---------------------------------------------------------------------------------------------------------------------

* ¿Qué es ***LinReg***?

**LinReg** es un asistente informático para el desarrollo de análisis de regresión lineal, el cual básicamente constituye una interfaz de usuario amigable para la utilización de determinadas librerías y paquetes estadísticos de Python. Mediante este programa es posible obtener y evaluar modelos de regresión lineal múltiple de una manera muy sencilla, sin la limitación de poseer conocimientos y habilidades de programación en Python, o en otros programas a fines como R o Matlab. ***LingReg*** está soportado en diferentes paquetes de procesamiento estadístico y de Ciencia de Datos, los que ponen a disposición del usuario un poderoso arsenal de pruebas estadísticas para garantizar el desarrollo exitoso y competente de análisis de regresión.

Tiene implementado múltiples pruebas para la evaluación de los modelos de regresión como:

-Pruebas de normalidad, independencia y homocedasticidad de los residuales

-Pruebas de linealidad y de especificación

-Análisis de adecuación de modelos de regresión y detección de *outliers*

-Pruebas de calidad de ajuste

-Pruebas de validación cruzada por K-folds y bootstrapping (*machine learning*)

-Análisis de extrapolación oculta para el dominio de definición (implementación informática novedosa)

-Comparación de diferentes alternativas de modelos de regresión (implementación informática novedosa)

Como usted comprenderá, por lo abarcador que resulta este programa desde un punto de vista estadístico, la documentación adjunta no puede explicar detalladamente cada análisis implementado La documentación está concebida para un usuario con previos conocimientos sobre regresión lineal y solo pretende guiarlo para su interacción con el programa. Muchas de las temáticas son complejas y muy especializadas, por lo cual sería imposible realizar un compendio mejor y más competitivo que el realizado por prestigiosos autores, tales como el profesor Douglas Montgomery en Montgomery et al. (2002). Además, una extensión innecesaria supondría alejar al documento del verdadero valor de un resumen conciso, cuya función es la de organizar el contenido temáticamente y servir de punto de partida y fin para una revisión bibliográfica más profunda que usted como usuario crítico y ávido de conocimientos debería realizar. No obstante, en esta se describe el objetivo y principio de cada prueba o análisis estadístico y se propone la literatura clave para ampliar sus conocimientos estadísticos sobre el tema.

Esta es la primera versión del *software*. Los desarrolladores desean que ante la duda, señalamiento o recomendación usted no dudase comunicarse a través de las direcciones de correos anexas. Sus criterios serán tenidos en cuenta para el perfeccionamiento del *software* en próximas versiones. En el nombre de los desarrolladores:

MSc. Jonathan Serrano Febles MSc. Andrés Valido Fajardo

*jonathan.serrano@umcc.cu* [*andres.valido@umcc.cu*](mailto:andres.valido@umcc.cu)

---------------------------------------------------------------------------------------------------------------------

* ¿Cómo usar ***LinReg***?

**LinReg** tiene el contenido compartimentado en un sistema de pestañas, que se vinculan mediante el botón ¨Siguiente¨, este botón a su vez las activa y hace que se pueda luego acceder a ellas directamente. Para el manejo simultáneo con otros programas usted puede presionar el botón *Windows* o conjuntamente el botón *Tab* y el botón *Alt*.

Debe disponer de la colección de datos en una hoja de Excel debidamente colocada en forma de columnas para el desarrollo de los análisis de regresión. Metodológicamente el análisis de regresión en este programa se divide en las siguientes etapas:

1. Importación de datos
2. Selección de datos y Análisis estadístico básico de las variables involucradas (Pestaña: *Datos filtrados*)
3. Análisis exploratorio de datos y detección de correlaciones o correspondencias entre las variables (Pestaña: *Correlación*)
4. Selección de la variable dependientes e independientes y conformación de los modelos matemáticos (Pestaña: *Conformación de los modelos*)
5. Análisis de los resultados generales de la regresión (Pestaña: *Resultados de la regresión*)
6. Comprobación de los supuestos de la regresión lineal, pruebas de especificación de los modelos lineales, multicolinealidad y detección de outliers (Pestaña: *Pruebas de adecuación*).
7. Análisis de la calidad de ajuste de los modelos matemáticos (Pestaña: *Pruebas de Calidad de Ajuste*)
8. Establecimiento del intervalo de aplicación a partir de análisis de extrapolación oculta (Pestaña: *Análisis de extrapolación oculta*)
9. Validación de los modelos matemáticos por técnicas de *machine learning* (Pestaña: *Pruebas de validación*)
10. Comparar diferentes alternativas de modelos matemáticos (Pestaña: *Comparación de modelos*)

---------------------------------------------------------------------------------------------------------------------

* Importación de datos

La información debe estar recogida en un archivo tipo Excel en forma de columnas. El programa importa el nombre de cada variable (nombre de la columna) y siempre la MISMA cantidad de datos por variables, de forma tal que la cantidad de observaciones vendrá dada por el menor valor a partir de un filtrado interno (Ejemplo: Si la columna A tiene 56 filas y la columna B 106, la tabla importada tendrá dos columnas y 56 filas). Para importar los datos se presiona en el ícono con forma de *Carpeta* que se encuentra en la parte superior izquierda. Al presionar este ícono sale una ventana para la búsqueda y selección del archivo en el equipo.

La información importada se muestra en la pestaña *Datos filtrados*.

---------------------------------------------------------------------------------------------------------------------

* Selección de datos y Análisis estadístico básico de las variables involucradas (Pestaña: *Datos filtrados*)

A partir de los datos importados del Excel y mostrados en forma de tabla en la pestaña *Datos filtrados* es necesario seleccionar las variables que se desean correlacionar. Para ello, en un recuadro se muestra un inventario de todas las variables importadas. Al seleccionar una variable, se obtiene un resumen estadístico básico de la misma como: cantidad de observaciones, máximo, mínimo, media, mediana, coeficiente de variación (CV), entre otras.

---------------------------------------------------------------------------------------------------------------------

* Análisis exploratorio de datos y detección de correlaciones o correspondencias entre las variables (Pestaña: *Correlación*)

A partir de las variables seleccionadas para la correlación en la pestaña anterior (Pestaña: *Datos filtrados*) y mediante el botón ¨*Siguiente paso*¨ se obtienen los diferentes gráficos de dispersión. Estos vienen dados por una dupla de las variables seleccionadas. Complementariamente en forma de tabla a la derecha se muestran los valores de los coeficientes de correlación.

Esta etapa es de gran importancia, ya que permite la detección de patrones gráficos o correspondencias para formular o corroborar estructuras matemáticas sobre las cuales se efectuará la regresión. En caso de necesitarse la linealización de un variable o la introducción de una nueva, debe realizarse ello en el Excel base, guardarse los cambios en este y repetir el procedimiento realizado hasta este punto.

---------------------------------------------------------------------------------------------------------------------

* Selección de la variable dependientes e independientes y conformación de los modelos matemáticos (Pestaña: *Conformación de los modelos*)

Esta pestaña es fundamental ya que aquí es en donde se crean las estructuras y se ajustan los datos a los modelos de regresión.

La introducción de cada modelo (especificación de su estructura) debe realizarse de forma individual y secuenciada. Para ello:

1. Debe asignarse un nombre único al modelo en el recuadro ¨Nombre del modelo¨
2. Especificar el nivel de significación para la estimación de los intervalos de confianza de coeficientes de regresión.
3. Seleccionar la variable dependiente en el recuadro superior izquierdo del listado de variables.
4. Seleccionar las variables independientes en el recuadro inferior izquierdo del listado de variables.
5. Presionar en el botón ¨Registrar modelo¨.

Una vez registrado el modelo, este puede ser editado (*Editar*), eliminado (*Eliminar*) o examinado a partir de la información especificada en *Detalles*. Notará al respecto que solamente puede para un modelo especificar una variable dependiente y que esta no puede ser al mismo tiempo un término independiente. Al presionar el botón ¨*Siguiente paso*¨ internamente se ajustan los datos a las estructuras matemáticas establecidas, cuyos resultados se muestran en la pestaña: *Resultados de la regresión*

*---------------------------------------------------------------------------------------------------------------------*

* Análisis de los resultados generales de la regresión (Pestaña: *Resultados de la regresión*)

En esta pestaña se muestran los principales resultados de la regresión y se encuentra subdividida en ¨General¨, ¨Coeficientes de Regresión¨ e ¨Intervalos de estimación de los coeficientes de regresión¨. En caso de disponerse de más de un modelo, deberá seleccionarse el modelo del cual se desee ver los resultados. Este detalle es importantísimo para el trabajo en las futuras pestañas, ya que el modelo seleccionado aquí condiciona los resultados que se mostrarán en las mismas.

El nombre del modelo seleccionado aparece con una franja oscura en la parte superior del recuadro principal, mientras que el nombre de las otras alternativas (en caso de proceder) se posiciona debajo del recuadro principal. La selección de uno nuevo para ver sus resultados hace que se desmarque el seleccionado previamente y que se actualice la información mostrada en la pestaña actual.

El ajuste de los datos por regresión lineal y la información mostrada en esta sección se fundamenta en el empleo de la librería *statsmodels (Seabold y Perktold, 2010)*, la cual es una de las más prestigiosas y reconocidas en Python para la cuestión.

-Información mostrada en la sub-pestaña ¨General¨:

* Variable dependiente
* Número de observaciones
* Grados de libertad (GL) de los residuales
* Grados de libertad (GL) del modelo
* Criterio de información de Akaike (AIC)
* Criterio de información bayesiano (BIC)
* Log-Likelihood
* Coeficiente de determinación (R-cua)
* Coeficiente de determinación ajustado (R-cua ajustada)
* Valor del estadígrafo de Fisher de la prueba de significación de la regresión
* P-valor de la prueba de significación de la regresión
* Error cuadrático medio del modelo (MSE-modelo)
* Error cuadrático medio de los residuales (MSE-residuales)
* Error cuadrático medio total (MSE-total)
* Raíz del error cuadrático medio del modelo (RMSE-modelo)

-Información mostrada en la sub-pestaña ¨ Coeficientes de Regresión ¨:

* Coeficientes de regresión
* ANOVA de los coeficientes de regresión (que involucra al valor asociado de la distribución de Student, el P-valor correspondiente y al error estándar de la estimación del coeficiente)

-Información mostrada en la sub-pestaña ¨Intervalos de estimación de los coeficientes de regresión¨:

* Límite superior e inferior de estimación de cada coeficiente de regresión

---------------------------------------------------------------------------------------------------------------------

* Comprobación de los supuestos de la regresión lineal, pruebas de especificación de los modelos lineales, multicolinealidad y detección de outliers (Pestaña: *Pruebas de adecuación*)

Esta pestaña es de gran importancia ya que forma parte del proceso de validación matemática de los modelos matemáticos. A través de sus diferentes análisis se evalúa si desde un punto de vista estadístico el modelo es adecuado, aunque sin hacer inferencias propiamente a pruebas formales de calidad de ajuste (las que se abordan debidamente en otra pestaña).

De gran importancia reviste la comprobación de los supuestos de la regresión lineal (normalidad, independencia y homocedasticidad de los residuales) ya que tal como plantea Montgomery et al. (2002) los modelos de regresión serán solamente válidos si estos son satisfechos. A su vez se disponen de pruebas de linealidad y especificación, a través de las cuales puede corroborarse la adecuada formulación de la estructura matemática y de diferentes análisis para la detección de outliers influyentes. Estos últimos no deberían tenerse en consideración ya que pueden conducir a explicaciones incorrectas del modelo y conllevar a desviaciones en su descripción, así como falsear la verdadera calidad de ajuste del mismo a la colección de datos de proceso. Se determina el valor de inflación de la varianza para detectar posibles problemas de multicolinealidad y por tanto, evitar el enmascaramiento de efectos, interferencias en la respuesta y mala especificación del modelo; por ello es necesario que un término independiente no pueda expresarse a partir de uno o más términos independientes del modelo (debe ser legítimamente independiente).

La información que se muestra es la asociada al modelo que se selecciona en la pestaña anterior (Pestaña: *Resultados de la regresión*), de forma tal que si se necesita ver la información de otro modelo, debe antes en esa pestaña seleccionarse y luego presionar el botón siguiente. Los resultados están divididos en las sub-pestañas siguientes:

* Análisis de normalidad (residuales no escalados)
* Análisis de normalidad (residuales estudentizados)
* Homocedasticidad e independencia de los residuales
* VIF, linealidad y especificación
* DFBETAS
* Otros gráficos

Los análisis de normalidad con residuales no escalados parten de los residuales directos, sin ningún tipo de transformación mientras que la de los residuales escalados corresponde a los residuales estudentizados (externamente); variante potencialmente útil para la detección de *outliers* de acuerdo a Montgomery et al. (2002). Las pruebas que se implementan al respecto se realizan para una confianza del 95% a partir dela librería *SciPy* (Virtanen, 2020) y son las siguientes:

* Prueba de Lilliefors
* Prueba de Shapiro-Wilk
* Prueba de Kolmogorov-Smirnov
* Prueba de Jarque-Bera
* Prueba de K2 de D´Angostino
* Prueba de Chi-Cuadrado
* Prueba de Anderson-Darling

Para cada prueba se determina el valor del estadígrafo correspondiente (que se presenta con siglas relacionadas con el nombre de la misma) y el P-valor. En estas, si el P-valor es mayor que el nivel de significación resulta que NO existen evidencias que supongan el rechazo de la idea que los residuales siguen una distribución normal. Alternativamente, se determinan los valores de los coeficientes de Curtosis y de asimetría, y se muestran los gráficos Q-Q de los residuales y de distribución de densidad.

Los valores de los coeficientes de Curtosis y asimetría deben encontrarse entre -2 y 2 para que los residuales presenten una distribución normal. La distribución de la densidad de los residuales debe asemejarse a la de distribución normal de referencia; y del mismo modo, los residuales deben de posicionarse sobre la línea de referencia en el gráfico Q-Q.

Los análisis de homocedasticidad (varianza constante de los residuales e independiente del valor ajustado) se realizan a partir de la prueba de Breush-Pagan, la prueba de Goldfeld-Quandt y la prueba de White. En las mismas se determina el valor del estadígrafo asociado (representado con las siglas de la prueba) y el P-valor para una confianza del 95% a través del paquete *statsmodels (Seabold y Perktold, 2010)*. No es probable que los residuales sean homocedásticos si el P-valor de cada prueba es menor que el nivel de significación.

Complementariamente se muestran los gráficos de *Residuales vs Valor ajustado,* en donde debe apreciarse una total dispersión y sin patrones definidos de los residuales, de forma tal que la línea de tendencia resultante de la nube de densidad sea lo más horizontal posible y coincidente a y=0.

La independencia de los residuos, o sea, que estos no se encuentren correlacionados entre sí se realiza mediante la prueba de Durbin-Watson y la de Breush-Godfrey, ambas realizadas mediante el paquete *statsmodels (Seabold y Perktold, 2010).* La prueba de Breush-Godfrey se realiza para una confianza del 95% en la cual el P-valor debe ser mayor que el nivel de significación para que los residuales sean independientes entre sí mientras que para la de Durbin-Watson se determina el valor de su estadígrafo. Se le recomienda seguir los criterios referidos por Montgomery et al. (2002) el cual establece que:

* d < dL: Se infiere que los errores tienen autocorrelación positiva
* d>dU: Se infiere que los errores no presentan autocorrelación entre ellos
* dL<d<dU: La prueba no es concluyente

Los valores críticos o cotas dependen del tamaño de la muestra, del número de parámetros del modelo y de la confianza requerida; los cuales puede encontrarlo según el caso en Kanji (2006).

Alternativamente para el análisis de independencia de los residuales, puede emplear la gráfica anexa de *Valores observados vs Valores predichos*, los cuales deben posicionarse adecuadamente sobre la diagonal de 45o sin la existencia de patrones anormales definidos que indiquen una influencia externa.

En la sub-pestaña ¨VIF, linealidad y especificación¨ se condensan análisis de gran importancia para la evaluación de la adecuación de los modelos. El valor de inflación de la varianza permite determinar si existe multicolinealidad si este es inferior a 10 aunque algunos autores consideran que deben ser inferior a 5 (Montgomery y Runger, 2018).

En esta sección se pueden realizar la prueba de especificación de Ramsey y la de White, la prueba de linealidad de Harvey-Collier, la prueba de los multiplicadores de Lagrange para linealidad y la prueba de Rainbow de linealidad. En general estas pruebas permiten explorar si hay errores de especificación, omisión de variables, poca dependencia lineal entre la variable dependiente y los predictores y realizar inferencias preliminares acerca del ajuste de los modelos matemáticos. Las mismas se implementan mediante la librería *statsmodels* (Seabold y Perktold, 2010) para una confianza del 95%. El P-valor de cada prueba debe ser mayor que al nivel de significación para inferir que no existen problemas de especificación y de poca dependencia lineal. Alternativamente, la media de los residuales debe presentar valores cercanos a cero con un bajo coeficiente de variación.

En la sub-pestaña DFBETAS se muestran los gráficos DFBETAS para cada coeficiente de variación, los cuales son potencialmente útiles en la detección de outliers influyentes. Básicamente este indica cuánto cambia (en unidades de desviación estándar) si dicha observación fuera eliminada. Generalmente el valor máximo valor aceptado es 1 para muestras pequeñas y medianas, y para grandes muestras [2/n^0.5] (Sullivan et al., 2021). Si fuera superior a este valor debería examinarse con más detenimiento esta observación.

En la sub-pestaña “Otros gráficos” se muestran correspondencias gráficas de los residuales y otros análisis para la detección de outliers influyentes. Existen diversos criterios para cada uno de estos análisis a la hora de establecer el valor crítico a partir del cual se consideraría una observación atípica (outlier) influyente, le recomendamos al respecto consultar como primera aproximación las siguientes fuentes: Sullivan et al. (2021), Rahman et al. (2022), y Arimie et al. (2020). Los outliers influyentes no deberían tenerse en consideración ya que pueden conducir a explicaciones incorrectas del modelo, conllevar a desviaciones en su descripción, así como falsear la verdadera calidad de ajuste del mismo a la colección de datos de proceso; generalmente estos son el resultado de una anormalidad del proceso o de errores de medición.

Para el caso de los residuales estudiantados generalmente se recomienda que estos deben ser inferiores a 3. La estadística DFFITS mide el número de desviaciones estándar asociadas a la eliminación de la observación en cuestión. Si es mayor que ¨2\*(p/n)^0.5¨ (siendo ¨p¨ el número de parámetros del modelo y ¨n¨ la cantidad de observaciones) esa observación se considera influyente.

A partir de los elementos diagonales de la matriz sombrero (matriz conocida en la literatura como matriz H) es posible establecer aproximaciones de la influencia o más bien, ¨cantidad de influencia¨ (Sullivan et al., 2021). De modo general, se considera influyente si es mayor que ¨2\*p/n¨ (siendo ¨p¨ el número de parámetros del modelo y ¨n¨ la cantidad de observaciones).

La distancia de Cook es una medida de influencia desarrollada por Cook sustentada en la matriz sombrero de la regresión. Existen diversos criterios (como usted comprenderá luego de la lectura recomendada al respecto), de los cuales el más generalizado es considerar que una observación es influyente si tiene un valor mayor que 1 (Montgomery y Runger, 2018; Serrano et al., 2022).

La estadística de DFBETAS y DFFITS ofrecen inferencias acerca del efecto de las observaciones sobre los estimados y valores ajustados, sin embargo obvian la precisión de la estimación. De ahí entonces la importancia de COVARATIO, que esencialmente determina el papel de la observación en dicha precisión. Si su valor es mayor que 1 entonces dicho punto mejora la precisión mientras que si es menor que 1, su inclusión disminuye la precisión. Por lo general, si es mayor o menor que ¨1+3\*p/n¨ (siendo ¨p¨ el número de parámetros del modelo y ¨n¨ la cantidad de observaciones) esa observación se considera influyente.

* Análisis de la calidad de ajuste de los modelos matemáticos (Pestaña: *Pruebas de Calidad de Ajuste*)

En esta pestaña se realiza la prueba de bondad de ajuste de Fisher, el cual es el análisis de su tipo más concluyente, reconocida en algunas fuentes como Montgomery et al. (2002) como ¨ una *prueba formal* de calidad de ajuste¨. La implementación computacional de esta técnica es un aporte de ***LinReg.*** Básicamente esta prueba analiza si existen o no diferencias significativas entre los errores de predicción introducidos por el modelo y los errores inherentes a la variabilidad natural del proceso (error puro). Por tanto, es un requisito indispensable contar con observaciones replicadas para la determinación de este último. En esta si el estadígrafo de Fisher calculado es menor que al tabulado se establece que no hay evidencias desde un punto de vista estadístico que denoten pérdida de calidad de ajuste por los errores de la predicción. La relación entre el estadígrafo calculado y tabulado da una medida potencial de este ajuste. Para esta prueba debe de aclararse el nivel de significación. Se ofrecen como parámetros intermedios de salida a la suma de cuadrados por falta de ajuste (SSfa) y a la suma de cuadrados del error puro (SSpe).

En igual sentido se implementa el criterio de Montgomery et al. (2002), el que recomienda de que el rango de los valores ajustados sea grande en relación con su error estándar promedio estimado. Este autor también resalta como criterio complementario al RMSE, el que debe ser lo más cercano a cero posible.

Complementariamente para el análisis de calidad de ajuste debe existir una adecuada disposición sobre la diagonal de 45o sin patrones anómalos de los valores observados y predichos en el gráfico correspondiente que se emplea en la comprobación de la independencia de los residuales.

* Establecimiento del intervalo de aplicación a partir de análisis de extrapolación oculta (Pestaña: *Análisis de extrapolación oculta*)

La implantación informática de esta técnica de gran complejidad matemática es un aporte de ***LinReg***, dado que no existen procesadores estadísticos ni librerías afines que permitan su realización, a pesar de ser de gran importancia para definir el intervalo de aplicación de los modelos (Montgomery et al., 2002; Montgomery y Runger, 2018; Serrano et al., 2022).

Los valores máximos y mínimos de cada variable independiente restringen el intervalo de aplicación de los modelos de regresión, pero solo en primera instancia. Ello no supone que dentro de tales limites se disponga de información suficiente para contemplar efectos entre diferentes niveles de los términos independientes y se puede incurrir por tal motivo a predicciones erróneas por ‘’extrapolación oculta’’ (Montgomery et al., 2002). Por tal razón, para la definición del intervalo de aplicación de los modelos se realiza un análisis que parte del mallado definido por la combinación de las variables independientes involucradas en cada modelo y que está acotado por los valores máximos y mínimos de cada una de ellas. Luego, se analiza la posible inclusión de los puntos del mallado dentro de la región envolvente o cáscara de variables regresoras (RVH) (Montgomery et al., 2002) si se satisface la siguiente condición :

x^(-1)\*(X´X)^(-1) <= hmxm

Donde:

X: Matriz de los predictores que se emplea en el desarrollo del modelo de regresión

x: Punto constituido para la nueva predicción con coordenadas desde 1 a k, donde k es el número de variables independientes del modelo según el orden de estos en la matriz X.

hmxm: Valor máximo de los elementos de la diagonal de la matriz sombrero.

En caso de incluirse todos los puntos o nodos dentro de la RVH, se valida el intervalo de aplicación propuesto para los modelos (que define la malla). De modo contrario, se restringe más el intervalo de aplicación propuesto y por tanto, se acota más la malla hasta satisfacer la condición anterior.

Estos intervalos se especifican en la parte visual con la cantidad de niveles de las variables correspondientes. Por cuestiones de costo computacional, la cantidad de niveles se encuentra restringida a un máximo de 5.

´

* Validación de los modelos matemáticos por técnicas de *machine learning* (Pestaña: *Pruebas de validación*)

La validación de los modelos matemáticos se puede realizar tanto por k-cross-folds (validación cruzada por folds o pliegues) o por bootstrapping. Ello se aclara en la parte visual, mediante los botones correspondientes.

La validación cruzada por k-folds se realiza con el fin de reducir los sesgos de muestreo aleatorio y el problema de sobreajuste en el modelo matemático (Shah et al., 2021). Este es un método estadístico para la evaluación del comportamiento predictivo de modelos matemáticos más exhaustivo y fiable que la simple división de los datos en dos conjuntos: uno de entrenamiento y otro de validación (método tradicional).

Generalmente se recomienda un número de pliegues (folds) igual a 10. Para este caso por ejemplo, ello que supone la división del conjunto de datos en 10 subgrupos de igual tamaño. A partir de ello y de forma secuenciada, nueve se emplean para entrenamiento y uno para la validación del modelo, cuyo algoritmo concluye cuando todos los subgrupos hayan sido empleados para la validación del modelo (Jung et al., 2020). Se emplea para su realización la librería Sklearn (Hao et al., 2019).

Se implementan dos métricas, el RMSE y el coeficiente de determinación. Por cuestiones obvias en cuanto a indicación de calidad de ajuste, le recomendamos que considere solamente la del RMSE. Para su análisis se muestran los resultados para el conjunto de validación, de entrenamiento y la prueba en general. Es deseable que no existan diferencias entre ellos ni mucha variabilidad (bajo coeficiente de variación CV), ya que ello supone estabilidad y robustez del modelo ante cambios en el set de validación. Los valores de RMSE deben ser bajos propiamente por cuestiones de calidad de ajuste. Alternativamente puede emplear la gráfica anexa de valores predichos por el modelo y los predichos por k-folds, los cuales deben ser lo más semejantes posibles y disponerse sin anomalías sobre la diagonal de 45o.

El proceso de *bootstraping* consiste en generar de forma iterativa diferentes modelos lineales, empleando en cada caso una *bootstrap-sample* creada mediante *resampling* del mismo tamaño que la muestra inicial. Para cada modelo ajustado se registran los valores de los coeficientes y finalmente se estudia su distribución. En general es deseable para una mayor robustez que exista poca variabilidad en los valores de los coeficientes de regresión y que el valor del RMSE de la prueba sea cercano a cero con bajo coeficiente de variación. Estos análisis se complementan con otros anteriores, por ejemplo si no se comportase adecuadamente puede deberse entre otras razones a la existencia de outliers influyentes, a errores de especificación, de multiolinealidad o propiamente de calidad de ajuste. Por cuestiones de costo computacional, la cantidad de *boots* se encuentra limitada a 1000 (o sea, 1000 iteraciones).

* Comparar diferentes alternativas de modelos matemáticos (Pestaña: *Comparación de modelos*)

Esta es una prestación especial de **LinReg** que la diferencia de otros procesadores semejantes. Mediante la misma es posible la comparación de diferentes alternativas de estructuras matemáticas para la regresión o consecuentemente, de diferentes modelos de regresión a través de un gráfico radar o de araña. Se muestran como criterios de comparación al AIC, BIC, RMSE, el coeficiente de determinación y el coeficiente de determinación ajustado.

La selección del modelo más adecuado depende de la ponderación personal que haga el usuario de cada indicador, del nivel de compromiso que establezca entre calidad de descripción y complejidad estructural, entre otras cuestiones relacionadas con la cuestión. La superioridad del modelo solo está reflejada en el gráfico en base al indicador, de forma tal que un modelo presentará mejor resultado con respecto a otro modelo, mientras MAYOR sea su radio para tal indicador.

Se le recomienda al usuario NO emplear como criterio de comparación a los referidos al coeficiente de determinación proque este no es indicador de la calidad de ajuste, se encuentra fuertemente influenciado por la data de valores ajustados, depende del número de predictores y NO puede ser empleado cuando hay transformaciones en la variable dependiente (Montgomery et al., 2002). Para más información sobre el coeficiente de determinación y su empleo en la evaluación de modelos consulte los siguientes materiales, en los cuales específicamente se trata esta cuestión: Montgomery et al. (2002) y Rodríguez (2005).

El valor de RMSE mientras más cercano sea a cero es mejor, lo que permite inferir por este sentido un mejor ajuste (Gharekhani et al., 2021). El criterio de información de Akaike y el de información bayesiano permiten analizar la predicción en base a la complejidad del modelo matemático (principio de parsimonia), para lo cual penalizan el incremento del número de parámetros (Mullor et al., 2019). El criterio de información de Akaike es una medida relativa de la calidad de ajuste de un modelo de ajuste que se basa en la entropía de información, o sea, muestra una estimación relativa de la información perdida por el modelo. El criterio de información bayesiano está muy relacionado con el de Akaike y parte de la función de probabilidad. Importante aclarar que estos son solamente válidos para comparar modelos, de modo contrario carecen de sentido de por sí solos, sin la comparación con otro valor.

Es deseable que el modelo presente un bajo valor de AIC (Suresh y Krishna Priya, 2011) y de BIC (Ali et al., 2018). De modo contrario, el valor del coeficiente de determinación debe ser alto. El RMSE se emplea como criterio fundamental en la selección de alternativas de modelos en numerosas investigaciones (Ayetigbo et al., 2021; Güven y Şimşir, 2020; Isa et al., 2020; Yang et al., 2020) y por Montgomery et al. (2002).

**Referencias bibliográficas:**

Ali, M., Deo, R.C., Downs, N.J., & Maraseni, T. (2018). Cotton yield prediction with Markov Chain Monte Carlo-based simulation model integrated with genetic programing algorithm: A new hybrid copula-driven approach. *Agricultural and Forest Meteorology* 263, 428-448. <https://doi.org/https://doi.org/10.1016/j.agrformet.2018.09.002>.

Arimie, C.O., Biu, E.O., & Ijomah, M.A. (2020). Outlier Detection and Effects on Modeling. *Open Access Library Journal* 7, e6619. <https://doi.org/https://doi.org/10.4236/oalib.1106619>.

Ayetigbo, O., Latif, S., Abass, A., & Müller, J. (2021). Dataset on influence of drying variables on properties of cassava foam produced from white- and yellow-fleshed cassava varieties. *Data in Brief* 37, 107192. <https://doi.org/https://doi.org/10.1016/j.dib.2021.107192>.

Gharekhani, M., Nadiri, A.A., Khatibi, R., & Sadeghfam, S. (2021). An investigation into time-variant subsidence potentials using inclusive multiple modelling strategies. *Journal of Environmental Management* 294, 112949. <https://doi.org/https://doi.org/10.1016/j.jenvman.2021.112949>.

Güven, İ., & Şimşir, F. (2020). Demand forecasting with color parameter in retail apparel industry using artificial neural networks (ANN) and support vector machines (SVM) methods. *Computers & Industrial Engineering* 147, 106678. <https://doi.org/https://doi.org/10.1016/j.cie.2020.106678>.

Hao, J., Ho, T.K.J.J.o.E., & Statistics, B. (2019). Machine learning made easy: a review of scikit-learn package in python programming language. 44(3), 348-361.

Isa, M.H., Wong, L.-P., Bashir, M.J., Shafiq, N., Kutty, S.R.M., Farooqi, I.H., & Lee, H.C. (2020). Improved anaerobic digestion of palm oil mill effluent and biogas production by ultrasonication pretreatment. *Science of The Total Environment* 722, 137833.

Jung, K., Bae, D.-H., Um, M.-J., Kim, S., Jeon, S., & Park, D.J.S. (2020). Evaluation of nitrate load estimations using neural networks and canonical correlation analysis with k-fold cross-validation. 12(1), 400.

Kanji, G. (2006). *100 STATISTICAL TESTS* (3ra ed. Sage, Londres

Montgomery, D., Peck, E., & Vining, G. (2002). *Introducción al análisis de regresión lineal*. Compañía Editorial Continental, México.

Montgomery, D., & Runger, G. (2018). *Applied Statistics and Probability for Engineers* (7ma ed. Wiley, Estados Unidos.

Mullor, R., Mulero, J., & Trottini, M. (2019). A modelling approach to optimal imperfect maintenance of repairable equipment with multiple failure modes. *Computers & Industrial Engineering* 128, 24-31. <https://doi.org/https://doi.org/10.1016/j.cie.2018.12.032>.

Rahman, M.F., Khayat, M.E., Abd EL-Mongy, M., Yakasai, H.M., Yasid, N.A., Shukor, M.Y.J.J.o.E.M., & Toxicology (2022). Central Composite Design-based Optimization of Staphylococcus sp. strain Amr-15 Growth on Acrylamide as a Nitrogen Source. 10(2), 13-22.

Rodríguez, E.M.J.A.j.y.e.e. (2005). Errores frecuentes en la interpretación del coeficiente de determinación lineal. (38), 315-331.

Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. *9th Python in Science Conference*.

Serrano, J., Orozco, J.L., Dueñas, J., & Ramírez, H. (2022). Obtaining Mathematical Models to Predict the Behaviour of the Extraction Stage of the Raw Sugar Production Process. *Sugar Tech*. <https://doi.org/10.1007/s12355-022-01236-x>.

Shah, M.I., Javed, M.F., & Abunama, T. (2021). Proposed formulation of surface water quality and modelling using gene expression, machine learning, and regression techniques. *Environmental Science and Pollution Research* 28(11), 13202-13220. <https://doi.org/10.1007/s11356-020-11490-9>.

Sullivan, J.H., Warkentin, M., & Wallace, L. (2021). So many ways for assessing outliers: What really works and does it matter? *Journal of Business Research* 132, 530-543. <https://doi.org/https://doi.org/10.1016/j.jbusres.2021.03.066>.

Suresh, K.K., & Krishna Priya, S.R. (2011). Forecasting Sugarcane Yield of Tamilnadu Using ARIMA Models. *Sugar Tech* 13(1), 23-26. <https://doi.org/10.1007/s12355-011-0071-7>.

Virtanen, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature methods* 17, 261-272.

Yang, Y., He, M., & Li, L. (2020). Power consumption estimation for mask image projection stereolithography additive manufacturing using machine learning based approach. *Journal of Cleaner Production* 251, 119710. <https://doi.org/https://doi.org/10.1016/j.jclepro.2019.119710>.