ntic master School

UNIVERSIDAD COMPLUTENSE DE MADRID

Regresión lineal

Minería de Datos y Modelización Predictiva

Máster en Big Data, Data Science & Inteligencia Artificial
Universidad Complutense de Madrid

Los modelos de regresión tienes por objetivo **predecir** una variable y (que recibe el nombre de dependiente) a partir de un conjunto de m variables x_i *independientes* a través de una **ecuación**:

$$y = f(x_1, x_2, \ldots, x_m) + \epsilon,$$

donde ϵ representa el **error cometido** o, equivalentemente, la parte de la variable dependiente **no explicada** a partir de las variables independientes.

En el caso particular de la **regresión lineal**, la ecuación anterior se reduce a:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \epsilon,$$

donde y es una variable aleatoria **continua**, β_0 representa el valor que toma la variable dependiente cuando todas las variables independientes toman el valor 0 y los parámetros restantes representan **cuánto aumenta o disminuye** la variable dependiente por cada **incremento unitario** de las variables independientes.

A partir de un modelo, es posible predecir el valor de y para un determinado individuo, **conocidos** los valores que toman las variables **independientes**:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

ESTIMACIÓN DE LOS PARÁMETROS

Para poder llevar a cabo la predicción, es necesario conocer el valor de los **parámetros**. Como no es así, debemos **estimarlos** a partir del conjunto de datos.

Para ello, buscaremos para qué valor de los parámetros se **minimiza el error** cometido por el modelo, que viene dado por: $y - \hat{y}$.

El **estimador de mínimos cuadrados**, aquel que minimiza la suma de cuadrados de los errores, viene dado por:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y},$$

donde

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_m \end{pmatrix} \quad \boldsymbol{Y} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{21} & \dots & x_{m1} \\ 1 & x_{12} & x_{22} & \dots & x_{m2} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{1n} & x_{2n} & \dots & x_{mn} \end{pmatrix}$$

Nótese que hasta el momento no se ha utilizado ninguna hipótesis, por lo que los estimadores minimo cuadráticos son siempre válidos.

No obstante, aunque no se imponga ninguna hipótesis, es importante realizar algunos comentarios sobre la **predicción de los parámetros**:

- No se pueden incluir en el modelo dos variables independientes contínuas que estén **muy correlacionadas** pues, en ese caso, no es posible invertir la matriz $\boldsymbol{X'X}$.
- El número de parámetros incluidos en el modelo ha de ser **muy inferior** al número de observaciones, para evitar problemas en la estimación y de sobreajuste.
- Las variables independientes **categóricas** tienen un tratamiento "especial" a la hora de incluirse en el modelo.

TRATAMIENTO DE VARIABLES CUALITATIVAS

La inclusión de variables categóricas en el modelo no es tan directa como la de variables de intervalo, para las que es suficiente con añadir una columna en la matriz $\boldsymbol{X}$ que contenga los valores numéricos que toman las observaciones.

Como las categorías de las variables de clase generalmente vienen dadas por **valores alfanuméricos**, no es posible incluir dicha información en la matriz $\boldsymbol{X}$. Por ello, se construyen tantas **variables auxiliares *dummy*** como categorías tenga la variable independiente. Las variables dummy son variables dicotómicas que siempre estan asociadas a una única categoría de una variable categórica, las cuales valen $\mathbf{1}$ si la observación correspondiente toma ese valor y, $\mathbf{0}$, en otro caso.

Estas variables, al ser numéricas, pueden **incluirse** como columnas en la matriz $\boldsymbol{X}$.

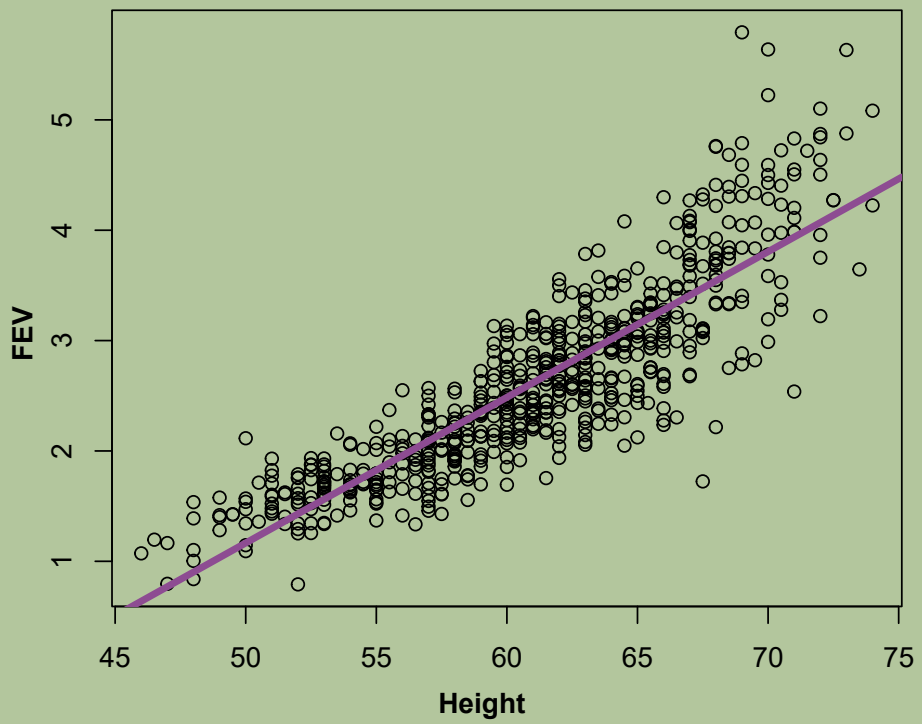
De esa forma, al incluir una variable independiente de clase, se incluyen al modelo **tantos parámetros como valores** tome dicha variable, los cuales representan lo que **cambia la media** de la variable (o lo que es lo mismo, el parámetro constante del modelo) según los valores de esta variable.

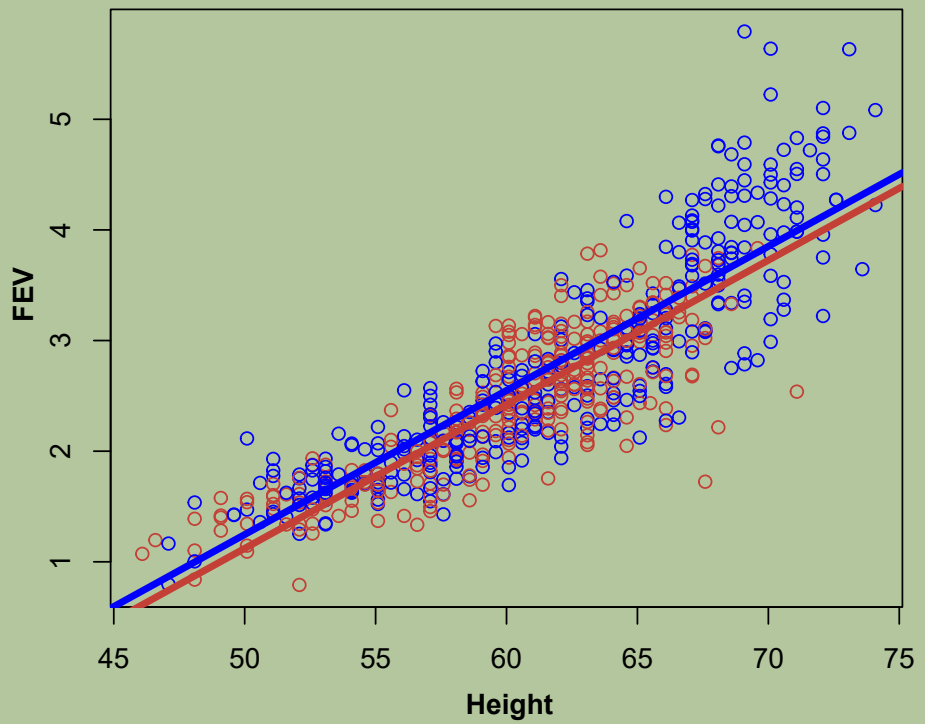
EJEMPLO

Se desea estimar el índice de funcionamiento pulmonar o FEV (Forced Expiratory Volume) en niños de edades entre 3 y 19 a partir de otras variables, como son la edad, la altura, el sexo y la presencia o no de un determinado gen.

Supongamos las siguientes regresiones:

$$FEV = \beta_0 + \beta_1 height + \epsilon \qquad FEV = \beta_0 + \beta_1 height + \beta_2 female + \beta_3 male + \epsilon$$





Es importante destacar que dadas todas **las variables dummys** menos una, es posible conocer perfectamente la restante (en términos matemáticos, las variables dummys son linealmente dependientes), lo que implica que, de esta forma, no se puede invertir la matriz $\boldsymbol{X'X}$ y, por tanto, **no se pueden estimar los parámetros**.

Lo que se hace en ese caso es **fijar** el parámetro de una de las categorías como $\mathbf{0}$ y **eliminar** de la matriz $\boldsymbol{X}$ la columna correspondiente.

De esta forma, el parámetro **constante** β_0 se corresponde con el valor de la variable objetivo que toma una observación de dicha categoría pero, que toma valor $\mathbf{0}$ para las demás variables independientes continuas.

Los parámetros restantes asociados a dicha variable miden, por tanto, la **diferencia** entre la categoría en cuestión y la **de referencia**.

INTERACCIONES ENTRE UNA VARIABLE CUALITATIVA Y UNA CONTINUA

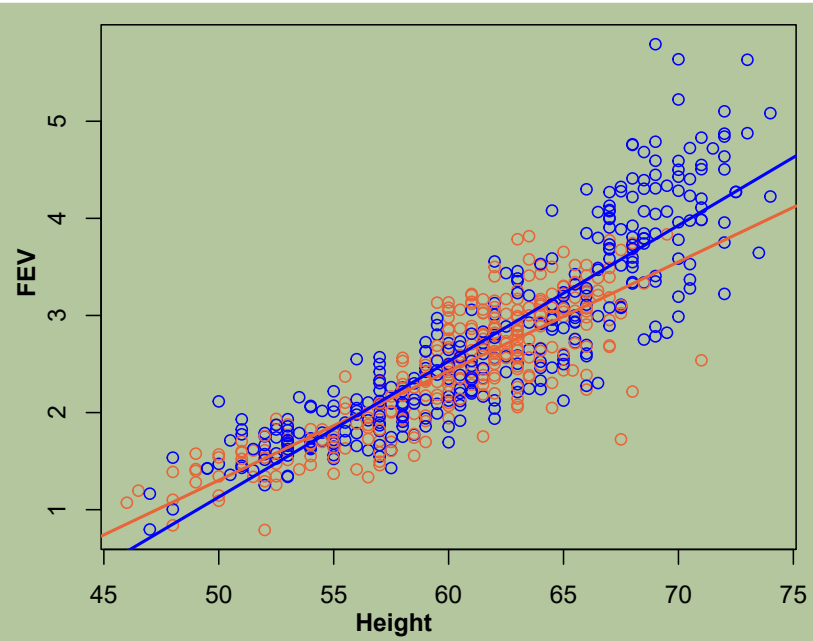
Como acabamos de ver, al **incluir** una variable categórica independiente lo que logramos es construir tantas **rectas de regresión paralelas** como categorías toma dicha variable. Sin embargo, hay ocasiones en las que el efecto de una determinada variable categorica se traduce en rectas de regresión con **distintas pendientes**.

La forma de modelizar este fenómeno es a partir de la inclusión de **interacciones** (producto) entre variables cuantitativas y cualitativas, donde, de nuevo, se recurre a variables dummys. Se presenta la misma problemática que con variables categóricas, por lo que el **número de parámetros** será uno menos que el número de categorías.

$$FEV = \beta_0 + \beta_1 height + \beta_2 female + \beta_3 height \star female + \epsilon$$

$$FEV_{male} = \beta_0 + \beta_1 height$$

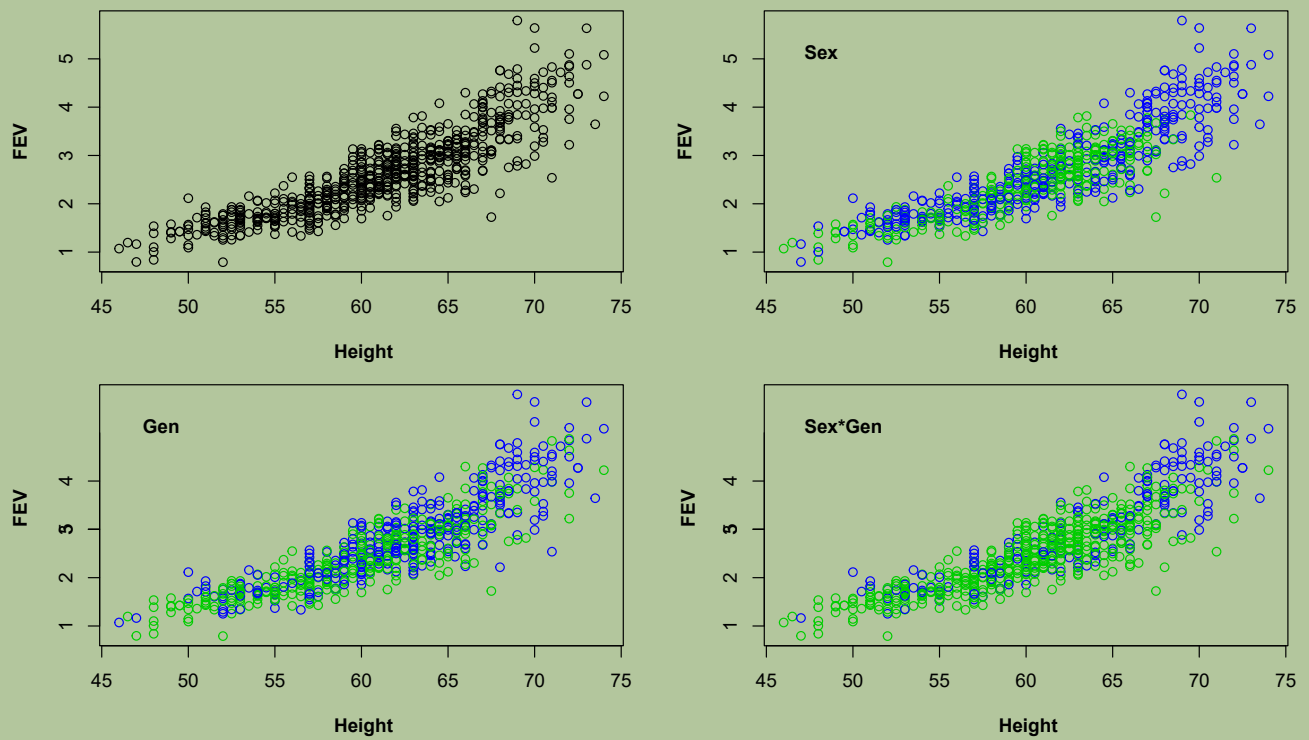
$$FEV_{fem} = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) height$$



INTERACCIONES ENTRE DOS VARIABLE CUALITATIVAS

En ocasiones, la **combinación** de algunas categorías de variables cualitativas origina un **efecto potenciador** (por el hecho de darse a la vez) mayor o menor que la suma de los efectos por separado.

En esos casos, se debe incluir el **producto** de ambas variables (en realidad, de las dummys) como un efecto más. El número de parámetros estimados es **menor** que el producto del número de categorías para evitar problemas de inversión de la matriz de diseño.

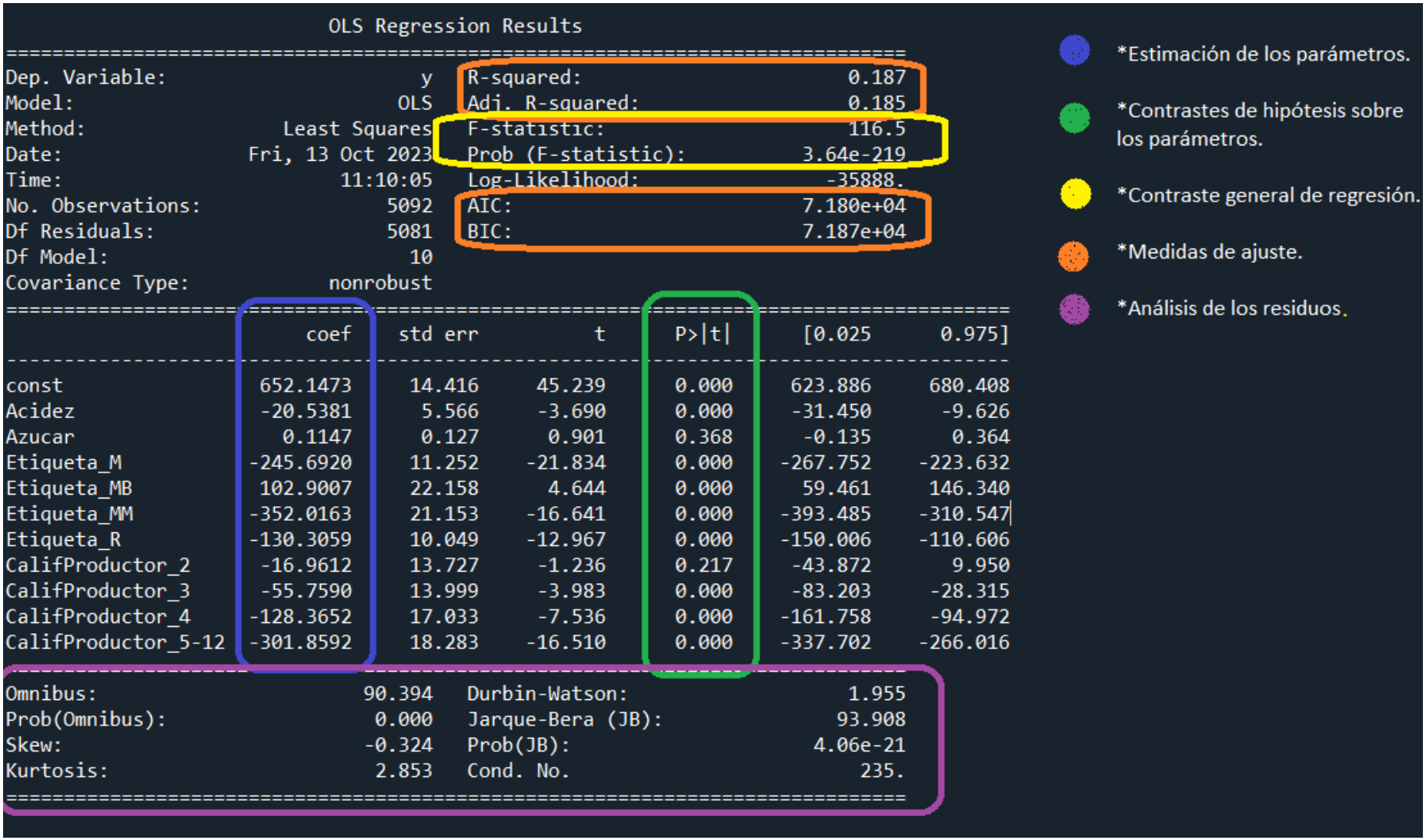


- Como ya se comentó en la fase de depuración, en general es recomendable evitar que haya categorías **poco representadas** (con pocas observaciones) para las variables de clase. En el caso de la regresión lineal este hecho da lugar a una **mala estimación** de dichos parámetros (se construyen en base a poca información).
- Al añadir **interacciones** al modelo se puede incurrir en problemas de estimación, pues puede no haber observaciones para **todas las combinaciones**, lo que hace imposible la estimación del parámetro.

Una forma de **evaluar el modelo** es a partir de la suma de cuadrados de los errores ***SSE*** (que mide el error cometido), de la suma de cuadrados explicada (o del modelo) ***SSM*** (que mide la información contenida en el modelo) y la suma de cuadrados total ***SST*** (que mide el error en el que se incurre si no hay modelo).

Hay que tener en cuenta que : $SST = SSE + SSM$.

Regresión lineal: resultados generales



OLS Regression Results

Dep. Variable:	y	R-squared:	0.187
Model:	OLS	Adj. R-squared:	0.185
Method:	Least Squares	F-statistic:	116.5
Date:	Fri, 13 Oct 2023	Prob (F-statistic):	3.64e-219
Time:	11:10:05	Log-Likelihood:	-35888.
No. Observations:	5092	AIC:	7.180e+04
Df Residuals:	5081	BIC:	7.187e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	652.1473	14.416	45.239	0.000	623.886	680.408
Acidez	-20.5381	5.566	-3.690	0.000	-31.450	-9.626
Azucar	0.1147	0.127	0.901	0.368	-0.135	0.364
Etiqueta_M	-245.6920	11.252	-21.834	0.000	-267.752	-223.632
Etiqueta_MB	102.9007	22.158	4.644	0.000	59.461	146.340
Etiqueta_MM	-352.0163	21.153	-16.641	0.000	-393.485	-310.547
Etiqueta_R	-130.3059	10.049	-12.967	0.000	-150.006	-110.606
CalifProductor_2	-16.9612	13.727	-1.236	0.217	-43.872	9.950
CalifProductor_3	-55.7590	13.999	-3.983	0.000	-83.203	-28.315
CalifProductor_4	-128.3652	17.033	-7.536	0.000	-161.758	-94.972
CalifProductor_5-12	-301.8592	18.283	-16.510	0.000	-337.702	-266.016

Omnibus:	90.394	Durbin-Watson:	1.955
Prob(Omnibus):	0.000	Jarque-Bera (JB):	93.908
Skew:	-0.324	Prob(JB):	4.06e-21
Kurtosis:	2.853	Cond. No.	235.

- *Estimación de los parámetros.
- *Contrastes de hipótesis sobre los parámetros.
- *Contraste general de regresión.
- *Medidas de ajuste.
- *Análisis de los residuos.

- El **contraste general de regresión** evalúa la bondad del modelo al completo, es decir, si la SSM es significativamente mayor que la SSE. Esto se realiza a través de un **test de la F**.
- Las **medidas de ajuste** que aparecen por defecto son R^2 y $R^2_{ajustado}$:
  - R^2: calcula la proporción de información explicada por el modelo

$$R^2 = \frac{SSM}{SST} = 1 - \frac{SSE}{SST}$$

  - $R^2_{ajustado}$: Es una modificación del anterior en el que se **penaliza** la presencia de muchos parámetros. Viene dado por:

$$R^2_{ajustado} = R^2 - (1 - R^2)\,\frac{p}{n - p - 1}$$

- Los **contrastes de hipótesis sobre los parámetros** evalúan si estos son **siginificativamente distintos de 0** pues, de no ser así, el efecto de la variable es despreciable y puede ser, por tanto, eliminada del modelo.

- El **análisis de los residuos** evalúa dos hipótesis de los modelos de regresión lineal. mostrandose dos estadísticos:
  - Estadístico de **Durbin-Watson** utilizado para determinar si los residuos del modelo son incorrelados. Dicho estadístico toma valores entre 0 y 4. Si el estadístico toma valores cercanos a 2 asumimos que los residuos son incorrelados. Para valores menores que 2, los errores muestran correlación positiva, y correlación negativa en caso de valores superiores a 2.
  - Estadístico de **Jarque-Bera (JB)** se utiliza para determinar si los residuos están normalmente distribuidos. El estadístico asociado al contraste sigue una distribución χ^2_2 bajo la hipótesis nula de normalidad. Se compara el valor del estadístico con el valor crítico de la distribucion para determinar si rechaza o no la hipótesis nula.

- El correcto funcionamiento de los contrastes anteriores depende del cumplimiento de ciertas hipótesis.
- Además, en mineria de datos, suele ocurrir que, si el tamaño muestral es muy grande, los contrastes tienen **demasiada potencia** y, por tanto, se rechazan incorrectamente las hipótesis.
- Por tando, dado que el objetivo principal de la minería de datos es la **predicción**, es habitual omitir la fase de evaluación de las hipótesis, así como el análisis de los p-valores de los contrastes, y sustituirla por **métodos de remuestreo**, como la división de los datos en submuestras entrenamiento y test, o la validación cruzada, donde se analiza el comportamiento del modelo bajo distintos escenarios.
  - Por ejemplo, se puede comprobar si eliminar una variable del modelo **empeora** los resultados significativamente y, de no ser así, se podría **eliminar** sin pérdida de información (sería el equivalente a los contrastes de hipótesis sobre los parámetros).
  - Estos métodos permiten evaluar la bondad del modelo a través de **índices** (como la SSE) que **no dependen** del cumplimiento de las hipótesis.

MEDIDAS DE AJUSTE Y COMPARACIÓN DE MODELOS

Al margen del estadístico R^2 y el $R^2_{ajustado}$ existen otras medidas que permiten **comparar modelos** y que también se basan en la suma de cuadrados de los **errores** (SSE).

Estas medidas, al contrario de lo que ocurre con R^2, **no están acotadas** y, por tanto, sólo sirven para comparar modelos. Además, cuanto **menor** sea el valor que tomen, **mejor** será el modelo que están evaluando.

Por último, cabe destacar que están formadas por **dos sumandos**, el primero de los cuales mide el **error del modelo** y el segundo, penaliza el número de **parámetros** (p) del mismo:

- **AIC** (Akaike information criterion): $n \ln(\frac{SSE}{n}) + 2p$
- **BIC** (Bayesian information criterion): $n \ln(\frac{SSE}{n}) + p \ln(n)$

Nótese que el **primer sumando coincide** para los dos criterios, por lo que la diferencia entre ellos se reduce a la penalización del número de parámetros: siendo el BIC el que más penaliza siempre y AIC generalmente el que menos.

Existen argumentos a favor y en contra de unos y otros criterios en la literatura, por lo que lo recomendable es **utilizarlos todos** a la hora de comparar modelos.

VALIDACIÓN CRUZADA

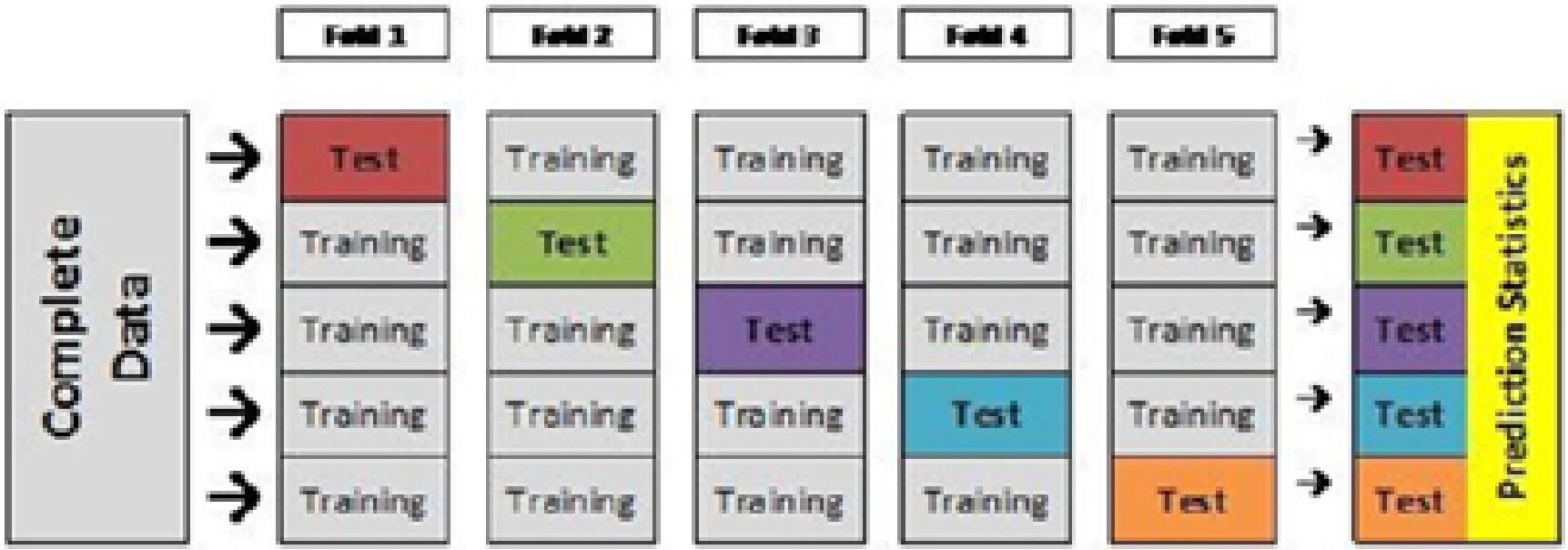
Este método consiste en dividir el conjunto de datos en **submuestras** e iterativamente construir el modelo con todas las observaciones menos las de una submuestra y evaluarlo a continuación con las observaciones de la submuestra **excluida**.

A diferencia de la división *Training - Test*, tiene la ventaja de que **todas las observaciones son predichas** una vez sin formar parte de la construcción del modelo pero **también contribuyen a la construcción** del modelo en el resto de iteraciones.

Es el método de remuestreo **más fiable y utilizado**. Solventa los inconvenientes de *Training - Test* cuando el tamaño del conjunto de datos es muy pequeño y se obtienen resultados muy diferentes según la partición aleatoria.

No obstante, requiere **mucho esfuerzo computacional** ya que es recomendable, al igual que con *Training - Test*, repetirlo con **distintas semillas**.

A la hora de **comparar modelos**, es posible representar en un *boxplot* los resultados obtenidos para validación cruzada, más concretamente, el comportamiento medio para cada semilla. Esto nos permite comprobar el **comportamiento medio** de cada modelo, junto con su **variabilidad**.


Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
Complete Data
Test
Training
Training
Training
Training
Training
Test
Training
Training
Training
Training
Training
Test
Training
Training
Training
Training
Training
Test
Training
Training
Training
Training
Training
Test
Test
Test
Test
Test
Test
Prediction Statistics

Algunos ejemplos de comparación de modelos: