UNIVERSIDAD MARIANO GÁLVEZ.

Facultad de Ingeniería

Minearía de Datos

# Algoritmo de regresión lineal

# Base de datos 2

# Ing.  [Jorge Salvador Santos Neill](http://licenciaturas.umg.edu.gt/moodle/user/view.php?id=122000&course=1)

# Julio Cesar Gil Alvarez 2890-09-3742

# 05 de junio de 2017

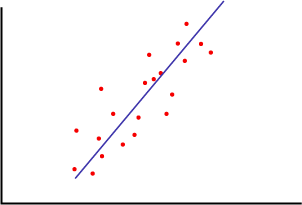
**Introducción**

El análisis estadístico detallado de una serie de cierta importancia, como el total de ventas de una cadena de supermercados o la tirada de un periodo en una ciudad, es de tal amplitud que el acceso a los datos es sólo una parte pequeña del problema. Pero cuando lo que se quiere analizar es, por ejemplo, las ventas porcada punto de venta, o aún más, las ventas de cada producto en cada punto; el número de datos crece vertiginosamente, al mismo tiempo que disminuye la importancia del nodo analizado. Entonces se hace necesaria la búsqueda de nuevas técnicas de estimación que impliquen sustanciosas rebajas en el consumo de recursos, para establecer una relación aceptable de coste y beneficio del análisis. Una posible estrategia consiste en desarrollar métodos de estimación que utilicen exclusivamente las herramientas de la propia base de datos.

**Algoritmo de regresión lineal de Microsoft**

El algoritmo de regresión lineal de Microsoft es una variación del algoritmo de árboles de decisión de Microsoft que ayuda a calcular una relación lineal entre una variable independiente y otra dependiente y, a continuación, utilizar esa relación para la predicción.

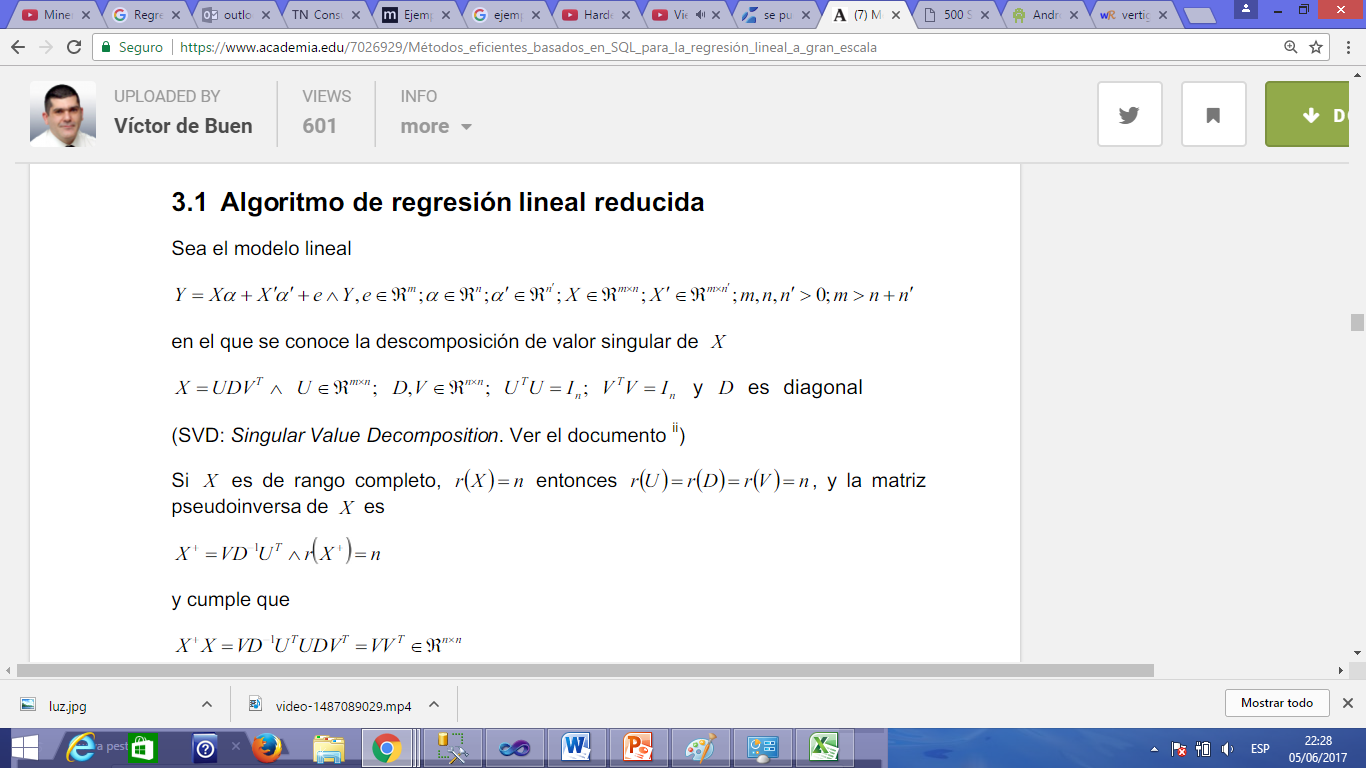
La relación toma la forma de una ecuación para la línea que mejor represente una serie de datos. Por ejemplo, la línea del siguiente diagrama muestra la mejor representación lineal de los datos.

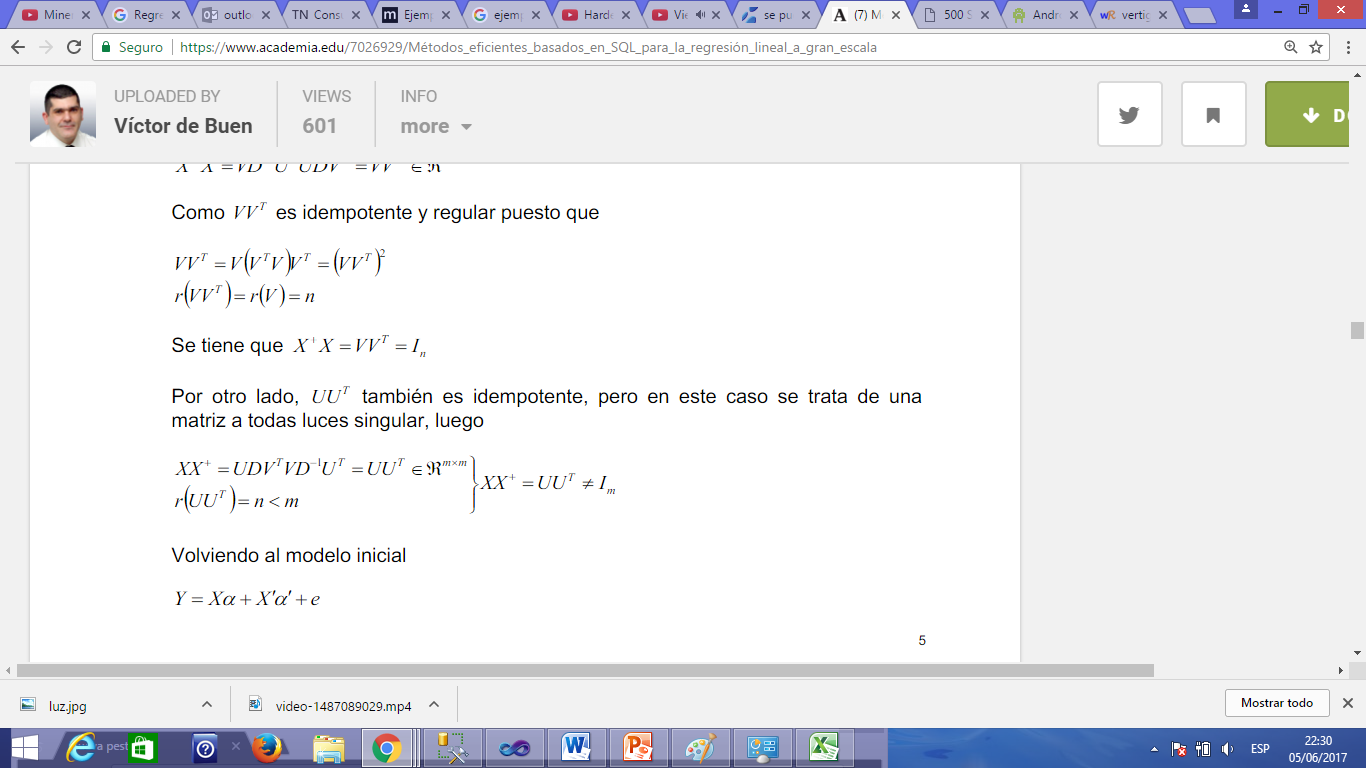


Cada punto de datos del diagrama tiene un error asociado con su distancia con respecto a la línea de regresión. Los coeficientes a y b de la ecuación de regresión ajustan el ángulo y la ubicación de la recta de regresión. Puede obtener la ecuación de regresión ajustando a y b hasta que la suma de los errores asociados a todos los puntos alcance su valor mínimo.

Hay otros tipos de regresión que utilizan varias variables y también hay métodos no lineales de regresión. Sin embargo, la regresión lineal es un método útil y conocido para modelar una respuesta a un cambio de algún factor subyacente.

Algoritmo de regresión lineal Reducida





**Modelos lineales en SQL**

En primer lugar se expone de forma teórica un método de estimación de regresiones lineales por mínimos cuadrados, y después se presenta el código TOL

Necesario para su implementación. Finalmente se muestran los resultados prácticos simulados para hacer patente la potencia del método expuesto. La regresión lineal es el modelo estadístico más sencillo pero no por ello deja de ser útil en multitud de ocasiones, pues bajo las transformaciones pertinentes es capaz de amoldarse a situaciones diversas y resulta de una claridad de exposición indudable.

**Ejemplo**

Puede utilizar la regresión lineal para determinar una relación entre dos columnas continuas. Por ejemplo, puede utilizar la regresión lineal para calcular una línea de tendencias en los datos de fabricación o ventas. También podría utilizar la regresión lineal como precursor para el desarrollo de modelos de minería de datos más complejos, con el fin de evaluar las relaciones entre las columnas de datos.

Aunque hay muchas maneras de calcular la regresión lineal que no requieren herramientas de minería de datos, la ventaja de utilizar el algoritmo de regresión lineal de Microsoft para esta tarea es que se calculan y se prueban automáticamente todas las posibles relaciones entre las variables. No tiene que seleccionar un método de cálculo, como por ejemplo para resolver los mínimos cuadrados. Sin embargo, la regresión lineal podría simplificar en exceso las relaciones en escenarios en los que varios factores afectan al resultado.

**Cómo funciona el algoritmo**

El algoritmo de regresión lineal de Microsoft es una variación del algoritmo de árboles de decisión de Microsoft . Al seleccionar el algoritmo de regresión lineal de Microsoft , se invoca un caso especial del algoritmo de árboles de decisión de Microsoft , con parámetros que restringen el comportamiento del algoritmo y requieren ciertos tipos de datos de entrada. Además, en un modelo de regresión lineal, el conjunto de datos completo se utiliza para calcular las relaciones en el paso inicial, mientras que en un modelo de árboles de decisión estándar los datos se dividen repetidamente en árboles o subconjuntos más pequeños.

**Datos requeridos para los modelos de regresión lineal**

Cuando se preparan datos para utilizarse en un modelo de regresión lineal, se deben entender los requisitos del algoritmo determinado. Esto incluye saber cuántos datos se necesitan y cómo se utilizan. Los requisitos para este tipo de modelo son los siguientes:

* **Una columna de una sola clave** : cada modelo debe contener una columna numérica o de texto que identifique cada registro de manera única. No están permitidas las claves compuestas.
* **Una columna de predicción** . Se requiere al menos una columna de predicción. Se pueden incluir varios atributos de predicción en un modelo, pero deben ser tipos de datos numéricos continuos. No se puede utilizar un tipo de datos de fecha y hora como atributo de predicción aunque el almacenamiento nativo para los datos sea numérico.
* **Columnas de entrada** Deben contener datos numéricos continuos y se les debe asignarse el tipo de datos adecuado.

Para obtener más información, vea la sección Requisitos de [Referencia técnica del algoritmo de regresión lineal de Microsoft](https://msdn.microsoft.com/es-es/library/cc645871.aspx).

Ver un modelo de regresión lineal

Para examinar el modelo, puede utilizar el **Visor de árboles de Microsoft**. La estructura de árbol de un modelo de regresión lineal es muy simple, con toda la información sobre la ecuación de regresión contenida en un nodo único. Para obtener más información, vea [Examinar un modelo usando el Visor de árboles de Microsoft](https://msdn.microsoft.com/es-es/library/ms174503.aspx).

Si desea obtener información más detallada sobre la ecuación, también puede ver los coeficientes y otros detalles utilizando el [Visor de árbol de contenido genérico de Microsoft](https://msdn.microsoft.com/es-es/library/bb677204.aspx).

En un modelo de regresión lineal, el contenido incluye metadatos, la fórmula de regresión y estadísticas sobre la distribución de los valores de entrada. Para obtener más información, vea [Contenido del modelo de minería de datos para los modelos de regresión lineal (Analysis Services - Minería de datos)](https://msdn.microsoft.com/es-es/library/cc645754.aspx).

Crear predicciones

Una vez procesado el modelo, los resultados se almacenan como un conjunto de estadísticas junto con la fórmula de regresión lineal, que se puede utilizar para calcular tendencias futuras. Para obtener ejemplos de consultas que se usan con un modelo regresión lineal, vea [Ejemplos de consultas de modelos de regresión lineal](https://msdn.microsoft.com/es-es/library/cc645906.aspx).

Para obtener información general sobre cómo crear consultas con modelos de minería de datos, vea [Consultas de minería de datos](https://msdn.microsoft.com/es-es/library/bb522459.aspx).

Además de crear un modelo de regresión lineal seleccionando el algoritmo de regresión lineal de Microsoft , si el atributo de predicción es un tipo de datos numéricos continuo, puede crear un modelo de árbol de decisión que contenga regresiones. En este caso, el algoritmo dividirá los datos cuando encuentre puntos de separación adecuados, pero en cambio creará una fórmula de regresión para algunas regiones de datos. Para obtener más información sobre los árboles de regresión en un modelo de árboles de decisión, vea [Contenido del modelo de minería de datos para los modelos de árboles de decisión (Analysis Services - Minería de datos)](https://msdn.microsoft.com/es-es/library/cc645758.aspx).

Comentarios

* No admite el uso del Lenguaje de marcado de modelos de predicción (PMML) para crear modelos de minería de datos.
* No admite la creación de dimensiones de minería de datos.
* Admite la obtención de detalles.
* Admite el uso de modelos de minería de datos OLAP.

# Ejemplos de consultas de modelos de regresión lineal

Cuando se crea una consulta en un modelo de minería de datos, puede tratarse de una consulta de contenido, que proporciona detalles de los patrones detectados durante el análisis, o de una consulta de predicción, que utiliza los patrones del modelo para realizar predicciones de los nuevos datos. Por ejemplo, una consulta de contenido podría proporcionar detalles adicionales sobre la fórmula de regresión, mientras que una consulta de predicción podría indicar si un nuevo punto de datos se ajusta al modelo. También se pueden recuperar metadatos sobre el modelo mediante una consulta.

## Buscar información sobre el modelo de regresión lineal

La estructura de un modelo de regresión lineal es sumamente simple: el modelo de minería de datos representa los datos como un nodo único, que define la fórmula de regresión. Para más información, vea [Contenido del modelo de minería de datos para los modelos de regresión logística (Analysis Services - Minería de datos)](https://msdn.microsoft.com/es-es/library/cc645751.aspx).

[Volver al principio](https://msdn.microsoft.com/es-es/library/cc645906.aspx#bkmk_top)

### Consulta de ejemplo 1: usar el conjunto de filas de esquema de minería de datos para determinar los parámetros que se usan para un modelo

Al consultar el conjunto de filas de esquema de minería de datos, puede buscar los metadatos acerca del modelo. Podría incluirse cuándo se creó el modelo, cuándo se procesó en último lugar, el nombre de la estructura de minería de datos en la que se basa y el nombre de la columna que se usa como atributo de predicción. También se pueden devolver los parámetros que se utilizaron cuando se creó el modelo por primera vez.

SELECT MINING\_PARAMETERS

FROM $system.DMSCHEMA\_MINING\_MODELS

WHERE MODEL\_NAME = 'TM\_PredictIncome'

Resultados del ejemplo:

| **MINING\_PARAMETERS** |
| --- |
| COMPLEXITY\_PENALTY=0.9, MAXIMUM\_INPUT\_ATTRIBUTES=255, MAXIMUM\_OUTPUT\_ATTRIBUTES=255, MINIMUM\_SUPPORT=10, SCORE\_METHOD=4, SPLIT\_METHOD=3, FORCE\_REGRESSOR= |

**Usar el algoritmo**

Utilice el Visor de árboles de Microsoft para explorar un modelo de minería de datos de regresión lineal.

Un modelo de regresión lineal debe contener una columna de clave, columnas de entrada y al menos una columna de predicción.

El algoritmo Regresión lineal de Microsoft admite los tipos de contenido de columna de entrada, tipos de contenido de columna de predicción e indicadores de modelado específicos que se enumeran en la siguiente tabla.

|  |  |
| --- | --- |
| Tipos de contenido de columna de entrada | Continuous, Cyclical, Key, Table y Ordered |
| Tipos de contenido de columna de predicción | Continuous, Cyclical y Ordered |
| Indicadores de modelado | NOT NULL y REGRESSOR |

Todos los algoritmos de Microsoft son compatibles con un conjunto común de funciones. No obstante, el algoritmo Regresión lineal de Microsoft admite las funciones adicionales que se enumeran en la siguiente tabla.

|  |  |
| --- | --- |
| [IsDescendant](https://msdn.microsoft.com/es-es/library/ms132097(v=sql.90).aspx) | [PredictStdev](https://msdn.microsoft.com/es-es/library/ms131998(v=sql.90).aspx) |
| [IsInNode](https://msdn.microsoft.com/es-es/library/ms132068(v=sql.90).aspx) | [PredictSupport](https://msdn.microsoft.com/es-es/library/ms132140(v=sql.90).aspx) |
| [PredictHistogram](https://msdn.microsoft.com/es-es/library/ms132060(v=sql.90).aspx) | [PredictVariance](https://msdn.microsoft.com/es-es/library/ms132165(v=sql.90).aspx) |
| [PredictNodeId](https://msdn.microsoft.com/es-es/library/ms132162(v=sql.90).aspx) |  |

Para consultar una lista de las funciones comunes a todos los algoritmos de Microsoft, vea [Algoritmos de minería de datos](https://msdn.microsoft.com/es-es/library/ms175595(v=sql.90).aspx). Para obtener más información acerca del modo de utilizar estas funciones, vea [Referencia de funciones de Extensiones de minería de datos (DMX)](https://msdn.microsoft.com/es-es/library/ms131996(v=sql.90).aspx).

El algoritmo Regresión lineal de Microsoft es compatible con varios parámetros que influyen en el rendimiento y la precisión del modelo de minería de datos resultante. Estos parámetros se describen en la tabla siguiente.

|  |  |
| --- | --- |
| **Parámetro** | **Descripción** |
| *MAXIMUM\_INPUT\_ATTRIBUTES* | Define el número de atributos de entrada que puede administrar el algoritmo antes de invocar la selección de características. Establezca este valor en 0 para desactivar la selección de características.  El valor predeterminado es 255. |
| *MAXIMUM\_OUTPUT\_ATTRIBUTES* | Define el número de atributos de salida que puede administrar el algoritmo antes de invocar la selección de características. Establezca este valor en 0 para desactivar la selección de características.  El valor predeterminado es 255. |
| *FORCED\_REGRESSOR* | Impone al algoritmo la utilización de las columnas indicadas como regresores, independientemente de su importancia según los cálculos del algoritmo. |