.730004 1.730004
```

Hide

```
S_R1=1-SCTR1/SCTY1
S_R2=1-SCTR2/SCTY2
S_R4=1-SCTR4/SCTY4
round(c(S_R1,S_R2,S_R4),2)
```

```
## [1] 0.63 0.10 0.43
```

Como puede verificarse, el primer modelo produce un R^2 más alto, lo cual indica que las cuatro variables en conjunto explican el 63% de la variabilidad de la respuesta, mientras que la edad por sí sola explica un 43%. Las otras tres variables que no son edad logran explicar por sí mismas el 10%, lo cual reafirma la evidencia encontrada antes respecto de la importancia de la edad y la poca relevancia estadística de las otras variables a nivel explicativo. El coeficiente de determinación puede también obtenerse de forma automática:

Hide

```
round(c(summary(mod1)$r.sq,summary(mod2)$r.sq,summary(mod4)$r.sq),2)
```

```
## [1] 0.63 0.10 0.43
```

v. Puede procederse a encontrar los efectos marginales de las variables explicativas vinculadas a la respuesta.

SOBRE LOS EFECTOS MARGINALES

Como se señala en https://www.statisticshowto.com/marginal-effects/ (https://www.statisticshowto.com/marginal-effects/), los efectos marginales explican cómo cambia una variable dependiente o de respuesta cuando cambia una variable independiente o explicativa específica. Se supone que las demás variables explicativas se mantienen constantes. Los efectos marginales a menudo se calculan al analizar los resultados del análisis de regresión. Debe decirse que los efectos marginales de las variables binarias miden cambios discretos, mientras que para las variables continuas miden la tasa de cambio instantánea (i.e., su derivada). De hecho, para una variable independiente X, podemos definir el efecto marginal como la derivada parcial, con respecto a X, de la función de predicción f. Existen distintos tipos de efectos marginales: 1) Efecto marginal promedio (AME, por su nombre en inglés). Como sugiere el nombre, puede pensarse en el AME como una “derivada promedio”. Para encontrar el AME, se debe calcular el efecto marginal de cada variable X para cada observación (teniendo en cuenta las covariables -las demás variables explicativas-) y, finalmente se calcula el promedio. 2) Efecto marginal en la media (MEM). Este efecto marginal es muy similar al AME, excepto que, en lugar de tomar sus valores observados, las covariables se mantienen en sus valores medios. 3) Efectos marginales a valores representativos (MER). La diferencia aquí es que se eligen valores representativos (es decir, valores de interés en su experimento o estudio) para las covariables de la variable X. Debe mencionarse que el orden en el que se ingresen las variables explicativas al especificar las variables involucradas en la regresión (por ejemplo, lm(veloc~brazo+pierna)) indica unívocamente el orden en su accionar, i.e., el orden en el que imprimen su efecto marginal sobre la respuesta o variable dependiente.

v.1. Por ejemplo, puede determinarse el aporte marginal de las variables “imc”, “pierna” y “brazo” asumiendo constante “edad”, cuantificado a través de la suma de cuadrados de regresión marginal. La “SCTY” es fija pues es la variabilidad en Y no depende de los predictores. Con más

predictores el R^2 casi siempre crece y nunca decrece. Esta estimación se suele definir en los contextos de Bioestadística como la suma de cuadrados de regresión marginal de (para este caso) “imc, pierna y brazo dada la edad” y, en general, es la suma de cuadrados totales correspondiente a la variable explicativa cuyo efecto marginal se desea analizar. Este valor marginal es presentado en tercera columna (en la fila correspondiente a la variable) de la tabla construida por el programa R al usar la sintaxis “anova()”. Para facilitar su lectura, puede extraerse el valor correspondiente mediante la sintaxis “c()” y colocarse en un vector numérico de un elemento, como se presenta a continuación. v.1. Puede determinarse, por ejemplo, el aporte marginal de las variables “imc”, “pierna” y “brazo” asumiendo constante “edad”, cuantificado a través de la suma de cuadrados de regresión marginal. Puesto que la “SCTY” es fija pues es la variabilidad en Y no depende de los predictores, con más predictores el R^2 casi siempre crece y nunca decrece. Esta estimación se suele definir en los contextos de Bioestadística como la suma de cuadrados de regresión marginal de (para este caso) “imc, pierna y brazo dada la edad” y, en general, es la suma de cuadrados totales correspondiente a la variable explicativa cuyo efecto marginal se desea analizar. Este valor marginal es presentado en tercera columna (en la fila correspondiente a la variable) de la tabla construida por el programa R al usar la sintaxis “anova()”. Para facilitar su lectura, puede extraerse el valor correspondiente mediante la sintaxis “c()” y colocarse en un vector numérico de un elemento, como se presenta a continuación.

Hide

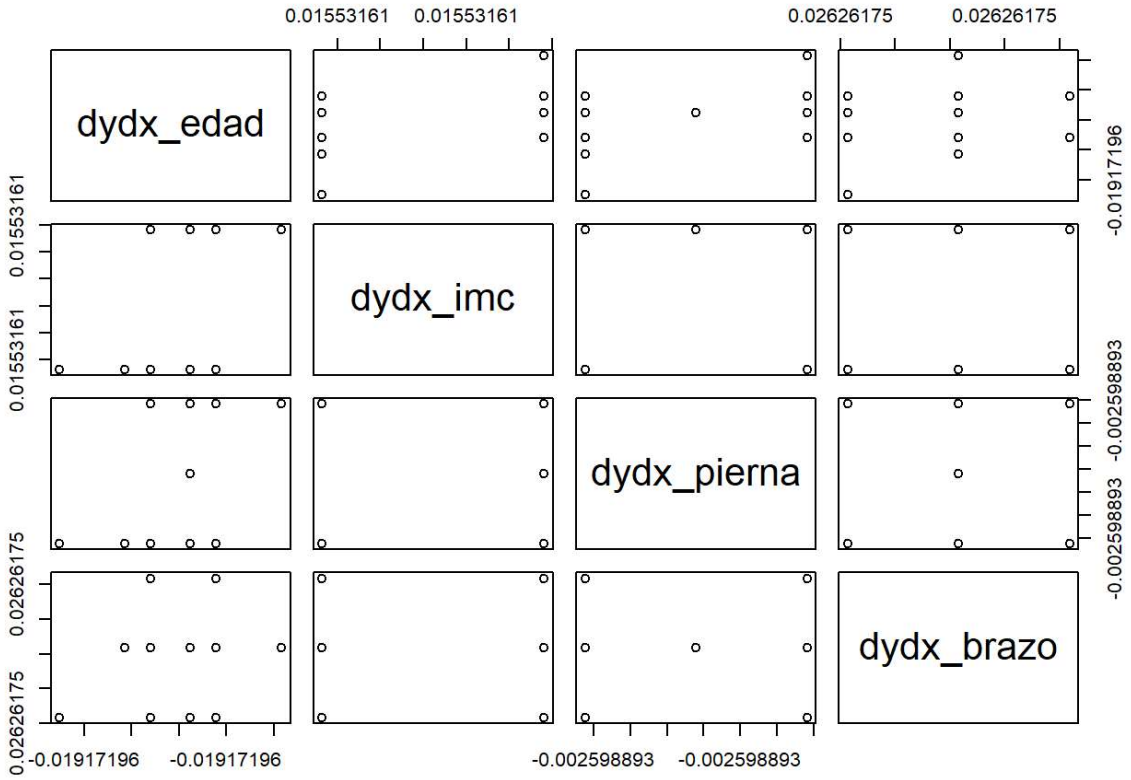
```
Efecto_Marginal_edad=c(anova(mod1)[1,2])
Efecto_Marginal_imc=c(anova(mod1)[2,2])
Efecto_Marginal_pierna=c(anova(mod1)[3,2])
Efecto_Marginal_brazo=c(anova(mod1)[4,2])
```

El aporte marginal de “pierna” es 0.073.

v.2. También se puede estimar el efecto marginal de cada observación de cada una de las variables explicativas a cada observación de la respuesta, por ejemplo, para el caso del modelo 1. Para ello, se puede utilizar la sintaxis presentada a continuación.

Hide

```
library(margins)
margins=marginal_effects(mod1,variables=NULL)
plot(
  margins
)
```



v.3.2. En general, el aporte marginal de X_i puede obtenerse manualmente como la diferencia entre el modelo que considera únicamente la variable (o conjunto de) que antecede en efecto marginal a la variable explicativa cuyo efecto marginal se desea analizar y el modelo que considera únicamente a la variable. Nótese que, por ejemplo, si se obtiene el aporte marginal de "pierna" después del efecto marginal de "brazo", se verifica que dicho resultado se reporta también en la tabla ANOVA del modelo 5 ("veloc~brazo+pierna") y ello es la consecuencia natural las implicaciones del orden de entrada de las variables explicativas del modelo antes mencionado. Se presentará antes de ello un listado de todos los modelos que se han construido en las diferentes secciones de este documento, con la finalidad de evitar confusiones o retrocesos innecesarios dentro de este documento.

Hide

```
mod1=lm(veloc~edad+imc+pierna+brazo)
mod2=lm(veloc~imc+pierna+brazo)
mod3=lm(veloc~edad+imc+pierna+brazomm)
mod4=lm(veloc~edad)
mod5=lm(veloc~brazo+pierna)
mod6=lm(veloc~brazo)
mod7=lm(veloc~pierna)
anova(mod6)[2,2]-anova(mod5)[3,2]
```

```
## [1] 0.01660077
```

Hide

```
anova(mod5)
```

```
## Analysis of Variance Table
##
## Response: veloc
##           Df Sum Sq  Mean Sq F value Pr(>F)
## brazo      1 0.1484 0.148403  2.5603 0.1212
## pierna     1 0.0166 0.016601  0.2864 0.5969
## Residuals 27 1.5650 0.057963
```

El aporte marginal de "pierna", tras considerar "brazo" es aproximadamente de 0.017, el cual es evidentemente menor que cuando se analiza sin considerar efectos marginales antecediendo a los suyos. Esto se debe a la alta correlación existente entre pierna y brazo, que además implica que no existe independencia lineal entre las variables regresoras, es decir, que el modelo posee multicolinealidad, como se vio antes.

Complementariamente debe decirse que también pueden obtenerse las sumas de cuadrados totales residuales relativas a la variable en cuestión a través de la matriz de residuos o errores, como se vio anteriormente.

Hide

```
r1=mod1$res; SCTR1=t(r1)%*%r1
r2=mod2$res; SCTR2=t(r2)%*%r2
r3=mod3$res; SCTR3=t(r3)%*%r3
r4=mod4$res; SCTR4=t(r4)%*%r4
r5=mod5$res; SCTR5=t(r5)%*%r5
r6=mod6$res; SCTR6=t(r6)%*%r6
r7=mod7$res; SCTR7=t(r7)%*%r7
SCTR6-SCTR5
```

```
##            [,1]
## [1,] 0.01660077
```

w. Adicionalmente, pueden compararse el modelo que contiene todas las variables (mod1) contra el modelo que tiene únicamente "edad" (mod4). Para ello debe plantearse antes la hipótesis nula y, por consiguiente, la alternativa. Existen dos formas de llevar a cabo esto, la primera es realizar manualmente la prueba F, mientras que la segunda es realizarlo de forma automatizada.

GENERALIDADES SOBRE EL VALOR p (véase Molina Arias, M. ¿Qué significa realmente el valor de p? Revista de Pediatría de Atención Primaria. 2017; 19:377-81):

1. El valor p no representa la probabilidad de que la hipótesis nula sea cierta, de hecho, se parte del supuesto de que la hipótesis nula es verdadera y es bajo ese supuesto que se estima el valor p.
2. El valor p está relacionado en sentido inverso a la fiabilidad del estudio, puesto que los resultados obtenidos tras el uso de un modelo estadístico son (estadísticamente) más fiable a medida que el valor p es menor.
3. La hipótesis nula bajo la cual se estima el valor p plantea de forma general que no existe diferencia estadística real entre los parámetros de la población de la que proceden las muestras estudiadas y los estadísticos de prueba obtenidos a través del estudio inferencial

de dichas muestras; las demás formas de expresar la hipótesis nula son variaciones de la forma general antes expuesta.

4. El valor p indica la probabilidad de obtener resultados semejantes a los obtenidos si el experimento se realizara en reiteradas ocasiones bajo las mismas condiciones o, dicho en otras palabras, de que los resultados obtenidos sean un mero producto del azar; por supuesto, hay muchos factores que pueden intervenir en los resultados que pudiesen ser obtenidos al repetir el experimento, además del hecho de que exista (o no) una diferencia estadística real entre los estadísticos de prueba y los parámetros, como lo son el tamaño de la muestra, la volatilidad de la variable medida, el tamaño del efecto, la distribución de probabilidad empleada y otros aspectos involucrados, que pueden variar (o no) en función del contexto de estudio.

5. Si el valor p es menor a un determinado nivel de significancia especoficado por el investigador (determinado este último en función de la información histórica disponible y/o de el criterio de un experto experimentado), significa que no se tiene la confianza necesaria como para poder negar que la diferencia observada (se plantee esta diferencia como sea y en el sentido que sea) sea resultado del azar.

La hipótesis nula es:

$H_0 : \beta_2 = \beta_3 = \beta_4 = 0$

La hipótesis alternativa es:

H_A : al menos uno es diferente de 0

w.1. Obtención manual del valor F y su probabilidad asociada

w.1.1. Valor F Crítico (F como estadístico de prueba estimado) Como se señala en https://www.wikiwand.com/en/F-test (https://www.wikiwand.com/en/F-test) y https://www.statisticshowto.com/probability-and-statistics/f-statistic-value-test/ (https://www.statisticshowto.com/probability-and-statistics/f-statistic-value-test/), el estadístico F es el resultado de dividir la variabilidad media local o al interior del grupo de interés (no debe confundirse con la varianza, que es el segundo momento -con centro en la media- de una distribución y como tal es la variabilidad global del modelo -el error cuadrático medio MSE-) entre la variabilidad global o existente entre los grupos (abreviada aquí como el “MSE”, “var_error” o “SCMres”). Como puede observarse en la última referencia realizada, la variabilidad del grupo de interés se estima como la suma de “las diferencias al cuadrado entre la media local o del grupo i-ésimo de interés y la media global o entre grupos multiplicadas por la cantidad de observaciones correspondientes al grupo de interés y dicho resultado dividido entre K-1, donde”K" es el número de grupos“. Complementariamente, la variabilidad media global o varianza de la respuesta (el error de predicción medio SCMres) se estima como la suma de las diferencias al cuadrado entre el valor observado de las observaciones y las predicciones realizadas sobre su valor esperado, divida entre el resultado de restarle al tamaño de muestra global (en este caso es 30 y debe especificarse que se ha trabajado únicamente con una muestra) la cantidad de parámetros a estimar (incluyendo al intercepto); esto último se expuso con antelación en términos teóricos y de su sintaxis de estimación”var_error=SCTR/(n-k)“, que también se dijo podía ser expresada como”SCMres=SCTR/(n-k)" o “MSE=SCTR/(n-k)”. En el contexto del análisis de regresión, la cantidad de grados de libertad del numerador es igual a la cantidad de coeficientes de los que dispone el modelo completo (mod1, en este caso) menos la cantidad de coeficientes de los que dispone el modelo de interés (mod4, en este caso), es decir, 5-2=3.

Finalmente, debe señalarse que los valores F contenidos en la tabla F son los valores F correspondientes al escenario en que la hipótesis nula es verdadera usualmente a niveles de significancia de 0.01, 0.05 y 0.10, así como a una determinada cantidad de grados de libertad df1 del numerador y una determinada cantidad de grados de libertad df2 del denominador de la variable estocástica $(U/df1)/(V/df2)$ cuya distribución caracteriza la distribución F; en la expresión anterior, U y V son variables aleatorias independientes que se distribuyen χ^2 con $df1$ y $df2$ grados de libertad, respectivamente. Por esta razón, como se verifica en http://www.biokin.com/tools/f-critical.html (http://www.biokin.com/tools/f-critical.html), sabiendo que la región de valores F de la tabla para algún nivel de significancia determinado constituye la región para los cuales se fallaría en rechazar H_0 (puesto que implica que la confiabilidad respecto de la probabilidad preestablecida α de cometer error tipo I es insuficiente para rechazar H_0), el valor más pequeño reportado en dicha tabla (a determinada significancia y determinados valores de $df1$ y $df2$) es el valor respecto del cual el valor F crítico debe ser menor o a lo sumo igual, para poder rechazar H_0 (la cual sostiene que no existen diferencias relevantes entre los valores de comparación).

Hide

```
numerador_F=(SCTR4-SCTR1)/3
denominador_F=SCMres1
F=numerador_F/denominador_F
round(F,2)
```

```
##      [,1]
## [1,] 4.46
```

w.1.2. Obtención de la probabilidad asociada al valor F Crítico

Puede obtenerse la probabilidad asociada al valor F obtenido en w.1.1. a través de su distribución de probabilidad acumulada. Para ello, es recomendable utilizar la sintaxis “1-pf(F,df1,df2)”, en donde “pf” es la probabilidad acumulada de la distribución F, “df1” representa los grados de libertad del numerador y “df2” expresa los grados de libertad del denominador; complementariamente a lo dicho antes, se puede concebir los grados de libertad como la cantidad de valores independientes que un análisis estadístico puede estimar (véase https://statisticsbyjim.com/hypothesis-testing/degrees-freedom-statistics/ (https://statisticsbyjim.com/hypothesis-testing/degrees-freedom-statistics/)). Como se adelantó, en el contexto del análisis de regresión (específicamente la comparación de modelos de regresión en el proceso de depuración del modelo -que implica eliminación de las variables cuyo aporte explicativo es estadísticamente trivial-), los grados de libertad del numerador al calcular la probabilidad asociada a “F” están dados por el número de coeficientes de regresión del modelo completo (mod1) menos el número de coeficientes de regresión del modelo de interés (mod4), lo que como se dijo equivale a 5-2=3; por otro lado, los grados de libertad del denominador de la distribución F están dados por el tamaño de la muestra menos el número de coeficientes de regresión a estimar (incluyendo el intercepto, salvo que mencione lo contrario), lo cual es igual a 30-5=25.

Hide

```
1-pf(F,3,25)
```

```
##            [,1]
## [1,] 0.01220434
```

w.2. Obtención manual del valor F y su probabilidad asociada

Hide

```
anova(mod4,mod1)
```

```
## Analysis of Variance Table
##
## Model 1: veloc ~ edad
## Model 2: veloc ~ edad + imc + pierna + brazo
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     28 0.99021
## 2     25 0.64517  3   0.34504 4.4568 0.0122 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Si se establece un valor de $\alpha = 0.05$, implica que bajo determinados parámetros objetivos (información histórica y/o criterio de un experto experimentado) el investigador considera como admisible como máximo una probabilidad de cometer error tipo I de 0.05 (rechazar H_0 siendo verdadera). Puesto que el estadístico F tiene un valor de probabilidad (valor p) igual a 0.012, el cual es menor a la probabilidad de rechazar H_0 siendo verdadera (fijado por el investigador hipotético de este ejemplo en 0.05 bajo los parámetros mencionados), se rechaza la hipótesis nula de que todos los parámetros son iguales entre sí y a su vez iguales a cero.

x. También es posible comparar el modelo que considera todas las variables (mod1) contra el modelo para el cual se ha eliminado "edad" del análisis.

La hipótesis nula es:

$H_0 : \beta_1 = 0$

La hipótesis alternativa es:

$H_A : \beta_1 \neq 0$

Hide

```
anova(mod2,mod1)
```

```
## Analysis of Variance Table
##
## Model 1: veloc ~ imc + pierna + brazo
## Model 2: veloc ~ edad + imc + pierna + brazo
##   Res.Df     RSS Df Sum of Sq     F    Pr(>F)
## 1     26 1.56336
## 2     25 0.64517  1   0.91819 35.58 3.153e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El valor p de la distribución asociado a determinado nivel de significancia puede obtenerse, en este caso, directamente utilizando la sintaxis "summary(mod1)".

Hide

```
summary(mod1)
```

```
##
## Call:
## lm(formula = veloc ~ edad + imc + pierna + brazo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32604 -0.10319  0.01196  0.08794  0.31486
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.135422   0.602032  -0.225   0.8239
## edad        -0.019172   0.003214  -5.965 3.15e-06 ***
## imc          0.015532   0.010023   1.550   0.1338
## pierna      -0.002599   0.009907  -0.262   0.7952
## brazo        0.026262   0.012982   2.023   0.0539 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1606 on 25 degrees of freedom
## Multiple R-squared:  0.6271, Adjusted R-squared:  0.5674
## F-statistic: 10.51 on 4 and 25 DF,  p-value: 3.906e-05
```

Que lo anterior sea posible implica que existe equivalencia numérica entre el estadístico de prueba t y el estadístico de prueba F, ¿por qué ocurre esto? Como se verifica en http://homepage.stat.uiowa.edu/~rdecook/stat3200/notes/t_and_F_4pp.pdf (http://homepage.stat.uiowa.edu/~rdecook/stat3200/notes/t_and_F_4pp.pdf), existe una relación teórica inmediata entre ambas distribuciones puesto que si se eleva al cuadrado una distribución t se obtiene una distribución F con $df1 = 1$ para el numerador y $df2$ para el denominador, y es precisamente ese nexo teórico el que explica los resultados antes obtenidos. En términos de una concepción más general de esta relación matemática, se señala en https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/epdf/10.1002/cem.2734 (https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/epdf/10.1002/cem.2734) (p. 582) que la distribución F puede ser considerada como la extensión equivalente de la distribución t cuando hay más de una variable de respuesta (siempre una única variable explicativa, debe recordarse), pero tamaños de muestra pequeños. Evidentemente, no por ello sustituye a la prueba t. Desde la perspectiva puramente algebraica, en este caso de estudio existió convergencia entre ambas distribuciones (F y t) porque la distribución F tenía únicamente un grado de libertad en el numerador.

En este caso de aplicación se obtiene una probabilidad asociada al estadístico F o al estadístico t sumamente baja (p<0.0001), por lo que al comparar un modelo que contiene las variables “imc”, “brazo” y “pierna” (mod2) contra un modelo que contiene las 4 variables (mod1), se concluye que el coeficiente de regresión correspondiente a “edad” es diferente de cero. A causa de lo anterior, se concluye que se prefiere el modelo que incluye la totalidad de las variables porque el aporte explicativo de “edad” es estadísticamente relevante o, lo que es lo mismo, que la fiabilidad que puede tenerse respecto de la probabilidad de cometer error tipo I $\alpha = 0.05$ (rechazar H_0 siendo verdadera) es muy baja, por lo que se rechaza H_0.

Utilizando la sintaxis “summary” para el modelo 1 es posible visualizar en cada día de la salida que se obtendrá la prueba de hipótesis correspondiente a elmiminar una variable a la vez, es decir, $H_0 : \beta_j = 0$, por lo que hacerlo manualmente no es necesario siempre que se tengan lo suficientemente claros los conceptos asociados; estos valores aparecen en la cuarta columna de la estructura de datos “summary(mod1)”. Por ejemplo, es posible verificar mediante “summary(mod1)” el valor p asociado a la prueba de hipótesis relativa a eliminar “pierna” como variable explicativa de la media condicional de “veloc” o, lo que es lo mismo, la relevancia estadística del parámetro de regresión correspondiente a “pierna” en el modelo de regresión; este valor es de 0.7952, por lo que el coeficiente de regresión relativo a “pierna” es muy poco relevante en términos estadísticos (puesto que el valor p es la fiabilidad que se debe tener respecto de la probabilidad de cometer error tipo I $\alpha = 0.05$), de lo cual se desprende que se falla en rechazar la H_0 que afirma que la relevancia estadística del coeficiente de regresión para “pierna” es nula. Lo mismo ocurre para con la relevancia estadística de los demás modelos (con excepción del coeficiente de “brazo”, que aunque es mayor que $\alpha = 0.05$ no lo es por amplio margen), se falla en rechazar H_0 que afirma que dichos coeficientes son nulos.

Complementariamente, pueden obtenerse los intervalos de confianza de cada uno de los coeficientes de regresión y con ello verificar que rechazar o no H_0 coincide con el hecho de que los extremos de cada intervalo de confianza (relativo a cada parámetro i-ésimo) tengan signos iguales o no.

Hide

```
round(confint(mod1),4)
```

```
##                2.5 %  97.5 %
## (Intercept) -1.3753  1.1045
## edad        -0.0258 -0.0126
## imc         -0.0051  0.0362
## pierna      -0.0230  0.0178
## brazo       -0.0005  0.0530
```

Ambos extremos del intervalo de confianza al 0.95 para “edad” tienen signos negativos (-0.0258,-0.0126), lo cual coincide con el hecho de que se rechazó la hipótesis nula $\beta_1 = 0$. En cambio, los coeficientes de las demás variables tienen intervalos de confianza (a la misma confianza) con un extremo negativo y otro positivo. Esto coincide también con el hecho de que en ninguno de estos casos se llega a rechazar la hipótesis nula correspondiente. Por lo tanto, si los extremos del intervalo de confianza tienen signos diferentes implica que el investigador no tiene evidencia suficiente para afirmar que la variable en cuestión esté relacionada de forma bien definida en algún sentido (de incrementos o disminuciones) con la respuesta condicional promedio o, lo que es lo mismo, se falla en rechazar H_0 que afirma que el coeficiente de regresión en cuestión (o el conjunto de coeficientes, para el caso del contraste global) es nulo.

y. El contraste global del modelo consiste en probar la hipótesis más extrema que establece que ninguna de las variables es útil en el modelo, conocida como prueba ominibus y se conoce también en el contexto de la econometría como contraste de hipótesis global del modelo de regresión lineal clásico. Aunque técnicamente no sea necesario puesto que se conoce ya la relevancia de la variable “edad”, así como la poca relevancia estadística de las demás variables explicativas de la velocidad Y, tal ejercicio tiene alto valor pedagógico y metodológico por cuanto es un contraste estadístico fundamental que debe realizarse al iniciar cualquier análisis con la finalidad

de valorar oportunamente si merece el esfuerzo seguir adelante afinando el modelo. En esta prueba, lo que se compara es el modelo completo (mod1) contra el modelo más simple que contiene solamente el intercepto; en tal caso, la media a priori de la respuesta es igual al intercepto y la suma de cuadrados totales a priori SCTY1 (de nuevo, "1" porque mod1 es el modelo de referencia) es igual a la suma de cuadrados totales residuales SCTR8, es decir, (SCreg8=0) + SCTR8 = SCTY1. Lo anterior obedece a que, al no existir variables para estimar la respuesta condicional cuando únicamente se considera el intercepto, este último es igual a la media incondicional o a priori de la respuesta, puesto que no podría ser igual a otra cosa debido que a que para estimar la media condicional se necesitan tener variables explicativas diferentes (que son diferentes de la respuesta a priori) y distintas entre sí (linealmente independientes) que en el teorema de Bayes desempeñan el rol de información adicional sobre la respuesta incondicional. Esto se justifica en cuanto al no disponerse de variables explicativas (o que su valor sea nulo) el intercepto de un hipotético modelo de regresión (recordando que el intercepto representa el valor de la respuesta condicional cuando las variables explicativas toman el valor de cero) solamente podría ser igual a la media a priori de Y dado que el escenario a posteriori no tiene sentido sin información adicional y son los datos a priori los únicos con los que se contaría. Puesto que la suma de cuadrados totales residuales SCTR8 se define como la suma de las distancias al cuadrado entre las observaciones reales de y∈Y y la media condicional de Y, mientras que la SCTY se define como suma de las distancias al cuadrado entre las observaciones reales y∈Y y la media a priori de Y, dado que el intercepto es igual a la media incondicional de Y, la variabilidad total de la respuesta incondicional será igual a la variabilidad total que no es explicada por el modelo de regresión que sólo considera un coeficiente de regresión de intercepción, el cual es desconocido y es estimado a partir de los datos (en este caso, de los datos a priori y por ello la mejor estimación posible de realizar es la media de la distribución a priori Y). De forma más simple, si no se realiza un pronóstico con información adicional, el modelo resultante será uno que no pueda explicar condicionalmente a la respuesta (porque no existen datos adicionales para estimar una media condicional -diferente de la media a priori-) y, por consiguiente, ¿cuánto sería lo que de la variabilidad total de la respuesta que el modelo de regresión (de naturaleza condicional) así configurado no podría explicar?, pues la totalidad de dicha variabilidad, por eso es que SCTR8=SCTY.

Hide

```
mod8=lm(veloc~1)
anova(mod8)
```

```
## Analysis of Variance Table
##
## Response: veloc
##           Df Sum Sq  Mean Sq F value Pr(>F)
## Residuals 29   1.73 0.059655
```

Hide

```
SCTY1=sum(anova(mod1)[,2])
SCTY8=sum(anova(mod8)[,2])
```

Sabiendo que $SCTY_8 = SCTY_1 = \ldots$. Así, se verifica que $SCTY1 = 1.730004$, donde $SCTY1$ es la suma de cuadrados a priori de la respuesta en mod1 que, como también ha dicho y se acaba de demostrar, es independiente de las variables explicativas utilizadas.

Hide

```
fit8=predict(mod8)
m=mean(veloc)
SCreg8=sum((fit8-m)^2)
```

se verifica que $SCreg8$ es aproximadamente cero. Debe decirse que no es exactamente cero debido a que R utiliza métodos numéricos para operativizar los postulados teóricos vistos, por lo que no se puede esperar completa coincidencia entre tales postulados (que son abstracciones del pensamiento) con los resultados empíricos que son aplicaciones concretas de los postulados teóricos, en cuanto las abstracciones son generalizaciones de las categorías construidas a través de la experiencia aplicada en alguna rama de las ciencias, y tales generalizaciones serían imposibles de realizar si se consideraran todos los elementos que caracterizan a cada uno de los hechos empíricos (fenómenos) que contribuyó a formar una imagen general (abstracta) sobre los mismos, lo cual puede verse (por ejemplo) en el proceso hegeliano conocido como Aufheben (que elimina a Lo Singular en el proceso del universal abstracto hacia el universal concreto en la combinación dialéctica entre Lo Universal y Lo Particular), en la naturaleza misma de las Matemáticas (que sólo considera Lo Universal y para ello requiere de eliminar Lo Singular y Lo Particular -porque no hay Aufheben, no hay dinámica en sus fundamentos teóricos-) o en términos aplicados en los métodos de numéricos de aproximación a funciones donde siempre existe pérdida de información, aunque esta pérdida pueda ser trivial en términos de su relevancia explicativa (que tiene mayor similitud con la lógica que sigue el Aufheben hegeliano que con la Matemática Pura, a pesar que los métodos numéricos se deriven de las Matemáticas -aunque por supuesto no en una forma que siga una lógica lineal).

Hide

```
r8=mod8$res
SCTR8=t(r8)%*%r8
```

Se puede verificar que $SCTR_8$ es igual a 1.730004, verificándose entonces que (SCreg8=0) + SCTR8 = SCTY1, tal como se estableció antes.

También puede compararse la estimación del intercepto con la estimación de la media de la respuesta incondicional y verificar la equivalencia entre ambas antes mencionada.

Hide

```
mod8$coef
```

```
## (Intercept)
##    1.373278
```

Hide

```
mean(veloc)
```

```
## [1] 1.373278
```

Con la finalidad de verificar empíricamente que (SCreg8=0) + SCTR8 = SCTY1, también se puede utilizar la SCTY1 para construir manualmente la prueba F y constatar cómo es indiferente la utilización de SCTR8 o de SCTY1 en los cálculos numéricos de la prueba en cuestión. Así como antes se estableció que “numerador_F=(SCTR4-SCTR1)/3”, ahora esta sintaxis adoptará la forma “numerador_F=(SCTY8-SCTR1)/4, en donde el cambio del denominador de”numerador_F" (antes de 3, ahora de 4) obedece a que la cantidad de parámetros que posee el modelo completo (mod1) es de 5 y la cantidad de parámetros que posee mod8 es igual a 1 (puesto que sólo se considera el intercepto), por lo que 5-1=4.

Hide

```
SCTY8=sum(anova(mod8)[,2])
numerador_F_mod8=(SCTY8-SCTR1)/4
denominador_F_mod8=SCMres1=SCTR1/(n-k1)
F=numerador_F_mod8/denominador_F_mod8
round(F,2)
```

```
##       [,1]
## [1,] 10.51
```

Adicionalmente, puede obtenerse como antes la probabilidad asociada al valor F obtenido antes.

Hide

```
1-pf(F,4,25)
```

```
##              [,1]
## [1,] 3.905766e-05
```

Los resultados antes expuestos pueden obtenerse de forma automatizada utilizando el comando “anova” y aninando mod8 y mod1, como se muestra a continuación.

Hide

```
anova(mod8,mod1)
```

```
## Analysis of Variance Table
##
## Model 1: veloc ~ 1
## Model 2: veloc ~ edad + imc + pierna + brazo
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     29 1.73000
## 2     25 0.64517  4    1.0848 10.509 3.906e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Finalmente, pueden también obtenerse los resultados anteriores, como se mostró antes, directamente del modelo completo (mod1) aplicando la sintaxis “summary” sobre este.

Hide

```
summary(mod1)
```

```
##
## Call:
## lm(formula = veloc ~ edad + imc + pierna + brazo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32604 -0.10319  0.01196  0.08794  0.31486
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.135422   0.602032  -0.225   0.8239
## edad        -0.019172   0.003214  -5.965 3.15e-06 ***
## imc          0.015532   0.010023   1.550   0.1338
## pierna      -0.002599   0.009907  -0.262   0.7952
## brazo        0.026262   0.012982   2.023   0.0539 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1606 on 25 degrees of freedom
## Multiple R-squared:  0.6271, Adjusted R-squared:  0.5674
## F-statistic: 10.51 on 4 and 25 DF,  p-value: 3.906e-05
```

Las comparaciones de modelos antes realizadas de forma manual (sobre mod4 y mod8) tienen como finalidad el verificar si al menos uno de los coeficientes de regresión generados a partir del conjunto de datos disponible tienen una relevancia estadística nula o no (lo que busca representar desde la teoría de las probabilidades que tienen -o no- un aporte explicativo relevante al fenómeno de estudio); como se vio, todos estos resultados individuales obtenidos manualmente se pueden obtener en su conjunto de forma automatizada y directa a través de la sintaxis “summary(mod1)”, donde “mod” debe ser siempre el modelo completo. Puesto que el estadístico de prueba F tiene una probabilidad asociada sumamente baja (p<0.0001), se decide rechazar H_0 que considera que todos los coeficientes de regresión poseen una relevancia estadística nula. Así, al comparar un modelo que considera únicamente el intercepto (mod8) contra el que contiene las 5 variables (1 de respuesta y 4 explicativas), se obtiene que al menos unos de los coeficientes de las variables “edad”, “imc”, “pierna” o “brazo” es diferente de cero. Por tanto, se prefiere el modelo completo al modelo que considera únicamente el intercepto. Por supeusto, lo anterior necesariamente no excluye el escenario en el que pueda existir un modelo apropiado (diferente a mod8 y mod1) para explicar el comportamiento de la variable de respuesta tan bien o mejor en comparación al modelo completo (en este caso mod1).

B. APLICACIONES EN ECONOMÍA POLÍTICA

I. OBJETIVO GENERAL

Estudiar estadísticamente, como parte de un ejercicio pedagógico, los determinantes fundamentales lineales de la tasa media de ganancia para el caso de Estados Unidos en el período 1964-2008 mediante un análisis de regresión lineal.

II. ANTECEDENTES TEÓRICOS Y EMPÍRICOS

La tasa media de ganancia ha sido estudiada por las distintas escuelas de pensamiento económico desde antes que la economía política fuese expresada por Adam Smith como un cuerpo teórico coherente por vez primera en su célebre obra *Riqueza de las Naciones* publicada el 9 de marzo de 1776, fundamentalmente en la vasta obra de William Petty (fundamentalmente en *Anatomía Política de Irlanda* y *Aritmética Política*) y en la tabla económica de François Quesnay (que fueron la antesala de las matrices insumo-producto en términos de estadísticas macroeconómicas alrededor del producto neto). ¿Cómo se abordará aquí este análisis?, al respecto debe decirse que esta investigación considerará únicamente las concepciones teóricas formuladas desde la teoría del valor-trabajo de los economistas clásicos y afines (fundamentalmente Smith, Ricardo y Marx), por motivos que se expondrán a continuación; debe decirse complementariamente que, pesar de las dificultades infranqueables para una unificación general de las escuelas de pensamiento económico, prácticamente la totalidad de las escuelas (y la totalidad de las relevantes) admiten que existe una tasa de ganancia promedio que describe globalmente al sistema macroeconómico, pero también existe convergencia general entre dichas escuelas respecto al hecho de la tendencia de largo plazo a decrecer de dicha tasa; en lo que se distinguen las escuelas es en las causas a las que atribuyen dicha tendencia a la baja en el largo plazo. La posición clásica, expresada nítidamente en Adam Smith y David Ricardo, es la misma que la posición de Karl Marx, aunque tras este último ocurriría una ruptura histórica que nacería con la escuela marginalista, que reina teóricamente en la academia ortodoxa hasta la actualidad; sin embargo, como se verá en breve, toda la evidencia estadística apunta a que las concepciones clásicas y marxistas son una mejor explicación de la tasa media de ganancia. En (Smith, 1977, pág. 127) se señala que: “El aumento y la caída de las ganancias de los stocks dependen de las mismas causas que el aumento y la caída de los salarios del trabajo, el estado creciente o decreciente de la riqueza de la sociedad; pero esas causas afectan a unos y a otros de manera muy diferente. El aumento de los stocks, que eleva los salarios, tiende a reducir las ganancias. Cuando las acciones de muchos comerciantes ricos se convierten en el mismo comercio, su competencia mutua tiende naturalmente a reducir sus ganancias; y cuando hay un aumento similar de acciones en todos los diferentes oficios que se llevan a cabo en la misma sociedad, la misma competencia debe producir el mismo efecto en todos ellos. No es fácil, ya se ha observado, determinar cuáles son los salarios medios del trabajo, incluso en un lugar y en un momento determinados.” Por su parte, (Ricardo, 2004, pág. 71) señala que “Habiéndose demostrado que las ganancias de las acciones, en diferentes empleos, guardan una proporción entre sí y tienen una tendencia a variar todas en el mismo grado y en la misma dirección, nos queda por considerar cuál es la causa de las variaciones permanentes en la tasa de ganancia, y las consiguientes alteraciones permanentes en la tasa de interés.”

Para (Marx, 2010, págs. 161-263), la tasa media de ganancia es el resultado de largo plazo, tras el proceso de competencia capitalista por la apropiación del producto neto generado por la sociedad, alrededor del cual se organiza la producción (y, por consiguiente, tratándose de Marx, también la circulación) en cada uno de los períodos de producción sucesivos, “Se presenta como una divergencia sistemática entre el valor producido en una rama y el valor apropiado en la circulación, de manera más precisa, la ganancia se compone por la plusvalía extraída en la rama, más un incremento (positivo o negativo), que resulta en la circulación. La realización de las mercancías por sus precios de producción, implica una apropiación o cesión sistemática de plusvalía para cada una de las ramas.” (Valle Baeza, 1978, pág. 200). La explicación de cómo ocurre este proceso por vez primera (antes que se forme por primera vez la tasa media de ganancia) es extremadamente compleja y no se abordará aquí, sin embargo, debe decirse que posee un carácter histórico-empírico y es conocida teóricamente como el *proceso de transformación de valores en precios de producción* que ocurre en la infancia del capitalismo, cuando este apenas se desarrollaba en el seno de la sociedad feudal que sustituiría; la

fundamentación a esta explicación histórica se encuentra en la obra del mismo Marx (fundamentalmente en el tomo I de *El Capital* -en el capítulo titulado "La llamada acumulación originaria" - y en los *Grundrisse* -en el capítulo titulado "El capítulo del capital"-), pero también implícita en la lógica de las simulaciones empíricas a través de métodos iterativos como en (Shaikh, 1998), así como también fundamentada por estudios estadísticos como la investigación de (Duménil & Lévy, 1998) para Estados Unidos con datos desde la Guerra Civil, además de otras múltiples investigaciones teóricas (algunas de ellas en la obra citada cuyo editor es Riccardo Bellofiore) como (Emmanuel, 1972, págs. 425-427), (Freeman & Carchedi, 1995, págs. 180-233), así como también es fundamentada historiográficamente en (Dobb, 2008), (Sweezy, y otros, 1978), (Aston & Philpin, 1987) y muchos otros, en donde la transformación histórica de las condiciones de producción feudales en condiciones de producción capitalistas recorre transversalmente los planteamientos y debates historiográficos que ahí se abordan. Complementariamente, en (Nabi, 2020) se compila a detalle el debate sostenido entre el Nobel de Economía Paul Krugman y el célebre macroeconometrista Gregory Mankiw (ambos pertenecientes a la escuela conocida como síntesis neoclásica) sobre la inherencia o no de las raíces unitarias en las series de tiempo y se explica en términos de que su inherencia a estas implicaría, por cuanto la producción interna bruta de una economía se puede descomponer en salarios y beneficios (los cuales pueden ser expresados a su vez como proporciones de participación en el producto global), una verificación empírica de la ley de la tendencia decreciente de la tasa media de ganancia armonizada con la teoría estadística. Este debate se inclinó, como el lector puede verificar en la fuente citada (y en las fuentes citadas por la fuente citada), a favor del macroeconometrista de Harvard; sin embargo, evidentemente Mankiw no es el único economista neoclásico de prestigio que considera esto, puesto que entre otros se le une (Blanchard, 2009, pág. 8), quien afirmó que "Si bien existe una gran variación entre países, la conclusión es que, en promedio, la producción no vuelve a su antigua trayectoria de tendencia, sino que permanece permanentemente por debajo de ella." en su etapa como economista jefe del Fondo Monetario Internacional (FMI).

Existe basta evidencia estadística de que la teoría del valor-trabajo de los economistas clásicos y Marx permite formular modelos matemáticos y estadísticos (más allá de las divergencias parciales que pueden existir en ciertos casos -que hasta cierto punto es un hecho natural de las ciencias normales, en el sentido de Thomas Kuhn-) que explican con mayor robustez empírica (a un determinado nivel de confianza) el patrón geométrico descrito por el conjunto de datos en relación a los que podrían formularse considerando como determinantes fundamentales de la creación de valor y de la formación de precios a otros factores productivos diferentes y distintos del trabajo bajo un análisis insumo-producto (lineal por definición, desde su aparición histórica en la obra del célebre economista Wassily Leontief). El listado de las investigaciones que muestran la mencionada evidencia estadística es cuantioso, por lo que se hará un breve sumario del mismo: 1. (Sánchez & Ferràndez, Valores, precios de producción y precios de mercado a partir de los datos de la economía española, 2010) -la forma clásica de este tipo de investigaciones, enfocada en España-, 2. (Cockshott & Cottrell, Robust correlations between prices and labor values, 2005) -enfocada en descartar teórica y empíricamente correlación espuria-, 3. (Cockshott, Cottrell, & Valle Baeza, The Empirics of the Labour Theory of Value: Reply to Nitzan and Bichler, 2014) -enfocada en descartar correlación espuria entre las variables involucradas-, 4. (Sánchez & Montibeler, La teoría del valor trabajo y los precios en China, 2015) -enfocada en el caso de China-, 5. (Zachariah, 2006) -se estudian 18 países de 1968 a 2000, además de estudiar la ecualización de las ganancias antes mencionada-, 6. (Işıkara & Mokreb, 2021) -se estudian 42 países de 2000 a 2017 con 36,000 vectores precio-. A esto debe agregarse que el histórico debate teórico que sostuvieron la escuela neoclásica y las escuelas postkeynesiana y neoricardiana, recopilado nítidamente en (Jiménez, 2011, págs. 183-298) y

conocido como la *Controversia del Capital de Cambridge* (pues enfrentaba a escuela neoclásica afincada en el MIT domiciliado en Cambridge de Estados Unidos contra las escuelas postkeynesiana y neoricardiana afincadas en University of Cambridge domiciliada en Cambridge de Inglaterra) sobre la teoría del capital neoclásica (cuya génesis histórica-teórica se localiza en la obra de Knut Wicksell), se saldó a favor de los postkeynesianos y neoricardianos, como el mismísimo Paul Samuelson (máximo exponente neoclásico del debate, participante más recurrente y relevante del bando neoclásico, así como fundador de la Síntesis Neoclásica en su obra *Foundations of Economic Analysis*) expresando que "El fenómeno de la reversión a una tasa de interés muy baja a un conjunto de técnicas que habían parecido viables solo a una tasa de interés muy alta implica más que tecnicismos esotéricos. Ello muestra que el cuento sencillo de Jevons, Böhm Bawerk, Wicksell y otros autores neoclásicos -según el cual a medida que baja la tasa de interés como consecuencia de la abtención del consumo presente a favor del consumo futuro, la tecnología debe volverse en algún sentido más *indirecta*, más *mecanizada* y más *productiva*- no puede ser universalmente válida (...) No hay manera inequívoca de caracterizar diferentes procesos como más *intensivos* en *capital*, más *mecanizados*, más *indirectos*, excepto en el sentido tautológico *ex post* de haber sido adoptados a una tasa de interés baja e involucrando un salario real alto. Este tipo de tautología ha mostrado, en el caso del *reswitching*, que lleva a una clasificación inconsistente entre pares de tecnologías constantes, dependiendo de cuál tasa de interés prevalecerá en el mercado. Si todo esto causa dolores de cabeza a quienes suspiran por las viejas parábolas de la teoría neoclásica, deberemos recordarles que los académicos no han nacido para llevar una existencia fácil. Debemos respetar y valorar los hechos de la vida." (Samuelson, The Quarterly Journal of Economics, págs. 568-583).

Como señala (Jiménez, 2011, pág. 228), lo anterior implica que la concepción neoclásica de que las remuneraciones de los factores de producción *capital* y *trabajo* se explican por sus respectivas productividades marginales ha sido invalidada teórica, matemática y lógicamente. Con ello se derrumba la explicación de que la distribución del ingreso vía oferta y demanda, la teoría de que los precios son indicadores de *escasez* y la concepción neoclásica de la producción. Adicionalmente, nótese que de lo que Samuelson afirma se deriva que los determinantes fundamentales del sistema económico no deben buscarse únicamente en factores económicos, puesto que bajo la lógica neoclásica ¿qué motivo no-técnico podría conducir a un agente a realizar la reversión descrita si este es racional?, esta pregunta bajo el marco neoclásico no tiene respuesta, pero bajo el marco marxista se responde inmediatamente que existen factores complementarios al económico que explican la dinámica del sistema económico, que son de carácter político. Esto, en congruencia con lo antes expuesto, reafirma al marco marxista como una teoría más adecuada para explicar la realidad económica objetiva.

En el marco de las investigaciones empíricas realizadas por la escuela marxista como las antes citadas, los modelos de regresión son utilizados para corroborar empíricamente las relaciones entre valores, precios de producción y precios de mercado (y en algunos casos, como en las investigaciones de Sánchez, se suelen incluir los precios sraffarianos -neoricardianos-). En esta investigación se buscará también tomar en consideración la productividad del capital, puesto que aunque tras la controversia de Cambridge se sepa que no explica teóricamente de forma general la formación de precios, es innegable que como señaló Marx sí transfiere valor al producto y multiplica las fuerzas productivas del trabajo vivo (pues Marx considera al capital trabajo cristalizado, trabajo pretérito), por lo que casi seguramente tendrá alguna incidencia estadística al explicar la tasa media de ganancia; aquí se considera la productividad media del capital. Por motivos teóricos y también de contraste con la productividad media del capital, se incluye la productividad media del trabajo, así como también la tasa de depreciación del capital y el acervo neto estandarizado de capital fijo para contrastar su poder explicativo y significancia

con la masa salarial promedio y la tasa media de salario. Los datos se han obtenido de (Marquetti & Foley, 2012), quienes elaboran las *Extended Penn World Tables*, que son una generalización de las Penn World Tables (véase https://www.rug.nl/ggdc/productivity/pwt/?lang=en (https://www.rug.nl/ggdc/productivity/pwt/?lang=en)) y otras bases de datos para realizar investigación sobre las variables económicas reales, comprendida del período 1963-2009 (aunque para algunas variables y/o países la disponibilidad puede variar). En esta investigación se escogió el período 1964-2008 para que todas las variables tengan la misma longitud.

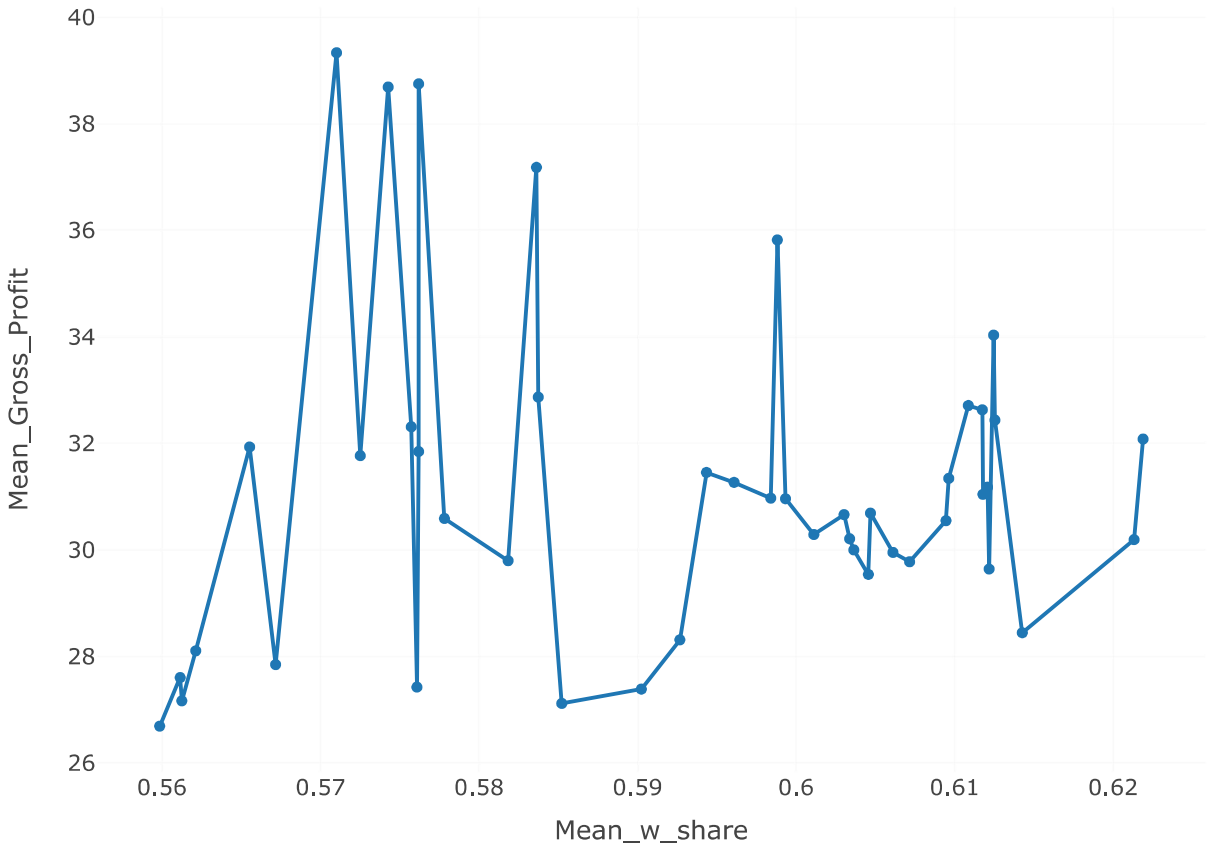
Complementariamente, debe especificarse que aquí no se ha dado un tratamiento a las variables mediante los métodos matriciales que la corriente marxista conocida como simultaneista-fisicalista (y que es la escuela líder en el tipo de investigaciones particulares que analizan la correlación entre valores, precios de producción, precios de mercado y precios sraffarianos). ¿Por qué?, porque además de que la aplicación de una metodología y no otra tiene implicaciones filosóficas que aún no parecen estar bien-definidas en términos de su alcance (cualitativa y cuantitativamente hablando), como se puede verificar en (citar a Nabi, creación y destrucción de valor), esta investigación no puede ser más que análisis preliminar sobre las variables reales, un ejercicio académico con el modelo clásico de regresión lineal múltiple y, además, orientado en otra dirección: específicamente en la de determinar (al menos preliminarmente) los predictores fundamentales de la tasa media de ganancia, no de medir la capacidad explicativa de los valores. ¿Por qué estudiar la tasa media de ganancia? Según investigaciones empíricas pioneras en tiempos actuales como (Tapia Granados, 2012) y (Wells, 2007), así como también investigaciones empíricas anteriores como la de (Farjoun & Marchover, 1983) y la investigación fundacional de Marx, la tasa media de ganancia es (como se adelantó) la variable alrededor de la cual se organizan los agentes dentro del sistema económico capitalista, por ello su estudio tiene una relevancia fundamental.

Finalmente, la razón por la que se estudió la economía estadounidense y no otra es porque para el período de análisis seleccionado esta era la economía capitalista de mayor importancia planetaria por su tamaño e influencia. Esto es importante de mencionar, puesto que fue precisamente la razón por la que Marx en sus análisis se enfocó exclusivamente en Estados Unidos, Inglaterra, Francia y Alemania (por el nivel de desarrollo de sus fuerzas productivas para ese entonces), porque eran la mejor representación del capitalismo de su época. De igual forma, todas las investigaciones neoclásicas de carácter teórico se acostumbran también a hacer para países de alto desarrollo industrial, especialmente Estados Unidos.

III. ANÁLISIS DESCRIPTIVO

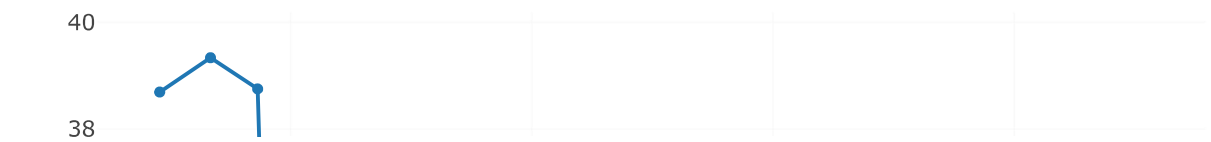
Hide

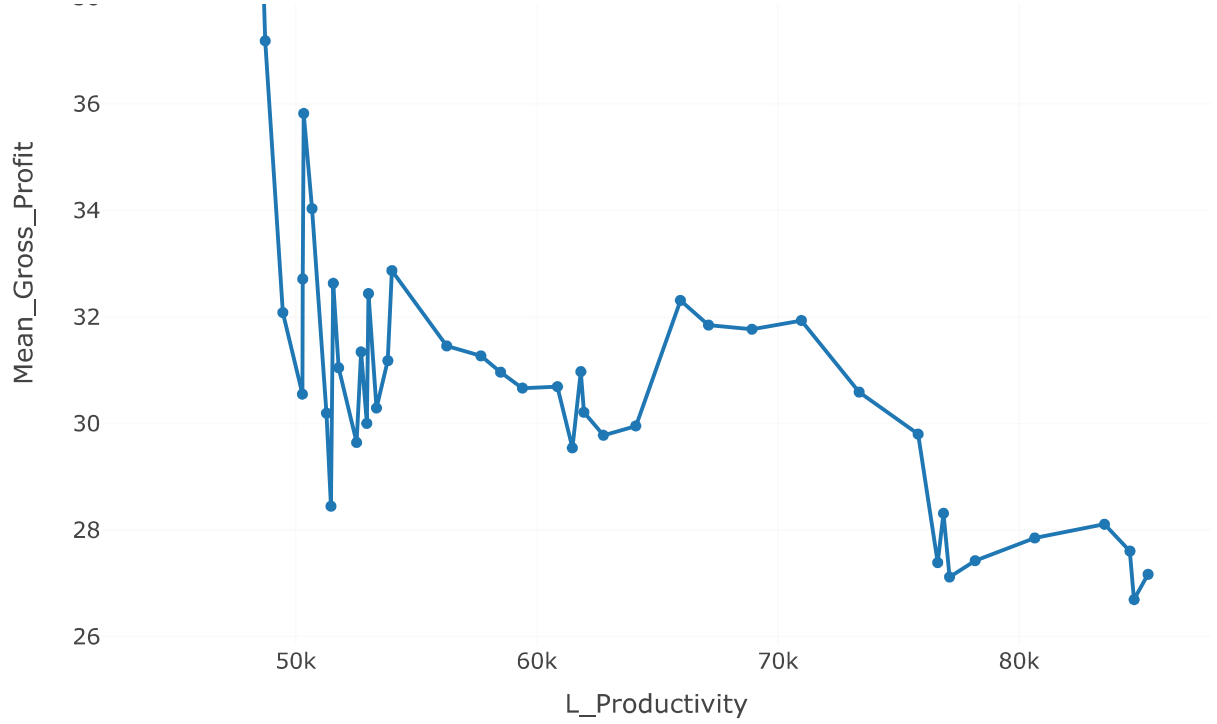
```
setwd("C:/Users/User/Desktop/Carpeta de Estudio/Maestría Profesional en Estadís
tica/Semestre II-2021/Métodos, Regresión y Diseño de Experimentos/2/Modelo de R
egresión, Maestría UCR")
library(readxl)
DATOS <- read_excel("C:/Users/User/Desktop/Carpeta de Estudio/Maestría Profesio
nal en Estadística/Semestre II-2021/Métodos, Regresión y Diseño de Experimento
s/2/Modelo de Regresión, Maestría UCR/DATOS.xlsx")
library(readxl)
library(plotly)
library(dplyr)
library(tidyr)
library(DT)

datos_ordenados <- DATOS[order(DATOS$Mean_w_share),]
fig <- plot_ly(datos_ordenados, x = ~Mean_w_share)
fig <- fig %>% add_trace(y = ~Mean_Gross_Profit, mode = 'lines+markers')
fig
```



Hide

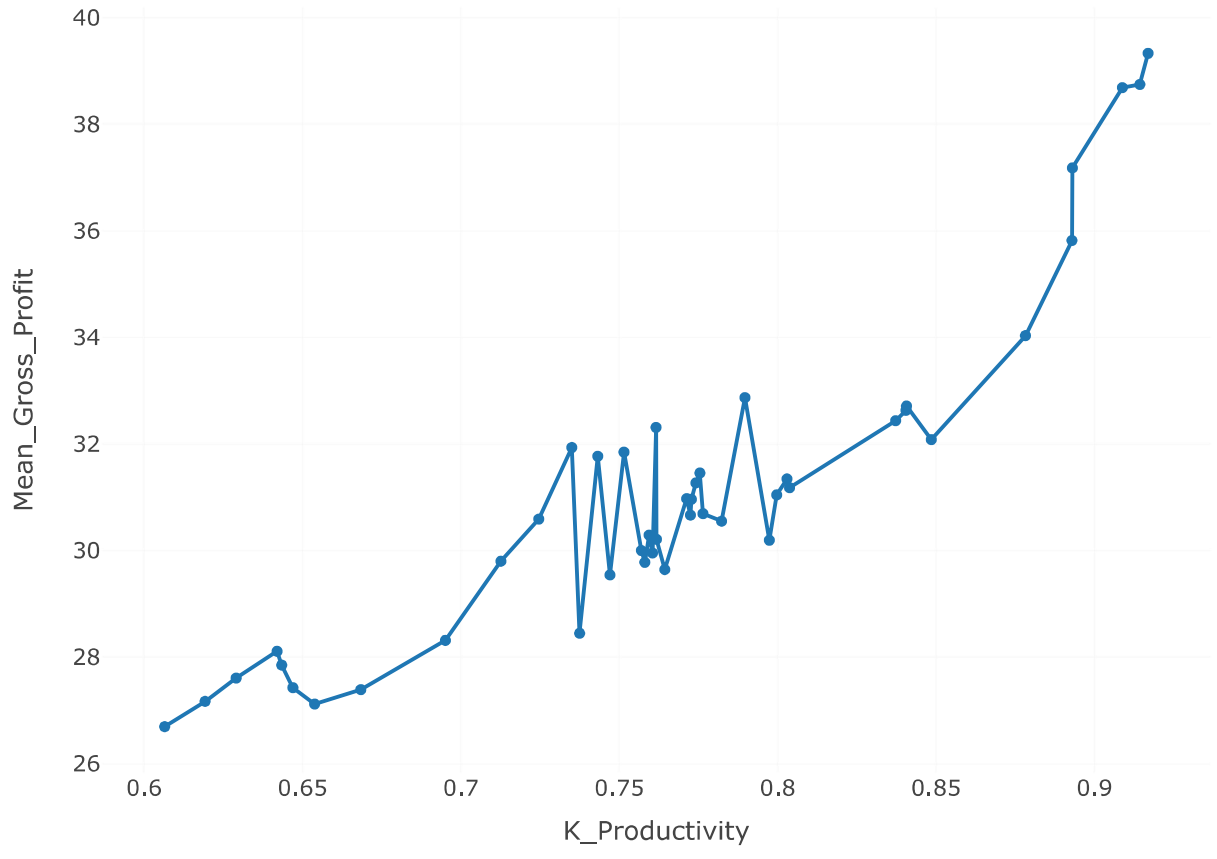
```
datos_ordenados <- DATOS[order(DATOS$L_Productivity),]
fig <- plot_ly(datos_ordenados, x = ~L_Productivity)
fig <- fig %>% add_trace(y = ~Mean_Gross_Profit, mode = 'lines+markers')
fig
```





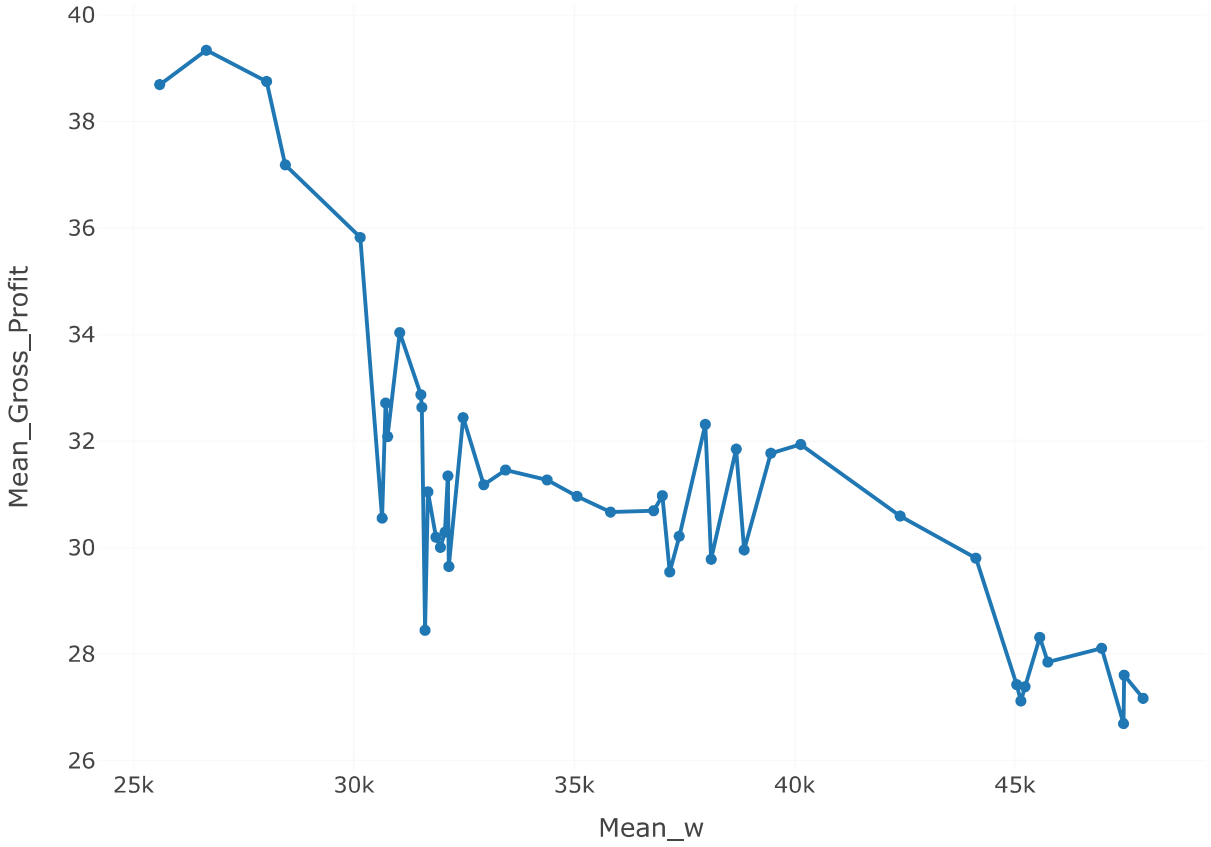
Hide

```
datos_ordenados <- DATOS[order(DATOS$K_Productivity),]
fig <- plot_ly(datos_ordenados, x = ~K_Productivity)
fig <- fig %>% add_trace(y = ~Mean_Gross_Profit, mode = 'lines+markers')
fig
```



Hide

```
datos_ordenados <- DATOS[order(DATOS$Mean_w),]
fig <- plot_ly(datos_ordenados, x = ~Mean_w)
fig <- fig %>% add_trace(y = ~Mean_Gross_Profit, mode = 'lines+markers')
fig
```



A nivel gráfico, se puede observar que la relación entre tasa media de ganancia tiene una relación inversa con la tasa media salarial, la productividad del trabajo y la masa salarial promedio, mientras que una relación directa con la productividad media del capital. Estos resultados del análisis exploratorio preliminar deben verificarse mediante procedimientos estadísticos formales.

IV. ANÁLISIS INFERENCIAL: REGRESIÓN LINEAL MÚLTIPLE

IV.I. Verificación del Modelo de Mejor Ajuste vía eliminación hacia atrás mediante el Criterio Bayesiano de Información (BIC)

Hide

```
lmfit2=lm(Mean_Gross_Profit~., DATOS)
n = nrow(DATOS)
lmfit2.1=step(lmfit2,k=log(n))
```

```
## Start:  AIC=-261.75
## Mean_Gross_Profit ~ Year + RealGDP + K_Net_Stock + DepreciationRate +
##     K_Productivity + L_Productivity + Mean_w_share + Mean_w
##
##                    Df Sum of Sq    RSS      AIC
## - DepreciationRate  1    0.0000 0.0626 -265.561
## <none>                          0.0626 -261.755
## - Year              1    0.0114 0.0740 -258.014
## - RealGDP           1    0.0120 0.0745 -257.697
## - K_Net_Stock       1    0.0133 0.0759 -256.901
## - Mean_w            1    0.2060 0.2686 -200.006
## - L_Productivity    1    0.2675 0.3301 -190.727
## - Mean_w_share      1    2.0933 2.1559 -106.277
## - K_Productivity    1    5.2538 5.3164  -65.661
##
## Step:  AIC=-265.56
## Mean_Gross_Profit ~ Year + RealGDP + K_Net_Stock + K_Productivity +
##     L_Productivity + Mean_w_share + Mean_w
##
##                   Df Sum of Sq    RSS      AIC
## <none>                         0.0626 -265.561
## - Year             1    0.0114 0.0740 -261.814
## - RealGDP          1    0.0122 0.0748 -261.337
## - K_Net_Stock      1    0.0140 0.0766 -260.295
## - Mean_w           1    0.2197 0.2823 -201.571
## - L_Productivity   1    0.3535 0.4161 -184.110
## - Mean_w_share     1    2.1636 2.2262 -108.641
## - K_Productivity   1    6.8435 6.9061  -57.695
```

Hide

```
summary(lmfit2.1)
```

```
##
## Call:
## lm(formula = Mean_Gross_Profit ~ Year + RealGDP + K_Net_Stock +
##     K_Productivity + L_Productivity + Mean_w_share + Mean_w,
##     data = DATOS)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.121127 -0.019434 -0.002306  0.018275  0.109041
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.325e+02  2.382e+01   5.560 2.47e-06 ***
## Year           -3.166e-02  1.217e-02  -2.600  0.01330 *
## RealGDP         3.935e-13  1.464e-13   2.689  0.01069 *
## K_Net_Stock    -1.035e-13  3.600e-14  -2.875  0.00666 **
## K_Productivity  3.887e+01  6.110e-01  63.613  < 2e-16 ***
## L_Productivity -3.801e-04  2.629e-05 -14.459  < 2e-16 ***
## Mean_w_share   -1.146e+02  3.204e+00 -35.768  < 2e-16 ***
## Mean_w          5.834e-04  5.119e-05  11.398 1.15e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04112 on 37 degrees of freedom
## Multiple R-squared:  0.9998, Adjusted R-squared:  0.9998
## F-statistic: 3.415e+04 on 7 and 37 DF,  p-value: < 2.2e-16
```

Hide

```
lmfit2=lm(Mean_Gross_Profit~., DATOS)
lmfit2.1=lmfit2
drop1(lmfit2.1,)
```

```
## Single term deletions
##
## Model:
## Mean_Gross_Profit ~ Year + RealGDP + K_Net_Stock + DepreciationRate +
##     K_Productivity + L_Productivity + Mean_w_share + Mean_w
##                   Df Sum of Sq    RSS      AIC
## <none>                         0.0626 -278.015
## Year               1    0.0114 0.0740 -272.467
## RealGDP            1    0.0120 0.0745 -272.150
## K_Net_Stock        1    0.0133 0.0759 -271.355
## DepreciationRate   1    0.0000 0.0626 -280.014
## K_Productivity     1    5.2538 5.3164  -80.114
## L_Productivity     1    0.2675 0.3301 -205.181
## Mean_w_share       1    2.0933 2.1559 -120.730
## Mean_w             1    0.2060 0.2686 -214.459
```

Por tanto, se conservarán el intercepto, las productividades medias del capital y del trabajo, así como la masa salarial promedio y la tasa media de salario para explicar la variable de respuesta *tasa media de ganancia*.

IV.II. El Modelo de Regresión

Hide

```
lmfit1 <- lm(Mean_Gross_Profit ~ Mean_w_share + L_Productivity + K_Productivity
+ Mean_w, data = DATOS)
predict_GP <- predict(lmfit1, newdata = data.frame(Mean_w_share = DATOS$Mean_w_
share,
                                                   L_Productivity = DATOS$L_Pro
ductivity,
                                                   K_Productivity = DATOS$K_Pro
ductivity,
                                                   Mean_w = DATOS$Mean_w))
summary(lmfit1)
```

```
##
## Call:
## lm(formula = Mean_Gross_Profit ~ Mean_w_share + L_Productivity +
##     K_Productivity + Mean_w, data = DATOS)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.108989 -0.025917 -0.002001  0.020774  0.114180
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     7.153e+01  1.483e+00   48.22   <2e-16 ***
## Mean_w_share   -1.188e+02  2.321e+00  -51.17   <2e-16 ***
## L_Productivity -3.951e-04  2.201e-05  -17.95   <2e-16 ***
## K_Productivity  3.990e+01  2.343e-01  170.34   <2e-16 ***
## Mean_w          6.536e-04  3.719e-05   17.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0439 on 40 degrees of freedom
## Multiple R-squared:  0.9998, Adjusted R-squared:  0.9998
## F-statistic: 5.244e+04 on 4 and 40 DF,  p-value: < 2.2e-16
```

El análisis de regresión confirma, salvo para el caso de la masa salarial promedio y a falta de realizar todas las pruebas de diagnóstico sobre la robustez estadística del modelo, la dirección de n-1 correlaciones de las que se observaron en el análisis descriptivo.

Adicionalmente, los coeficientes de regresión del modelo son significativos a cualquier nivel de confianza que se elija, el error de estimación promedio es de 0.0439 (el cual es bajo, puesto que la tasa media de ganancia en la base de datos está dada para todas las observaciones en decenas), el coeficiente de determinación es equivalente al ajustado (lo que parecería indicar una adecuada selección de las variables, porque no hay penalización por sobreajuste) y el intercepto es positivo y significativo (por lo cual no es estadísticamente válido eliminarlo), lo que implica que existe una tasa media de ganancia autónoma en el sistema que no depende de los

predictores seleccionados; esto parece ser, a falta de verificar la robustez del modelo, alguna fuerza de evidencia para la afirmación antes hecha de que los determinantes de la dinámica del sistema económico no son estrictamente económicos. La estimación preliminarmente parece ser confiable, por la elevada cantidad de grados de libertad (40) con la que se estimaron los diversos estadísticos de prueba.

IV.III. Tabla Comparativa: tasa media de ganancia observada versus la estimada con el modelo

Hide

```
new_predict <- cbind(DATOS, predict_GP)
new_predict <- cbind(new_predict, lmfit1$residuals)
library(DT)
datatable(new_predict[,c(1,9,10,11)])
```

Show 10 entries Search:

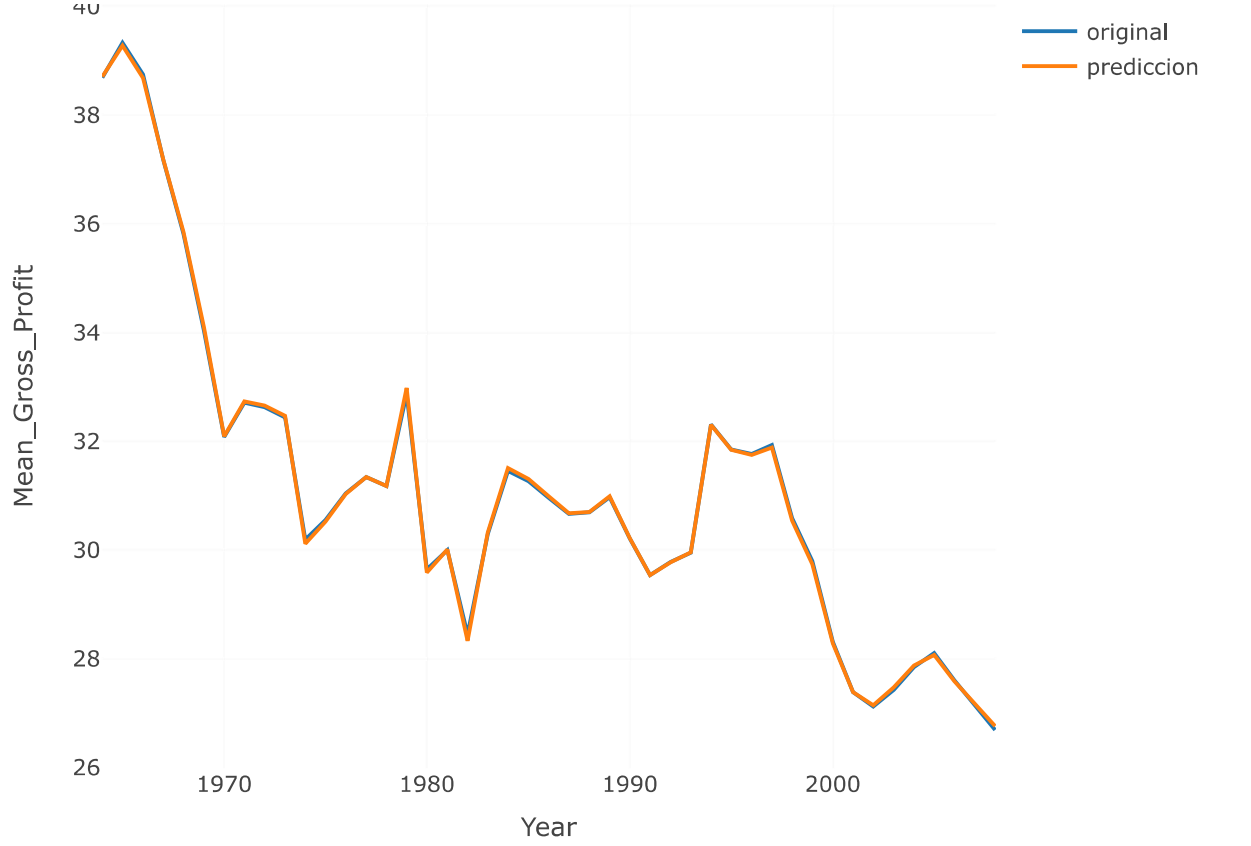
	Year	Mean_Gross_Profit	predict_GP	lmfit1$residuals
1	1964	38.6905152205469	38.7105475308051	-0.0200323102581783
2	1965	39.3346882320302	39.2801978282104	0.0544904038198301
3	1966	38.7501607969993	38.6836370943449	0.0665237026544362
4	1967	37.1828579973273	37.1848592657186	-0.00200126839126084
5	1968	35.8208877528316	35.8532595933479	-0.0323718405163146
6	1969	34.0344264980962	34.0984025865532	-0.0639760884569992
7	1970	32.0827658115209	32.091467200081	-0.00870138856004431
8	1971	32.7128384754236	32.7345969347068	-0.0217584592832761
9	1972	32.631627506655	32.6575444517936	-0.0259169451385439
10	1973	32.4390636717603	32.4680444177329	-0.0289807459726017

Showing 1 to 10 of 45 entries Previous 1 2 3 4 5 Next

IV.IV. Gráfico Comparativo: tasa media de ganancia observada versus la estimada

Hide

```
nuevos_datos <- new_predict[order(new_predict$Year),]
library(plotly)
fig <- plot_ly(nuevos_datos, x = ~Year)
fig <- fig %>% add_trace(y = ~Mean_Gross_Profit, name = "original", mode = 'lin
es')
fig <- fig %>% add_trace(y = ~predict_GP, name = "prediccion", mode = 'lines')
fig
```



A continuación, se realizarán pruebas de diagnóstico sobre la robustez estadística preliminar del modelo arrojada por el análisis de regresión antes expuesto.

IV.V. Verificando el supuesto de normalidad mediante un ajuste de distribución

Para esto, se comparará su ajuste Normal con su ajuste gamma y su ajuste LogNormal, en donde la calidad del ajuste se determina con base en alguno de los criterios de información.

Hide

```
library(fitdistrplus)
attach(DATOS)
#Ajuste Normal
fit_normal_MGP<-fitdist(Mean_Gross_Profit, "norm")
summary(fit_normal_MGP)
```

```
## Fitting of the distribution ' norm ' by maximum likelihood
## Parameters :
##       estimate Std. Error
## mean 31.126348  0.4468179
## sd    2.997346  0.3159478
## Loglikelihood:  -113.25   AIC:  230.4999   BIC:  234.1132
## Correlation matrix:
##      mean sd
## mean    1  0
## sd      0  1
```

Hide

```
#Ajuste Gamma
fit_gamma_MGP<-fitdist(Mean_Gross_Profit, "gamma")
summary(fit_gamma_MGP)
```

```
## Fitting of the distribution ' gamma ' by maximum likelihood
## Parameters :
##         estimate Std. Error
## shape 114.069449 24.0126750
## rate    3.664755  0.7731586
## Loglikelihood:  -111.856   AIC:  227.712   BIC:  231.3254
## Correlation matrix:
##           shape      rate
## shape 1.0000000 0.9978091
## rate  0.9978091 1.0000000
```

Hide

```
#Ajuste LogNormal
fit_lognormal_MGP<-fitdist(Mean_Gross_Profit, "lnorm")
summary(fit_lognormal_MGP)
```

```
## Fitting of the distribution ' lnorm ' by maximum likelihood
## Parameters :
##           estimate  Std. Error
## meanlog 3.43366471 0.013792536
## sdlog   0.09252314 0.009747669
## Loglikelihood:  -111.2538   AIC:  226.5076   BIC:  230.1209
## Correlation matrix:
##         meanlog sdlog
## meanlog       1     0
## sdlog         0     1
```

Los resultados indican que, aunque los datos se ajustan de forma estadísticamente significativa a una distribución normal, el ajuste gamma y lognormal es más adecuado. Existe una justificación teórica (desde la misma teoría de las probabilidades) para este fenómeno empírico, que también puede permitir explicar (como se verá más adelante) por qué la calidad de la predicción es tan alta a pesar de que se viola flagrantemente el supuesto de independencia lineal entre predictores (que el producto escalar entre los vectores que contienen los predictores, considerándolos de dos en dos, es nulo). Sin embargo, antes de ello se procederá a realizar pruebas de normalidad complementarias.

Hide

```
#Contraste de Normalidad Shaphiro-Wilk
shapiro.test(DATOS$Mean_Gross_Profit)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  DATOS$Mean_Gross_Profit
## W = 0.89447, p-value = 0.0006404
```

Hide

```
#Contraste de Normalidad de Kolmogórov-Smirnov
##Este contraste requiere de comparar con una distribución que se sepa normal
variablenormal<-rnorm(0,1, n=44)
ks.test(DATOS$Mean_Gross_Profit,variablenormal)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  DATOS$Mean_Gross_Profit and variablenormal
## D = 1, p-value = 2.22e-16
## alternative hypothesis: two-sided
```

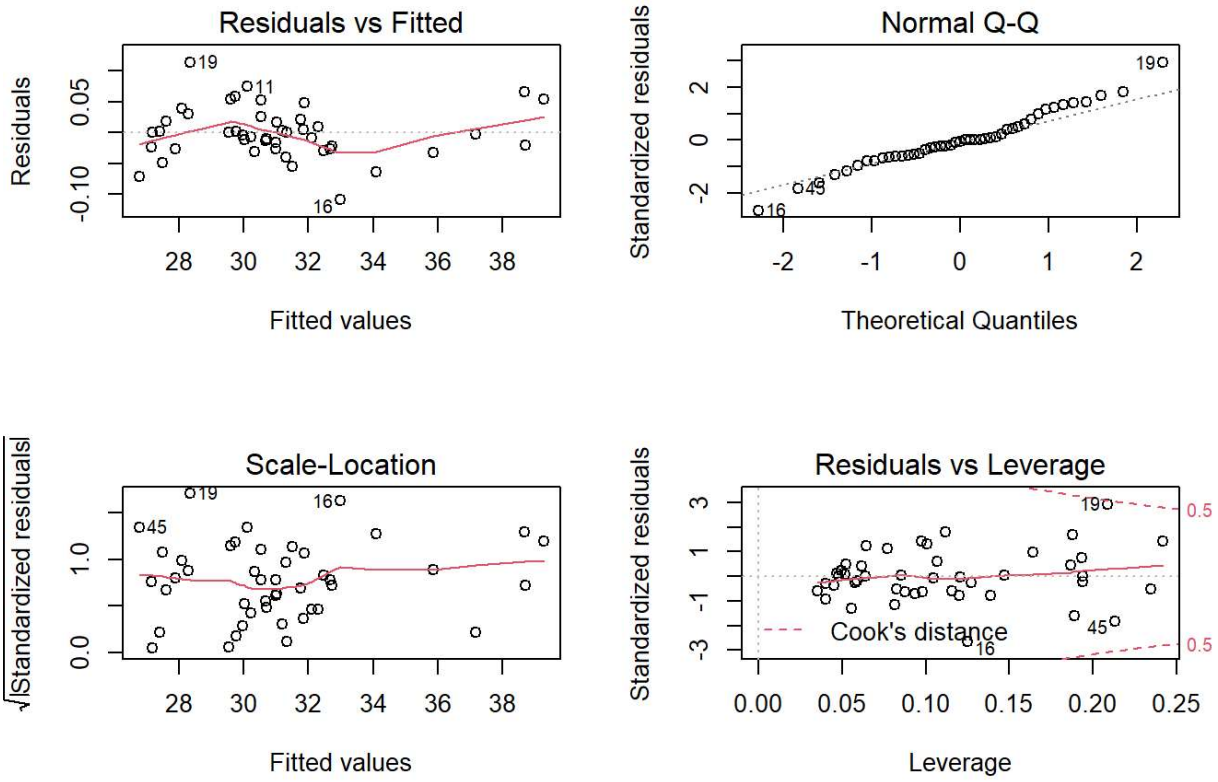
Puesto que según las pruebas antes especificadas no sólo la distribución normal no es la mejor distribución para el ajuste de los datos, sino que además, a juzgar por el bajísimo valor p en relación a niveles de significancia de 0.10, 0.05 y 0.01, por ejemplo. Por ello, se procederá a realizar pruebas de diagnóstico más profundas.

IV.VI. Pruebas de Diagnóstico de un Modelo de Regresión

IV.VI.I. Gráficos de Diagnóstico: Distancia D de Cook, Gráfico Q-Q y otros

Hide

```
par(mfrow=c(2,2))
plot(lmfit1)
```



Hide

```
library(mvinfluence)
CooksD <- cooks.distance(lmfit1)
influential <- CooksD[(CooksD > (3 * mean(CooksD, na.rm = TRUE)))]
influential
```

```
##         2         3         6        16        19        45
## 0.1294442 0.1306934 0.1219310 0.2012187 0.4505922 0.1783685
```

La distancia de Cook o la D de Cook es una estimación de uso común sobre la influencia de una observación cuando se realiza un análisis de regresión de mínimos cuadrados. En un análisis aplicado de mínimos cuadrados ordinarios, la distancia de Cook se puede utilizar de varias formas: para indicar observaciones influyentes cuyo impacto en el modelo merece la pena estudiar; o para indicar regiones del espacio de diseño (construido mediante las covariables) donde sería deseable poder obtener más observaciones. Puede entenderse como el resumen de qué tanto un modelo de regresión cambiaría cuando la i-ésima observación es removida o, de manera formal, mide el efecto de eliminar una observación en el vector de parámetros combinados.

(Statistics How To, 2016) señala que existen diversas interpretaciones de la distancia de Cook:

1. Una regla general es que las observaciones con una D de Cook mayor que 3 veces la media del conjunto de observaciones, μ, son posibles valores atípicos. Es la regla que se usará aquí y suele ser utilizada en el contexto del aprendizaje automático (véase https://towardsdatascience.com/identifying-outliers-in-linear-regression-cooks-distance-9e212e9136a (https://towardsdatascience.com/identifying-outliers-in-linear-regression-cooks-distance-9e212e9136a))

2. Una interpretación alternativa es investigar cualquier punto por encima de 4 / n, donde n es el número de observaciones.
3. Otros autores sugieren que se debe investigar cualquier D “grande”. ¿Qué tan grande es “demasiado grande”? El consenso parece ser que un valor D mayor que 1 indica un valor influyente, pero es posible que se deseen observar valores por encima de 0,5. También se debe investigar cualquier valor que sobresalga del otro.
4. Una forma alternativa (pero un poco más técnica) de interpretar D es encontrar el valor del percentil del valor atípico potencial utilizando la distribución F. Un percentil de más de 50 indica un punto muy influyente.

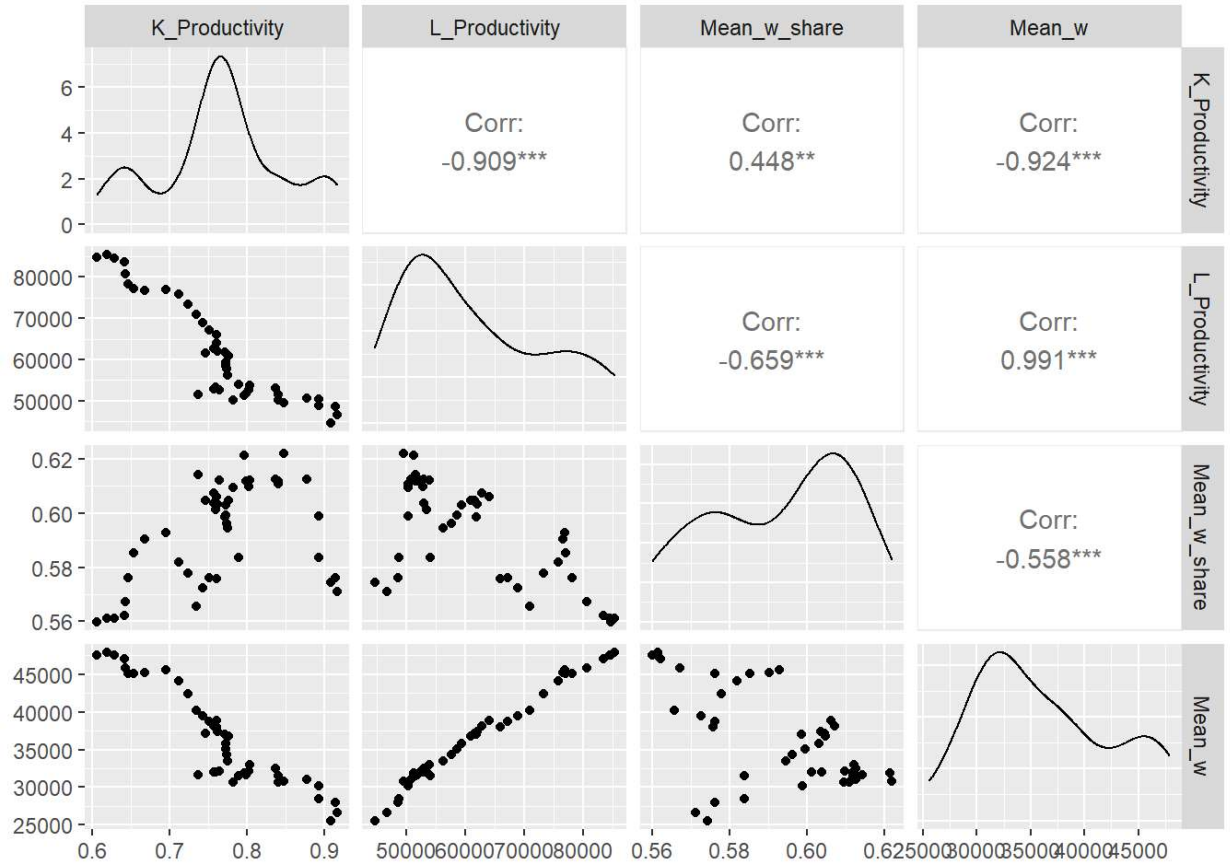
Complementariamente, otros textos como (Fox, 1991, pág. 34) establecen un umbral que puede ser no solo de 4 / N sino también de 4 / (N – k – 1), donde N es el número de observaciones y k el número de variables explicativas. En su caso, la última fórmula debería producir un umbral de alrededor de 0,1. Sin embargo, John Fox es bastante cauteloso cuando se trata de dar umbrales numéricos. En el lugar citado, aconseja el uso de gráficos y examinar con más detalle los puntos con “valores de D que son sustancialmente mayores que el resto”. Según Fox, los umbrales solo deberían usarse para mejorar las pantallas gráficas.

En los gráficos generados anteriormente, aquellas observaciones de alta *influencia* en el modelo (a esto en inglés se conoce como alto *leverage*) son aquellas que se encuentran fuera de la región comprendida entre las dos líneas rojas punteadas en la parte superior e inferior del gráfico titulado “Residuals vs Leverage”. En este caso, no parecen haber puntos fuera de tales líneas, sin embargo, el análisis formal revela (aunque desaconsejado por Fox) que las observaciones 2, 3, 6, 16, 19 y 45 son de alta influencia.

IV.VI.II. Prueba de Multicolinealidad

Hide

```
library(GGally)
ggpairs(nuevos_datos[,c(5:8)])
```



Hide

```
library(mctest)
omcdiag(lmfit1)
```

```
##
## Call:
## omcdiag(mod = lmfit1)
##
##
## Overall Multicollinearity Diagnostics
##
##                        MC Results detection
## Determinant |X'X|:         0.0001         1
## Farrar Chi-Square:       407.0887         1
## Red Indicator:             0.7756         1
## Sum of Lambda Inverse:  2939.2447         1
## Theil's Method:            0.8514         1
## Condition Number:        943.0074         1
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
```

Hide

```
imcdiag(lmfit1)
```

```
##
## Call:
## imcdiag(mod = lmfit1)
##
##
## All Individual Multicollinearity Diagnostics Result
##
##                      VIF    TOL         Wi         Fi Leamer    CVIF Klein
## Mean_w_share     41.3407 0.0242   551.3226   847.1542 0.1555 -0.0085     0
## L_Productivity 1621.2446 0.0006 22143.3424 34025.1358 0.0248 -0.3320     0
## K_Productivity    8.0950 0.1235    96.9647   148.9945 0.3515 -0.0017     0
## Mean_w         1268.5645 0.0008 17323.3810 26618.8537 0.0281 -0.2598     0
##                   IND1   IND2
## Mean_w_share    0.0018 1.0136
## L_Productivity 0.0000 1.0381
## K_Productivity 0.0090 0.9104
## Mean_w         0.0001 1.0379
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## * all coefficients have significant t-ratios
##
## R-square of y on all x: 0.9998
##
## * use method argument to check which regressors may be the reason of colline
arity
## ===================================
```

Como resultado del hecho de que únicamente el trabajo crea valor de uso y de cambio, mientras que el capital se limita a transferirlo, así como también por el hecho de los estímulos psicológicos que representa para el rendimiento de un trabajador el salario del cual este goza, existe alta multicolinealidad para 3 de las 4 variables estudiadas (salvo la productividad marginal del capital, cuyo FIV es menor a 10), que, como se adelantó y profundizará más adelante, se explica por el alto dependencia lineal entre las variables. Más adelante se profundizará sobre este aspecto. ##### IV.VI.III. Heterocedasticidad

Hide

```
library(lmtest)
bptest(lmfit1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lmfit1
## BP = 7.8008, df = 4, p-value = 0.09915
```

La prueba de White, conocida también como contraste de Breusch-Pagan Studentizado, establece como H_0 que la variabilidad de la varianza es nula, es decir, que la varianza del error de estimación para cualquier observación que se desee estimar será la misma, lo que se conoce

en homocedasticidad; como H_A establece lo opuesto. Así, sólo se puede descartar heterocedasticidad para un nivel de confianza del 0.90. Este problema se corregirá en el último inciso de esta sección.

IV.VI.IV. Pruebas de Autocorrelación

IV.VI.IV.I. Contraste de Durbin-Watson

Hide

```
library(car)
durbinWatsonTest(lmfit1)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1467137      1.636473   0.072
##  Alternative hypothesis: rho != 0
```

Hide

```
library(lmtest)
dwtest(lmfit1)
```

```
##
##  Durbin-Watson test
##
## data:  lmfit1
## DW = 1.6365, p-value = 0.03317
## alternative hypothesis: true autocorrelation is greater than 0
```

IV.VI.IV.II. Contraste de Breusch-Godfrey

Hide

```
library(lmtest)
bgtest(lmfit1, order=1)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  lmfit1
## LM test = 1.1327, df = 1, p-value = 0.2872
```

Hide

```
bgtest(lmfit1, order=2)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 2
##
## data:  lmfit1
## LM test = 1.2789, df = 2, p-value = 0.5276
```

Hide

```
bgtest(lmfit1, order=3)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 3
##
## data:  lmfit1
## LM test = 8.5259, df = 3, p-value = 0.03631
```

De lo anterior se verifica que existe autocorrelación serial, al menos de orden 1, 2 y 3, entre los términos del error estocástico.

IV.VI.V. Errores Estándar Robustos en presencia de Heterocedasticidad (Errores Estándar HAC)

Recuérdese que no es necesario el supuesto de homogeneidad de varianza para demostrar que los estimadores MCO son insesgados siempre que se trate de una muestra finita, así como tampoco para probar su consistencia asintótica,lo que sí es importante es lograr corregir los errores estandarizados obtenidos por esa vía, por MCO (véase http://www3.grips.ac.jp/~yamanota/Lecture_Note_9_Heteroskedasticity (http://www3.grips.ac.jp/~yamanota/Lecture_Note_9_Heteroskedasticity)); esto se logra precisamente a través de los llamados *errores estándar consistentes ante heterocedasticidad* (véase https://www.wikiwand.com/en/Heteroscedasticity-consistent_standard_errors (https://www.wikiwand.com/en/Heteroscedasticity-consistent_standard_errors) y https://www.r-econometrics.com/methods/hcrobusterrors/ (https://www.r-econometrics.com/methods/hcrobusterrors/)). Lo anterior se justifica por el hecho de que, aunque la heterocedasticidad no produce estimaciones de MCO sesgadas, conduce a un sesgo en la matriz de varianza-covarianza. Esto significa que ya no se puede confiar en los métodos de prueba de modelos estándar, como las pruebas t o las pruebas F.

Hide

```
library(sandwich)
coeftest(lmfit1)
```

```
##
## t test of coefficients:
##
##                   Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)     7.1532e+01  1.4834e+00  48.221 < 2.2e-16 ***
## Mean_w_share   -1.1876e+02  2.3208e+00 -51.175 < 2.2e-16 ***
## L_Productivity -3.9514e-04  2.2014e-05 -17.949 < 2.2e-16 ***
## K_Productivity  3.9903e+01  2.3425e-01 170.339 < 2.2e-16 ***
## Mean_w          6.5361e-04  3.7195e-05  17.573 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Puesto que la sintaxis `coeftest` es una función genérica de R que realiza contrastes z de Wald y cuasi-t de Wald, que son contrastes destinados a verificar la significancia global y local de los coeficientes de regresión (véase https://www.statisticshowto.com/wald-test/ (https://www.statisticshowto.com/wald-test/)) utilizando en lugar de los errores de la regresión mínimo cuadrática ordinaria aquellos errores robustos de White HC, de los resultados

estadísticos anteriores se deriva que sí es posible realizar una estimación de parámetros robustos (donde su significancia se preserve, en este caso hasta un nivel de confianza menor que 0.0001, aún en presencia de heterocedasticidad y autocorrelación -como puede verificarse en https://cran.r-project.org/web/packages/sandwich/sandwich.pdf (https://cran.r-project.org/web/packages/sandwich/sandwich.pdf), p.13-).

Con los errores así calculados, en adición a su robustez en presencia de heterocedasticidad, se añade la propiedad de que no-correlación serial de orden n-ésimo entre los términos del error, así como también la propiedad de normalidad en la distribución de los mismos, ambas cuestiones a cualquier nivel de significancia que se escoja.

V. CONCLUSIONES

Como se señala en (Nabi, 2021, págs. 9-10), desde el punto de vista formal de la teoría estadística, especialmente la clásica, el teorema central del límite, la ley de los grandes números y el teorema ergódico exigen inexorablemente que las variables aleatorias posean independencia estocástica, sin embargo, también es importante preguntarse ¿cuál es el desempeño de estos teoremas en ausencia de independencia estocástica entre las variables de estudio? Se podría hacer aquí un repaso exhaustivo de las múltiples investigaciones, como por ejemplo la de (Greene, 2012, págs. 655-658), en las que se observa un buen desempeño en estudios Monte Carlo de los estadísticos de prueba a pesar de la violación de ciertos supuestos que teóricamente condicionan significativamente su robustez empírica. Otro ejemplo de lo anterior podría formularse para el caso de la violación del supuesto de independencia estocástica, sin embargo, para este caso en particular es ampliamente conocido que, en ocasiones, es posible economizar los esfuerzos que tal tarea demanda gracias a que esta dependencia estocástica no es relevante a nivel aplicado siempre que esta dependencia no sea significativa en términos del impacto que tiene en la de la varianza (de ahí la importancia de las pruebas que consisten en medir los factores de inflación de varianza), puesto que la dependencia lineal es un problema en la predicción fundamentalmente por cuanto puede volver indeseablemente volátil el comportamiento de la varianza, lo que resta poder predictivo en el marco de los modelos estadísticos clásicos, como se adelantó. Así, los factores de inflación de varianza (FIV) entran en juego para determinar si esta correlación entre los predictores está controlada o no. En este caso en particular los FIV muestran que la multicolinealidad no está bajo control, aunque a pesar de ello la capacidad predictiva del modelo es sumamente alta, la muestra es grande (al menos tratándose de series temporales, distinto caso sería si este tamaño de muestra se utilizase en el contexto de la bioestadística, por ejemplo) y se pueden obtener estimadores robustos HAC. Complementariamente, debe mencionarse que, en el contexto de la Bioestadística (de su generalidad, al menos), al igual que como se mencionó en el contexto de la econometría, la multicolinealidad se considera un problema únicamente si es moderada o alta, como se verifica en (Penn State University, Eberly College of Science, 2018) y (Simon Fraser University, 2011). Como señalan (Gujarati & Porter, 2010, pág. 328), este indicador toma su nombre del hecho que al incrementar la colinealidad incrementa la varianza del estimador también y, en el límite, se vuelve infinito. Ello es así puesto que $VIF_{ij} = 1/(1 - r_{ij}^2)$, en donde r_{ij}^2 es el coeficiente de determinación entre la variable explicativa i-ésima y la variable explicativa j-ésima.

Lo anterior no significa que la teoría estadística, en última instancia, no tenga relevancia y que lo único que verdaderamente importa son las aplicaciones. Lo que ocurre es que, al igual que la luz y el sonido cuando la Naturaleza parte con furia eléctrica el cielo, la práctica “viaja más rápido” que la teoría, por lo que las deficiencias del paradigma (en el sentido de Thomas Kuhn) se evidencian primero en las aplicaciones y es tras ello que se teoriza sobre los nuevos hallazgos y se consolida el nuevo paradigma sobre la base del paradigma anterior, puesto que los

paradigmas no son más que "(...) una o más realizaciones científicas pasadas, realizaciones que alguna comunidad científica particular reconoce, durante cierto tiempo, como fundamento para su práctica posterior." (Kuhn, 2011, pág. 33). Así las cosas, en la investigación (Dedecker & Prieur, 2007) se demuestra la existencia empírica de un teorema central del límite para distribuciones multidimensionales, en la investigación (Andrews, 1991) se demuestra la existencia del TCL para procesos empíricos indizados por funciones suaves (en las que las variables estocásticas subyacentes pueden ser temporalmente dependientes y distribuidas no-idénticamente), en la investigación (KO, RYU, KIM, & CHOI, 2007) se plantea una versión general del TCL para sumas ponderadas de variables aleatorias con dependencia lineal cuadrática negativa (la idea central de esta investigación es ponderar los elementos de una matriz triangular no negativa, cuyos elementos poseen una medida finita, ponderando los elementos de esta matriz con una sucesión de dependencias lineales cuadráticas negativas), en la investigación (LI, 2015) se plantea una demostración del teorema central del límite bajo las condiciones especificadas (m-ésimas variables dependientes bajo estructuras matemáticas conocidas como espacios de expectativas sublineales), en la investigación (Berk, 1973) se plantea una generalización del TCL para m-ésimas variables dependientes sin considerar ningún acotamiento local del espacio en donde se plantea tal generalización, en la investigación (Parzen, 1957) se plantea una demostración del cumplimiento del TCL para el caso de procesos estocásticos multilineales, en la investigación (Godwin & Zaremba, 1961) se plantea la verificación del TCL cuando las variables, aunque son dependientes entre sí, sólo lo son parcialmente (la razón por la que los autores usan "partly dependent" en lugar de "partially dependent" obedece a que, según (Alan, 2011), "parly" se usa cuando un objeto es parte de un todo tangible y parcialmente cuando no lo es); estas investigaciones únicamente son algunas de las múltiples realizadas en la misma dirección (el lector puede encontrar gran vastedad de ellas en algún repositorio virtual, como por ejemplo www.jstor.org). Por tanto, el estado del arte parecería revelar que existe evidencia académica sobre la problemática planteada que apunta a que es cuestión de tiempo antes que los paradigmas de la Estadística Clásica se generalicen y robustezcan a nivel global.

Por supuesto, esto no contradice el hecho de que son poco usuales los escenarios en que se puede tener un alto poder predictivo y una alta confiabilidad en dicho poder en presencia de multicolinealidad. Como señalan (Gujarati & Porter, 2010, págs. 320-342), la multicolinealidad puede obedecer a los siguientes factores:

1. El método de recolección de la información. Por ejemplo, la obtención de muestras en un intervalo limitado de valores tomados por las regresoras en la población (muestra pequeña). Este no es el caso aquí, como se indicó antes.

2. Restricciones en el modelo o en la población objetivo de muestreo. Por ejemplo, en la regresión del consumo de electricidad sobre el ingreso y el tamaño de las viviendas hay una restricción física en la población, pues las familias con ingresos más altos suelen habitar en viviendas más grandes que las familias con ingresos más bajos. **Esta es la fuente de problema de multicolinealidad estructural del modelo, puesto que el capital es crisálida del trabajo pasado**.

3. Especificación del modelo. Por ejemplo, la adición de términos polinomiales a un modelo de regresión, en especial cuando el rango estadístico de las variables explicativas es pequeño. Esto puede verificarse mediante el contraste RESET diseñado por J. B. Ramsey para detectar errores de especificación por omisión de variables, por correlación entre los predictores o la forma funcional. A juzgar por el elevado coeficiente de determinación obtenido (0.9998) es poco verosimil esperar que se hayan omitido variables relevantes,

mientras que a juzgar por el bajo error de predicción no parecería haber un problema con la forma funcional del modelo, aunque no puede decirse lo mismo con el caso de correlación entre predictores (que ya se conoce como cierta).

Hide

```
library(fRegression)
resettest(lmfit1, power = 2:3, type = c("fitted", "regressor",
  "princomp"), data = list())
```

```
##
##  RESET test
##
## data:  lmfit1
## RESET = 12.497, df1 = 2, df2 = 38, p-value = 6.75e-05
```

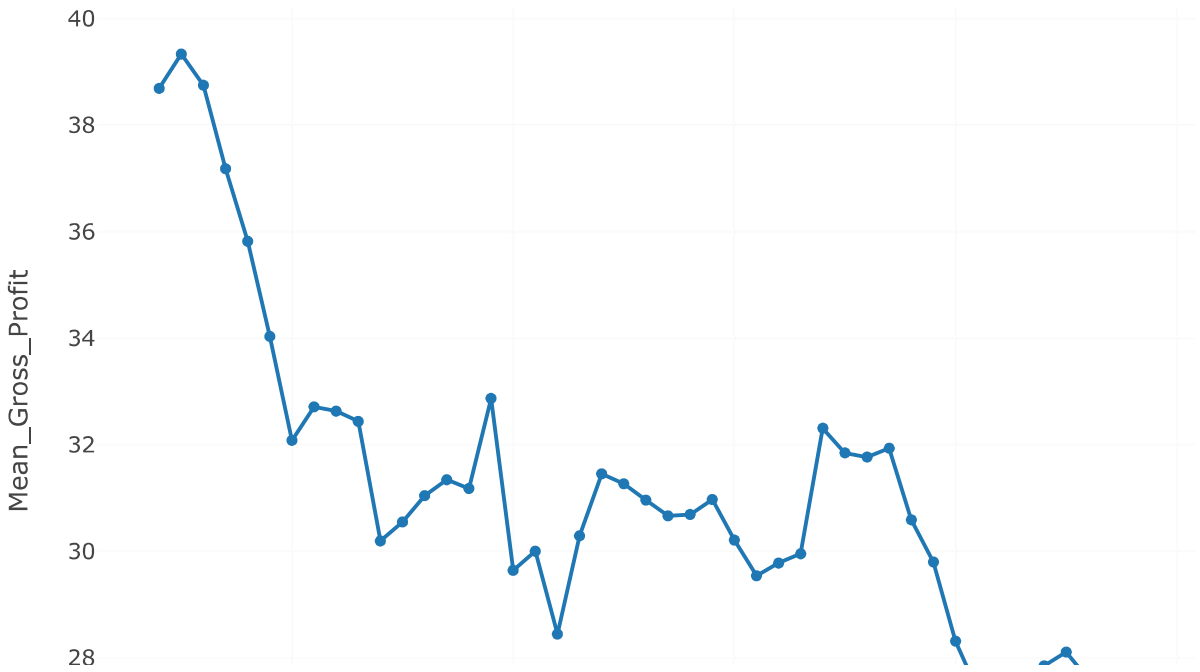
Se verifica que existe un problema de especificación, que parecería ser debido al problema de multicolinealidad abordado.

4. Un modelo sobredeterminado. Esto sucede cuando el modelo tiene más variables explicativas que el número de observaciones de las que hace uso. Esto puede suceder en investigación médica, donde en ocasiones hay un número reducido de pacientes sobre quienes se reúne información respecto de un gran número de variables. Esto está descartado aquí por la equivalencia que se presenta entre el coeficiente de determinación inicial y su versión ajustada.

5. Que las regresoras compartan una tendencia común, es decir, que todas aumenten o disminuyan a lo largo del tiempo.

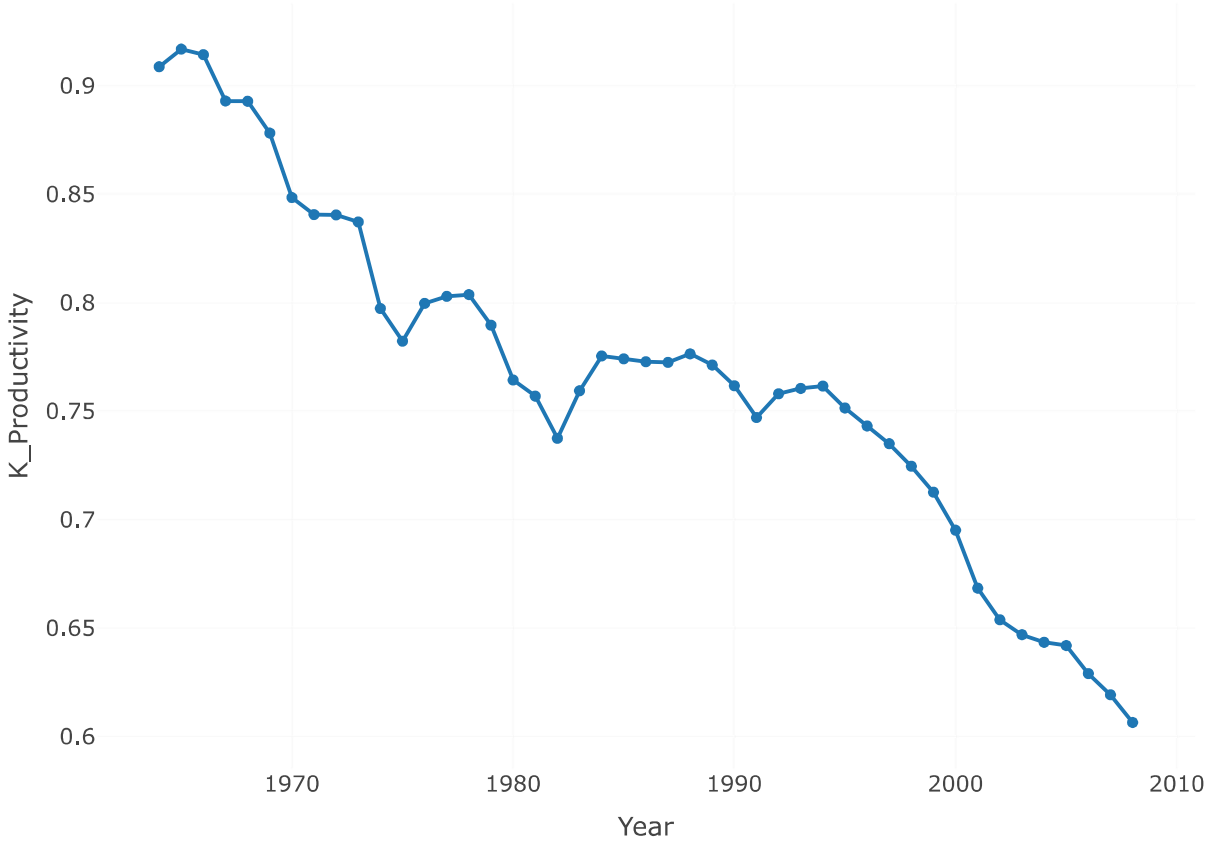
Hide

```
attach(DATOS)
fig <- plot_ly(DATOS, x = ~Year)
fig <- fig %>% add_trace(y = ~Mean_Gross_Profit, mode = 'lines+markers')
fig
```



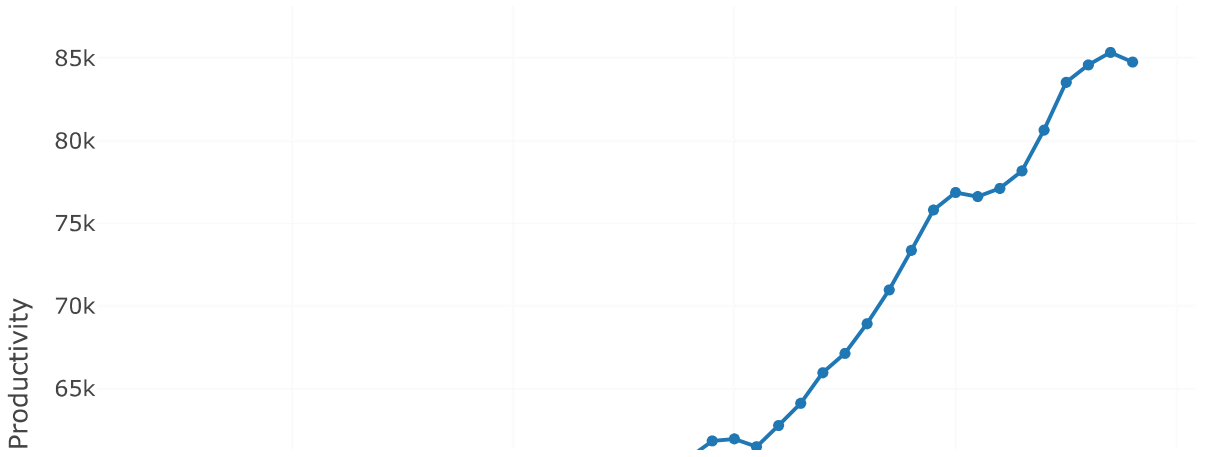
Hide

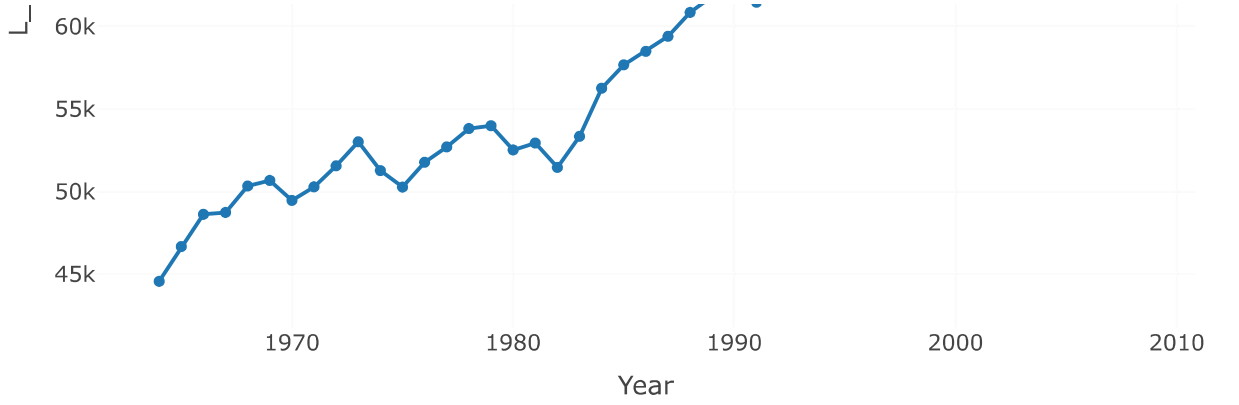
```
fig <- plot_ly(DATOS, x = ~Year)
fig <- fig %>% add_trace(y = ~K_Productivity, mode = 'lines+markers')
fig
```



Hide

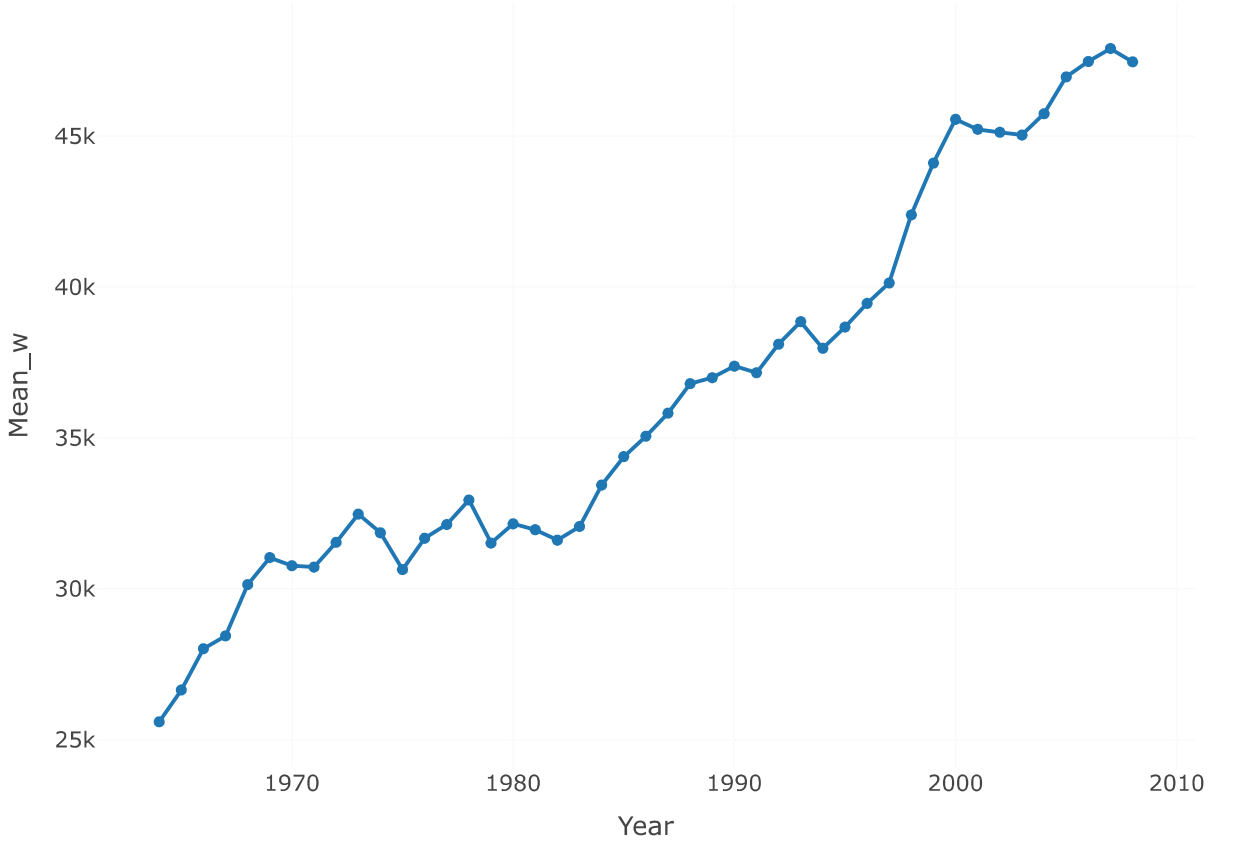
```
fig <- plot_ly(DATOS, x = ~Year)
fig <- fig %>% add_trace(y = ~L_Productivity, mode = 'lines+markers')
fig
```





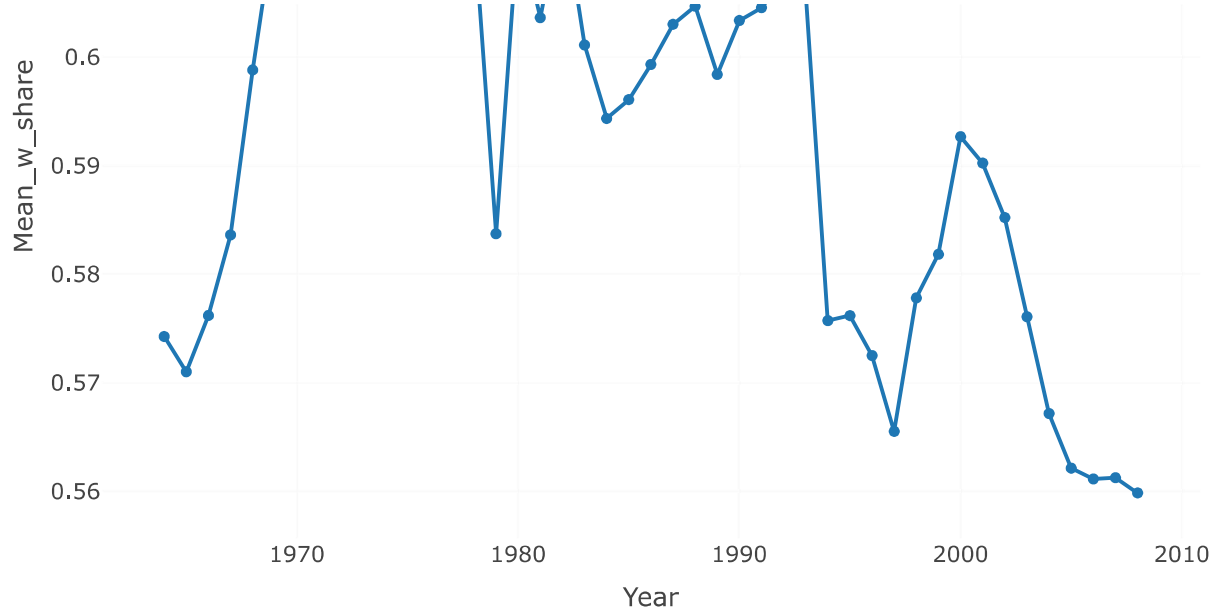
Hide

```
fig <- plot_ly(DATOS, x = ~Year)
fig <- fig %>% add_trace(y = ~Mean_w, mode = 'lines+markers')
fig
```



Hide

```
fig <- plot_ly(DATOS, x = ~Year)
fig <- fig %>% add_trace(y = ~Mean_w_share, mode = 'lines+markers')
fig
```



Como puede observarse, la participación media de salarios en el producto interno bruto (Mean_w_share) va a la baja (igual que la tasa media de ganancia, dicho sea de paso -lo que es congruente con la tesis de las raíces unitarias y de la no-restauración del producto interno bruto medio que relataba Blanchard-), la masa salarial promedio se eleva (lo que no es extraño considerando que el poder adquisitivo de la clase trabajadora puede subir sin que por ello deje de implicar un empobrecimiento relativo a los valores producidos por la misma y sin olvidar que con independencia de lo que ocurre en), la productividad media del trabajo va al alza, mientras que la productividad media del capital tiende a la baja (esto es congruente con la financiarización de la economía; nótese que las productividades medias del trabajo y del capital van en sentido opuesto, lo cual podría deberse (habría que realizar una investigación más amplia, *i.e.*, para más países y si fuese posible para una mayor cantidad de años) a que, como es conocido entre algunos economistas, Estados Unidos en los últimos años se ha enfocado en la financiarización especulativa de su economía en lugar del desarrollo tecnológico que lo caracterizó (junto la financiarización especulativa y las guerras expansionistas, estas últimas dos cosas las conserva vigorosamente) hasta más o menos pasado el primer lustro de este milenio. Así, no parece ser que todas las variables compartan una tendencia común (aunque es cierto que existen métodos más robustos, *i.e.*, no-descriptivos, para extraer tendencias).

Matemáticamente hablando, las estimaciones en presencia de multicolinealidad perfecta impiden la estimación de los coeficientes de regresión, puesto que algebraicamente hablando el resultado de su estimación será la operación cero entre cero, cuyo resultado es indeterminado. Más en profundidad, este cociente refleja que en presencia de multicolinealidad perfecta no puede obtenerse una solución única para las ecuaciones por medio de las cuales se encuentran los coeficientes de regresión o parámetros del modelo. Por su parte, la estimación en presencia de multicolinealidad alta, pero imperfecta, es posible de realizar matemáticamente hablando, con independencia de la validez de los resultados de la estimación. Finalmente, las consecuencias prácticas de la multicolinealidad son:

1. Aunque los estimadores de MCO son MELI, presentan varianzas y covarianzas grandes que dificultan la estimación precisa. De lo anterior se desprende que el hecho de que los estimadores sean eficientes, no significa que sean precisos. Para el caso del modelo aquí planteado, todo parecería indicar que los estimadores son precisos.
2. Debido a lo anterior, los intervalos de confianza tienden a ser mucho más amplios, lo cual propicia una aceptación más fácil de la “hipótesis nula cero” (es decir, que el verdadero coeficiente poblacional es cero). Como se presenta a continuación, no parecería ser este el

caso, al menos no a un nivel que pudiese parecer preocupante a simple vista (por supuesto, deben considerarse también las demás pruebas antes realizadas).

Hide

```
round(confint(lmfit1),4)
```

```
##                     2.5 %    97.5 %
## (Intercept)      68.5337   74.5299
## Mean_w_share   -123.4550 -114.0742
## L_Productivity   -0.0004   -0.0004
## K_Productivity   39.4291   40.3760
## Mean_w            0.0006    0.0007
```

3. También debido a la primera consecuencia, ocurre que la razón t de uno o más coeficientes tiende a ser estadísticamente no significativa. No es el caso en el modelo aquí presentado.
4. Aunque la razón t de uno o más coeficientes sea estadísticamente no significativa, el coeficiente de determinación (que es la medida global de bondad de ajuste) puede ser muy alto. En este caso, las razones t de los coeficientes son altamente significativas a nivel estadístico en conjunción con un coeficiente de determinación ajustado significativamente elevado (.
5. Los estimadores MCO y sus errores estándar son sensibles a pequeños cambios en los datos. Este aspecto parecería descartarse a juzgar por lo observado al estimar las distancias D de Cook.

Como se señala en (Salmerón Gómez, Blanco Izquierdo, & García García, 2016, págs. 3-4), existen dos tipos de multicolinealidad:

1. La multicolinealidad sistemática, la cual es debida a un problema estructural, es decir, a la alta correlación lineal de las variables exógenas que en concreto se han especificado. A juicio del autor de la presente investigación, este es el tipo de multicolinealidad que exhibe el presente modelo.
2. La multicolinealidad errática, que es debidamente a un problema numérico. Para este segundo tipo de multicolinealidad es que, según los autores citados en el lugar referido, el famoso econometrista Arthur Goldberger acuñó el término micronumerosidad. Es importante destacar que, según (Gujarati & Porter, 2010, pág. 332), Goldberger acuñó dicho término como una parodia de las consecuencias de la multicolinealidad.

La parodia de Goldberger a las consecuencias de la multicolinealidad se complementan con otras referencias presentadas por (Gujarati & Porter, 2010, pág. 320), específicamente una de Edward E. Leamer (reconocido economista, estadístico y catedrático de UCLA), otra de Christopher Achen (reconocido politólogo, estadístico y catedrático de Princeton), otra de Olivier Blanchard (anterior economista jefe del Fondo Monetario Internacional, catedrático de Harvard y MIT, así como uno de los economistas más citados del mundo) y otra de Peter Kennedy (economista, estadístico y catedrático de la Simon Fraser University).

1. Leamer señala que no hau una expresión más errónea, tanto en los libros de texto de econometría como en la bibliografía aplicada, que la de “problema de multicolinealidad”. Es un hecho que muchas variables explicativas presentan un alto grado de colinealidad; asimismo, resulta muy claro que existen diseños experimentales que serían mucho más convenientes que los diseños que proporciona la experimentación natural (es decir, la

muestra disponible). No obstante, no es nada constructivo quejarse de la aparente malevolencia de la naturaleza, y los remedios ad hoc para un mal diseño -como una regresión por pasos o una regresión en cadena- pueden ser desastrosamente inapropiados. Es mejor aceptar de plano que los datos que no se recopilaron mediante experimentos diseñados a veces no proporcionan mucha información sobre los parámetros de interés.

2. Por su parte, Achen señala que los novatos en el estudio de la metodología en ocasiones se preocupan porque sus variables independientes estén correlacionadas: el llamado problema de multicolinealidad. Sin embargo, la multicolinealidad no viola los supuestos básicos de regresión. Se presentarán estimaciones consistentes e insesgadas y sus errores estándar se estimarán de forma correcta. El único efecto de la multicolinealidad tiene que ver con la dificultad de obtener coeficientes estimados con errores estándar pequeños [aquí “pequeños” no hace referencia a mínimo, que es un concepto matemático, sino que hace referencia a un intervalo de confianza que contenga los valores de los parámetros que para fines de estimación y control de políticas pueda ser considerado pequeño (según el marco científico de referencia, las necesidades objetivas que motivaron la investigación, el criterio experto y, por supuesto, la evidencia objetiva -que la constituyen los hechos, más allá de los datos-)]. Sin embargo, se presenta el mismo problema al contar con un número reducido de observaciones o al tener variables independientes con varianzas pequeñas. De hecho, en el nivel teórico, los conceptos de multicolinealidad, número reducido de observaciones y varianzas pequeñas en las variables independientes forman parte esencial del mismo problema. Por tanto, la pregunta “¿qué debe hacerse entonces con la multicolinealidad?” es similar a “¿qué debe hacerse si no se tienen muchas observaciones?” Al respecto no hay una respuesta estadística.

3. Blanchard señala que “Cuando los estudiantes efectúan por primera vez la regresión de MCO, el primer problema que suelen afrontar es el de multicolinealidad. Muchos concluyen que hay algo malo con los MCO; otros recurren a nuevas y con frecuencia creativas técnicas a fin de darle la vuelta al problema. Pero eso está mal. La multicolinealidad es la voluntad de Dios, no un problema con los MCO ni con la técnica estadística general.” Esto podría ser el caso para la multicolinealidad exhibida entre los predictores aquí utilizados.

4. Kennedy señala que en el trabajo aplicado que se nutre de información secundaria (la información recopilada por alguna institución, como la información del producto nacional bruto recopilada por el gobierno), es posible que un investigador por sí solo no pueda hacer gran cosa sobre el tamaño de la información muestral, y quizá deba enfrentar la estimación de problemas lo bastante importantes (en términos de las consecuencias prácticas de una mala estimación) para justificar su tratamiento (el de la multicolinealidad) como una violación del modelo clásico de regresión lineal.

Adicionalmente, señalan Gujarati y Porter (G&P) que:

1. Es cierto que aún en el caso de casi multicolinealidad los estimadores de MCO son insesgados. Pero el insesgamiento es una propiedad multimuestral o de muestreo repetido. Esto significa que, si mantenemos fijos los valores de X, si obtenemos muestras repetidas y calculamos los estimadores de MCO para cada una de esas muestras, el promedio de los valores muestrales se aproximará a los verdaderos valores poblacionales de los estimadores a medida que aumenta el número de las muestras. Pero esto nada dice sobre las propiedades de los estimadores en una muestra concreta. Este punto en particular señalado por G&P es de especial interés en investigaciones como esta, que en general es una de esas investigaciones en las que no es posible obtener varias muestras; esto es así

para esta investigación porque los datos macroeconómicos, aunque son de carácter objetivo, no son recogidos directamente de la Naturaleza, sino que son construidos por el ser humano mediante la inversión de cantidades de recursos (técnicos, económicos y financieros) significativas en ello, por lo que su disponibilidad y su tamaño está sujeto a múltiples restricciones que otras ciencias no enfrentan respecto a la calidad y cantidad de los datos que manipulan. Por supuesto, que sean de carácter objetivo no significa que deban estudiarse necesariamente sin procesamiento alguno, ello dependerá de múltiples factores.

2. La multicolinealidad no destruye la propiedad de varianza mínima; en la clase de los estimadores lineales insesgados, los estimadores de MCO tienen varianza mínima; es decir, son eficientes. Pero esto no significa que la varianza de un estimador de MCO necesariamente sea pequeña (en relación con el valor del estimador) en cualquier muestra dada, como demuestran los autores citados en el lugar citado.

3.La multicolinealidad es en esencia un fenómeno de regresión muestral en el sentido en que, aunque las variables X no estén linealmente relacionadas en la población, pueden estarlo en la muestra particular disponible: cuando se postula la función de regresión teórica o poblacional (FRP), se considera que todas las variables X incluidas del modelo ejercen una influencia separada o independiente sobre la variable dependiente Y. Pero puede suceder que en cualquier muestra dada con que se pruebe la FRP, alguna o todas las variables X sean tan colineales que no sea posible aislar su influencia individual sobre Y.

Finalmente, es recomendable investigar la viabilidad relativa a la utilización de los modelos lineales generalizados, puesto que los ajustes de distribución hechos y la literatura disponible, específicamente (Farjoun & Marchover, 1983) y (Wells, 2007) indican que la tasa media de ganancia se distribuye como una gamma, por lo que con un modelo lineal generalizado con función enlace canónica para la distribución gamma podrían obtenerse aún mejores resultados; otras opciones son, por ejemplo, probar el ajuste con modelos de cópulas, modelos con heterocedasticidad condicional autorregresiva (modelos ARCH).

VI. REFERENCIAS

Alan. (25 de Julio de 2011). ENGLISH LENGUAGE & USAGE. Obtenido de Stack Exchange: https://english.stackexchange.com/questions/35508/difference-between-partly-and-partially# (https://english.stackexchange.com/questions/35508/difference-between-partly-and-partially#):~:text=Use%20partly%20when%20the%20%22in,it's%20also%20%22partly%20closed%22.
Andrews, D. W. (1991). An Empirical Process Central Limit Theorem for Dependent Non-identically Distributed Random Variables . Journal of Multivariate Analysis, 187-203.

Aston, T. H., & Philpin, C. H. (1987). The Brenner Debate: Agrarian Class Structure and Economic Development in Pre-industrial Europe. Cambridge: Past and Present Publications. Obtenido de https://www.cambridge.org/core/books/brenner-debate/A44B7FC72563D885578E901E188924EF (https://www.cambridge.org/core/books/brenner-debate/A44B7FC72563D885578E901E188924EF)

Berk, K. (1973). A CENTRAL LIMIT THEOREM FOR m-DEPENDENT RANDOM VARIABLES WITH UNBOUNDED m. The Annals of Probability, 1(2), 352-354.

Blanchard, O. (2 de Septiembre de 2009). Sustaining a Global Recovery. Obtenido de Finance & Development - International Monetary Fund: https://www.imf.org/external/pubs/ft/fandd/2009/09/pdf/blanchard.pdf

(https://www.imf.org/external/pubs/ft/fandd/2009/09/pdf/blanchard.pdf)

Cockshott, P., & Cottrell, A. (2005). Robust correlations between prices and labor values. Cambridge Journal of Economics, 309-316.

Cockshott, P., Cottrell, A., & Valle Baeza, A. (2014). The Empirics of the Labour Theory of Value: Reply to Nitzan and Bichler. Investigación Económica, 115-134.

Dedecker, J., & Prieur, C. (2007). An empirical central limit theorem for dependent sequences. Stochastic Processes and their Applications, 117, 121-142.

Dobb, M. (2008). Studies in the Development of Capitalism. Whitefish, Montana, United States: Kessinger Publishing, LLC. Obtenido de http://digamo.free.fr/dobb1946.pdf (http://digamo.free.fr/dobb1946.pdf)

Duménil, G., & Lévy, D. (1998). The Dynamics of Historial Tendencies in Volume III of Capital: An Application to the US Economy since the Civil War. En R. Bellafoiore, Marxian Economics, A Reappraisal (Vol. II). Essays on Volume III of Capital Profit, Prices and Dynamics (págs. 209-224). New York: MACMILLAN PRESS LTD.

Emmanuel, A. (1972). El Intercambio Desigual. Ensayo sobre los antagonismos en las relaciones económicas internacionales. México, D.F.: Siglo Veintiuno Editores, S.A.

Farjoun, E., & Marchover, M. (1983). Laws of Chaos. A Probabilistic Approach to Political Economy. Londres: Verso Editions and NLB.

Fox, J. (1991). Regression Diagnostics. California: SAGE Publications.

Freeman, A., & Carchedi, G. (1995). Marx and Non Equilibrium Economics. (A. Freeman, & G. Carchedi, Edits.) Aldershot, Hampshire, United Kingdom: Edward Elgar Publishing Limited.

Godwin, H., & Zaremba, S. (1961). A Central Limit Theorem for Partly Dependent Variables. The Annals of Mathematical Statistics, 32(3), 677-686.

Greene, W. (2012). Econometric Analysis (Seventh ed.). Harlow, Essex, England: Pearson Education Limited.

Gujarati, D., & Porter, D. (8 de Julio de 2010). Econometría (Quinta ed.). México, D.F.: McGrawHill Educación. Obtenido de Homocedasticidad.

Işıkara, G., & Mokreb, P. (2021). Price-Value Deviations and the Labour Theory of Value: Evidence from 42 Countries, 2000–2017. Review of Radical Political Economy, 1-15. Obtenido de https://www.tandfonline.com/doi/full/10.1080/09538259.2021.1904648 (https://www.tandfonline.com/doi/full/10.1080/09538259.2021.1904648)

Jiménez, F. (2011). Crecimiento Económico. Enfoques y Modelos. Lima: Pontificia Universidad Católica del Perú.

KO, M.-H., RYU, D.-H., KIM, T.-S., & CHOI, Y.-K. (2007). A CENTRAL LIMIT THEOREM FOR GENERAL WEIGHTED SUMS OF LNQD RANDOM VARIABLES AND ITS APPLICATION. ROCKY MOUNTAIN JOURNAL OF MATHEMATICS, 37(1), 259-268.

Kuhn, T. (2011). La Estructura de las Revoluciones Científicas. México, D.F.: Fondo de Cultura Económica.

LI, X.-p. (2015). A Central Limit Theorem for m-dependent Random Variables under Sublinear Expectations. Acta Mathematicae Applicatae Sinica, 31(2), 435-444. doi:10.1007/s10255-015-0477-1 (doi:10.1007/s10255-015-0477-1)

Marquetti, A., & Foley, D. (25 de Marzo de 2012). Extended Penn World Tables. Obtenido de Economic Growth Data assembled from the Penn World Tables and other sources : https://sites.google.com/a/newschool.edu/duncan-foley-homepage/home/EPWT (https://sites.google.com/a/newschool.edu/duncan-foley-homepage/home/EPWT)

Marx, K. (2010). El Capital (Vol. III). México, D.F.: Fondo de Cultura Económica.

Nabi, I. (2020). SOBRE LA LEY DE LA TENDENCIA DECRECIENTE DE LA TASA MEDIA DE GANANCIA. Raíces Unitarias y No Estacionariedad de las Series de Tiempo. Documento Inédito. Obtenido de https://marxianstatistics.files.wordpress.com/2020/12/analisis-del-uso-de-la-prueba-de-hipotesis-en-el-contexto-de-la-especificacion-optima-de-un-modelo-de-regresion-isadore-nabi-2.pdf (https://marxianstatistics.files.wordpress.com/2020/12/analisis-del-uso-de-la-prueba-de-hipotesis-en-el-contexto-de-la-especificacion-optima-de-un-modelo-de-regresion-isadore-nabi-2.pdf) Nabi, I. (11 de Abril de 2021). Sobre la Creación y Destrucción de Valor en los Sistemas de Economía Política Capitalista en Particular y en los Sistemas Económicos En General. Obtenido de El Blog de Isadore Nabi: https://marxianstatistics.com/2021/04/11/sobre-la-creacion-y-destruccion-de-valor-en-los-sistemas-de-economia-politica-capitalista-en-particular-y-en-los-sistemas-economicos-en-general/ (https://marxianstatistics.com/2021/04/11/sobre-la-creacion-y-destruccion-de-valor-en-los-sistemas-de-economia-politica-capitalista-en-particular-y-en-los-sistemas-economicos-en-general/)

Parzen, E. (1957). A Central Limit Theorem for Multilinear Stochastic Processes. The Annals of Mathematical Statistics, 28(1), 252-256.

Penn State University, Eberly College of Science. (2018). 10.4 - Multicollinearity. Obtenido de Lesson 10: Regression Pitfalls: https://online.stat.psu.edu/stat462/node/177/ (https://online.stat.psu.edu/stat462/node/177/)

Ricardo, D. (2004). The Principles of Political Economy and Taxation. Mineola, New York, United States: Dover Publications.

Salmerón Gómez, R., Blanco Izquierdo, V., & García García, C. (2016). Micronumerosidad aproximada y regresión lineal múltiple. Anales de ASEPUMA(24), 1-17. Obtenido de https://dialnet.unirioja.es/descarga/articulo/6004585.pdf (https://dialnet.unirioja.es/descarga/articulo/6004585.pdf)

Samuelson, P. (The Quarterly Journal of Economics). A Summing Up. Oxford: Oxford University Press. Obtenido de https://www.jstor.org/stable/1882916 (https://www.jstor.org/stable/1882916)

Sánchez, C., & Ferràndez, M. N. (Octubre-diciembre de 2010). Valores, precios de producción y precios de mercado a partir de los datos de la economía española. Investigación Económica, 87-118. Obtenido de https://www.jstor.org/stable/42779601?seq=1 (https://www.jstor.org/stable/42779601?seq=1)

Sánchez, C., & Montibeler, E. E. (2015). La teoría del valor trabajo y los precios en China. Economia e Sociedade, 329-354.

Shaikh, A. (1998). The Empirical Strength of the Labour Theory of Value. En R. Bellofiore, Marxian Economics: A Reappraisal. Essays on Volume III of Capital Profit, Prices and Dynamics (págs. 225-251). New York: MACMILLAN PRESS LTD. Obtenido de https://link.springer.com/chapter/10.1007/978-1-349-26121-5_15 (https://link.springer.com/chapter/10.1007/978-1-349-26121-5_15)

Simon Fraser University. (30 de Septiembre de 2011). THE CLASSICAL MODEL. Obtenido de http://www.sfu.ca/~dsignori/buec333/lecture%2010.pdf (http://www.sfu.ca/~dsignori/buec333/lecture%2010.pdf)

Smith, A. (1977). An Inquiry into the Nature and Causes of The Wealth of Nations. (E. Cannan, Ed.) Chicago: University Of Chicago Press.

Statistics How To. (13 de Julio de 2016). Cook's D: Definition, Interpretation. Obtenido de Cook's Distance: https://www.statisticshowto.com/cooks-distance/ (https://www.statisticshowto.com/cooks-distance/)

Sweezy, P., Dobb, M., Takahashi, K., Hilton, R., Hill, C., Lefebvre, G., . . . Merrington, J. (1978). The Transition from Feudalism to Capitalism. London: Verso. Obtenido de https://www.versobooks.com/books/2179-the-transition-from-feudalism-to-capitalism (https://www.versobooks.com/books/2179-the-transition-from-feudalism-to-capitalism)

Tapia Granados, J. A. (Diciembre de 2012). Does Investment Call the Tune? Empirical Evidence and Endogenous Theories of the Business Cycle. Research in Political Economy, 1-23. Obtenido de https://deepblue.lib.umich.edu/handle/2027.42/94442 (https://deepblue.lib.umich.edu/handle/2027.42/94442)

Valle Baeza, A. (1978). Valor y Precios de Producción. Investigación Económica, 169-203.

Wells, J. (2007). The rate of profit as a random variable. The Open University, School of Management. Milton Keynes: Munich Personal RePEc Archive. Obtenido de https://mpra.ub.uni-muenchen.de/98235/ (https://mpra.ub.uni-muenchen.de/98235/)

Zachariah, D. (Junio de 2006). Labour value and equalisation of profit rates: a multi-country study. Indian Development Review, 4, 1-20.