Interpretación de los resultados de Regresión Lineal OLS

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               strength   R-squared (uncentered):                   0.933
Model:                            OLS   Adj. R-squared (uncentered):              0.932
Method:                 Least Squares   F-statistic:                              1376.
Date:                Mon, 13 Feb 2023   Prob (F-statistic):                        0.00
Time:                        20:20:28   Log-Likelihood:                         -2990.2
No. Observations:                 800   AIC:                                      5996.
Df Residuals:                     792   BIC:                                      6034.
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
cement         0.1183      0.005     24.663      0.000       0.109       0.128
slag           0.0969      0.006     16.548      0.000       0.085       0.108
ash            0.0823      0.009      9.170      0.000       0.065       0.100
water         -0.1719      0.018     -9.493      0.000      -0.207      -0.136
superplastic   0.2598      0.101      2.560      0.011       0.061       0.459
coarseagg      0.0073      0.003      2.126      0.034       0.001       0.014
fineagg        0.0109      0.004      2.751      0.006       0.003       0.019
age            0.1132      0.006     18.850      0.000       0.101       0.125
==============================================================================
Omnibus:                        4.731   Durbin-Watson:                   1.890
Prob(Omnibus):                  0.094   Jarque-Bera (JB):                4.607
Skew:                          -0.159   Prob(JB):                       0.0999
Kurtosis:                       3.193   Cond. No.                         367.
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
"""
```

Dep variable

“Strength”: Es la variable dependiente cuyo valor queremos calcular a partir de las demás

Model y Method

OLS “Ordinary Least Square.” Este modelo trata de encontrar una expresión en forma de combinación lineal de las columnas del Dataframe que minimice la suma de los cuadrados de los residuos. Los residuos son las distancias que van desde los puntos que representan cada observación hasta la recta obtenida en el modelo.

DF residuals y **DF model**

Tenemos una base de datos de entrenamiento que contiene 800 observaciones, y de cada observación hemos tomado el valor de 8 variables independientes y una dependiente. DF Model son esas 8 variables independientes. DF residual es la resta de 800 – 8 = 792

Covariance type

El tipo de covarianza normalmente es “no robusto” (nonrobust), lo que significa que no hay eliminación de datos para calcular la covarianza entre características. La covarianza muestra cómo dos variables se mueven una con respecto a la otra. Si este valor es mayor a 0, ambas se mueven en la misma dirección y si es menor a 0, las variables se mueven en dirección opuesta.
(No confundir la covarianza con la correlación. La covarianza no proporciona la fuerza de la relación, solo la dirección del movimiento, mientras que el valor de la correlación está normalizado y oscila entre -1 y +1 y la correlación da idea de la fuerza de la relación.)

R-squared

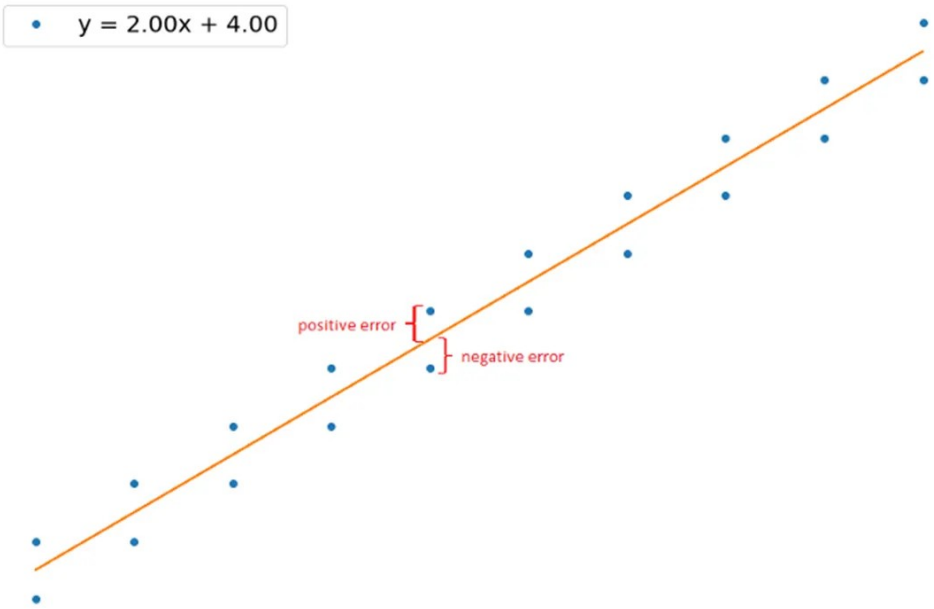
El valor R-cuadrado es el coeficiente de determinación que indica el porcentaje de la variabilidad de los datos explicados por las variables independientes seleccionadas.

Adj. R-squared

A medida que agregamos más y más variables independientes a nuestro modelo, los valores de R-cuadrado aumentan, pero en realidad, esas variables no siempre contribuyen a obtener mejor la variable dependiente. Por lo tanto, la adición de cada variable innecesaria necesita algún tipo de penalización. Por tanto, sólo tendremos en cuenta el ajuste de R-cuadrado cuando estemos realizando una regresión lineal múltiple. Para una sola variable independiente, tanto el valor de R-cuadrado como el de R-cuadrado ajustado son iguales.

coef y std err

La columna coef representa los coeficientes para cada variable independiente junto con el valor de intersección. Std err es la desviación estándar del coeficiente de la variable correspondiente en todos los puntos de datos. Cuando se usa solo una variable de predicción, el error estándar se puede obtener de este espacio bidimensional como se muestra a continuación.



Estadísticos t

El estadístico t es un valor estandarizado que se calcula a partir de los datos de la muestra durante una prueba de hipótesis.
Está basado en la distribución t de Student, que es una distribución normal con una variación en función de los grados de libertad.
El estadístico t es utilizado para determinar si hay una diferencia significativa entre dos o más muestras o para comparar una media poblacional con un valor hipotético.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               strength   R-squared (uncentered):                   0.933
Model:                            OLS   Adj. R-squared (uncentered):              0.932
Method:                 Least Squares   F-statistic:                              1376.
Date:                Mon, 13 Feb 2023   Prob (F-statistic):                        0.00
Time:                        20:20:28   Log-Likelihood:                         -2990.2
No. Observations:                 800   AIC:                                      5996.
Df Residuals:                     792   BIC:                                      6034.
Df Model:                           8
Covariance Type:            nonrobust
================================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
cement           0.1183      0.005     24.663      0.000       0.109       0.128
slag             0.0969      0.006     16.548      0.000       0.085       0.108
ash              0.0823      0.009      9.170      0.000       0.065       0.100
water           -0.1719      0.018     -9.493      0.000      -0.207      -0.136
superplastic     0.2598      0.101      2.560      0.011       0.061       0.459
coarseagg        0.0073      0.003      2.126      0.034       0.001       0.014
fineagg          0.0109      0.004      2.751      0.006       0.003       0.019
age              0.1132      0.006     18.850      0.000       0.101       0.125
==============================================================================
Omnibus:                        4.731   Durbin-Watson:                   1.890
Prob(Omnibus):                  0.094   Jarque-Bera (JB):                4.607
Skew:                          -0.159   Prob(JB):                       0.0999
Kurtosis:                       3.193   Cond. No.                         367.
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
"""
```

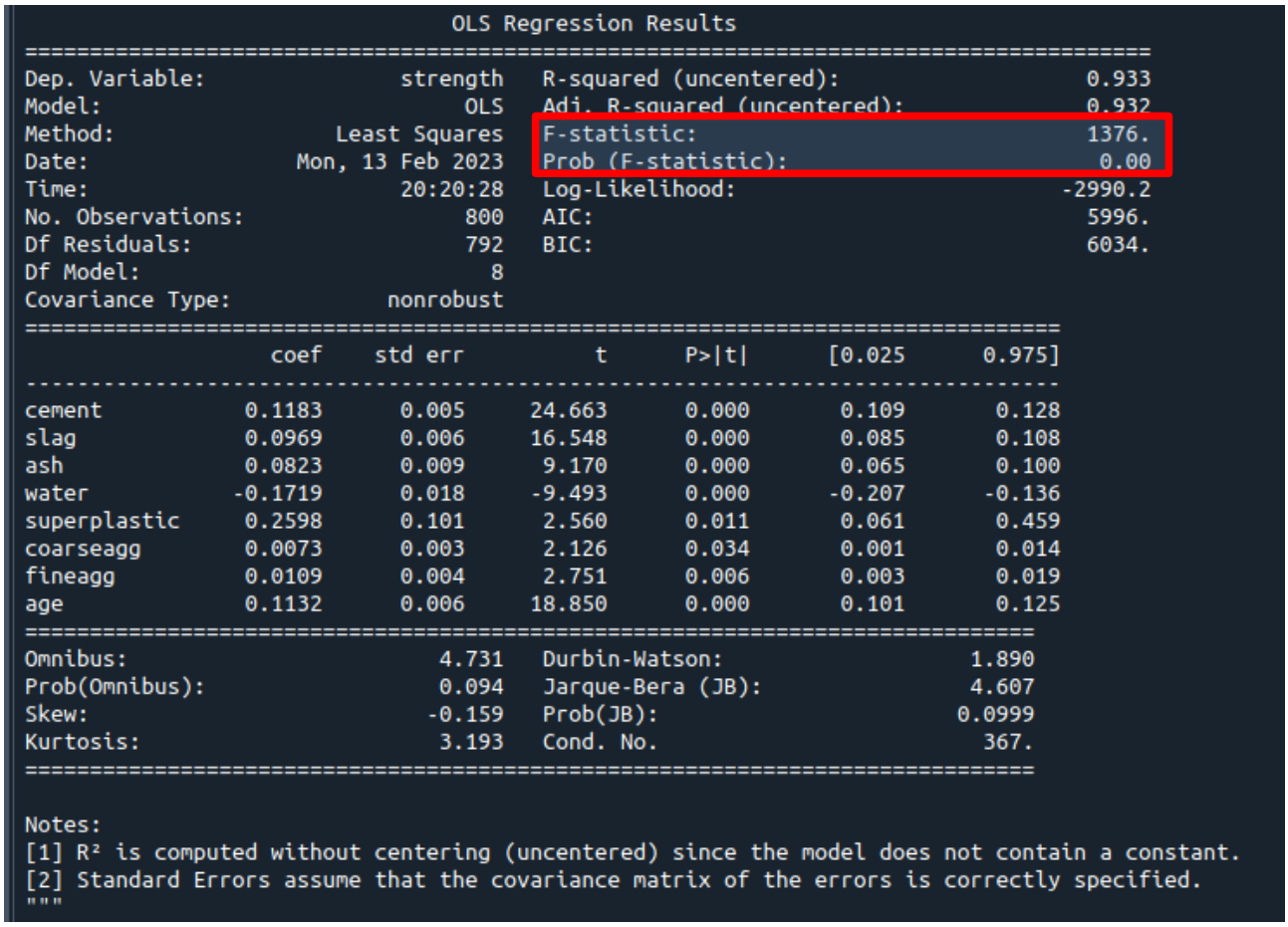
t-values y P>|t|

La columna t proporciona los valores t (t de Student) correspondientes a cada variable independiente. Por ejemplo, aquí c*ement, slag, ash, water*… tienen diferentes valores t, así como diferentes valores p asociados con cada variable. Las estadísticas T se utilizan para calcular los valores p.

Por lo general, **cuando el valor p es inferior a 0,05**, indica una fuerte evidencia en contra de la hipótesis nula que establece que la variable independiente correspondiente no tiene efecto sobre la variable dependiente. El valor P de 0,034 para *coarseagg* nos dice que hay un 3,4 % de posibilidades de que la variable *coarseagg* no tenga ningún efecto sobre la dureza del hormigón. Parece que *cement* obtuvo un valor p de 0, lo que indica que los datos de *cement* son estadísticamente significativos, ya que son inferiores al límite crítico (0,05). En este caso, podemos rechazar la hipótesis nula y decir que los datos de *cement* controlan significativamente la dureza, *strength*.

Estadísticos F

El estadístico F es una medida estadística utilizada para evaluar la capacidad explicativa que tienen un grupo de variables independientes sobre la variabilidad de una variable dependiente. Esta medida se calcula como un cociente entre la varianza de las medias del grupo y la media de las varianzas dentro del grupo.
En estadística, se utiliza para realizar pruebas F para evaluar la significación estadística de una hipótesis.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:               strength   R-squared (uncentered):                   0.933
Model:                            OLS   Adj. R-squared (uncentered):              0.932
Method:                 Least Squares   F-statistic:                              1376.
Date:                Mon, 13 Feb 2023   Prob (F-statistic):                        0.00
Time:                        20:20:28   Log-Likelihood:                         -2990.2
No. Observations:                 800   AIC:                                      5996.
Df Residuals:                     792   BIC:                                      6034.
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
cement         0.1183      0.005     24.663      0.000       0.109       0.128
slag           0.0969      0.006     16.548      0.000       0.085       0.108
ash            0.0823      0.009      9.170      0.000       0.065       0.100
water         -0.1719      0.018     -9.493      0.000      -0.207      -0.136
superplastic   0.2598      0.101      2.560      0.011       0.061       0.459
coarseagg      0.0073      0.003      2.126      0.034       0.001       0.014
fineagg        0.0109      0.004      2.751      0.006       0.003       0.019
age            0.1132      0.006     18.850      0.000       0.101       0.125
==============================================================================
Omnibus:                        4.731   Durbin-Watson:                   1.890
Prob(Omnibus):                  0.094   Jarque-Bera (JB):                4.607
Skew:                          -0.159   Prob(JB):                       0.0999
Kurtosis:                       3.193   Cond. No.                         367.
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
"""
```

La prueba F proporciona una forma de verificar al mismo tiempo todas las variables independientes para conocer si alguna de ellas está relacionada con la variable dependiente.

Si `Prob(F-statistic)` es mayor que 0,05, no hay evidencia de relación entre ninguna de las variables independientes con la de salida.

Si es menor a 0.05, podemos decir que hay al menos una variable que está significativamente relacionada con la salida.

En nuestro ejemplo, el valor es inferior a 0,05 y, por lo tanto, una o más de las variables independientes están relacionadas con la variable de salida *strength*. Por lo tanto, los datos de la prueba F respaldan los resultados de la prueba t. Sin embargo, puede haber algunos casos en los que la probabilidad (estadística F) sea superior a 0,05 pero una de las variables independientes muestre una fuerte correlación. Esto se debe a que cada prueba t se lleva a cabo con un conjunto diferente de datos, mientras que la prueba F verifica el efecto combinado que incluye todas las variables globalmente.

Log-likelihood

El valor de log-verosimilitud es una medida del ajuste del modelo con los datos dados. Es útil cuando comparamos dos o más modelos. **Cuanto mayor sea el valor** de log-verosimilitud, **mejor se ajustará el modelo a los datos proporcionados.** Puede variar desde infinito negativo hasta infinito positivo.
Si alguna de las variables que hemos utilizado para la regresión lineal no tuviese buenos datos en el valor p podríamos descartar esa columna y repetir el ensayo para ver si aumenta el valor de Log-Likelihood

AIC y BIC

AIC (abreviatura de Criterios de información de Akaike desarrollados por el estadístico japonés Hirotugo Akaike) y BIC (abreviatura de Criterios de información bayesianos) también se utilizan como criterios para la solidez del modelo. El objetivo es minimizar estos valores para obtener un mejor modelo.

Ómnibus y Prob (Ómnibus)

La prueba ómnibus verifica la normalidad de los residuos una vez que se implementa el modelo. Si el valor es cero, significa que los residuos son perfectamente normales. Aquí, en el ejemplo, prob(Omnibus) es 0,094, lo que indica que hay un 9,4 % de probabilidad de que los residuos se distribuyan normalmente. Para que un modelo sea robusto, además de verificar R-cuadrado y otras rúbricas, también se requiere que la distribución residual sea idealmente normal. En otras palabras, el residual no debe seguir ningún patrón cuando se grafica contra los valores ajustados.

Skew y Kurtosis
Los valores de sesgo nos dicen el sesgo de la distribución residual. Las variables normalmente distribuidas tienen 0 valores de sesgo. La curtosis es una medida de la distribución de cola ligera o cola pesada en comparación con la distribución normal. Una curtosis alta indica que la distribución es demasiado estrecha y una curtosis baja indica que la distribución es demasiado plana. Un valor de curtosis entre -2 y +2 es bueno para probar la normalidad. En nuestro caso el dato no es bueno

Durbin-Watson
La estadística de Durbin-Watson proporciona una medida de autocorrelación en el residual. Si los valores residuales están autocorrelacionados, el modelo se vuelve sesgado. Esto significa que un valor no debe depender de ninguno de los valores anteriores. Un valor ideal para esta prueba oscila entre 0 y 4. Hemos obtenido 1,89

Jarque-Bera (JB) y Prob (JB)
Jarque-Bera (JB) y Prob(JB) es similar a la prueba Omni midiendo la normalidad de los residuales.

Condition number
Un número de condición alto indica que existe una posible multicolinealidad presente en el conjunto de datos. (Una relación de dependencia lineal fuerte entre más de dos variables explicativas en una regresión múltiple) Si solo se usa una variable como predictor, este valor es bajo y puede ignorarse. Podemos proceder como una regresión por pasos y ver si se agrega alguna multicolinealidad cuando se incluyen variables adicionales.