ntic master

School

UNIVERSIDAD COMPLUTENSE DE MADRID

Selección de variables en modelos de regresión

Minería de Datos y Modelización Predictiva

Master Big Data y Data Science. Aplicaciones al Comercio, Empresa y Finanzas
Universidad Complutense de Madrid
Curso 2021-2022

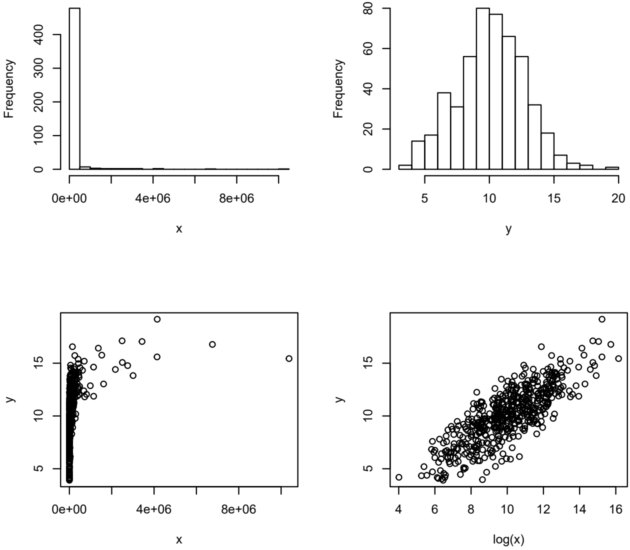
ntic master
revolucionamos la comunicación

UNIVERSIDAD COMPLUTENSE DE MADRID

TRANSFORMACIONES

A menudo es necesario **transformar las variables** implicadas en el modelo para **mejorar la bondad** del mismo:

- Variables independientes:
  - **Cualitativas**: en ocasiones, cuando alguna de las categorías no resulta **significativa**, conviene **agrupar las categorías** (teniendo en cuenta su relación con la variable objetivo) para así **simplificar el modelo** sin perder poder predictivo. Esto puede hacerse a partir de la función *recode*.
  - **Continuas**: en este caso, se puede llevar a cabo **cualquier transformación** que se desee siempre y cuando no de lugar a missings (lo que incluye la discretización de la variable). En particular, si la relación con la variable objetivo no es lineal, se puede realizar alguna transformación sobre la variable independiente que de lugar a linealidad. Nótese que en ese caso la **interpretación de los parámetros** se modifica según la transformación. Cabe destacar, además, que, en ocasiones, puede ocurrir que la variable independiente aporte información en su forma **original y transformada**, por lo que se **mantienen ambas** en el modelo.



Frequency
0
100
200
300
400
0e+00
4e+06
8e+06
x
Frequency
0
20
40
60
80
5
10
15
20
y
y
5
10
15
0e+00
4e+06
8e+06
x
y
5
10
15
4
6
8
10
12
14
16
log(x)

Uno de los pasos más importantes en la fase de modelización es la **selección de variables** que van a formar parte del modelo, pues esto será lo que determine su calidad.

El análisis de las relaciones entre las variables input y objetivo a través de estadísticos (como la correlación o la V de Cramer) o de gráficos es altamente **recomendable** de cara a detectar variables que realmente **no aportan nada** a la variable objetivo y que puedan, por tanto, **rechazarse** en la fase de modelización. De esta forma, conseguimos reducir el tiempo de computación de la selección de variables que estudiaremos a continuación. No obstante, esta “selección previa” debe realizarse con cautela y únicamente cuando estemos trabajando con muchas variables y/o el *no efecto* de las variables sea muy evidente.

Existen **3 métodos** de selección de variables *clásicos*:

- **Forward o hacia delante**: Este método consiste en, partiendo desde cero, ir introduciendo **una a una** las variables que mayor mejora produzcan en el modelo hasta que no haya ninguna variable más fuera del modelo que **aporte información**. La mejora en el modelo se puede medir de distintas formas, como veremos más adelante.

  Una vez que una variable entra en el modelo, **no puede salir**.

- **Backward o hacia atrás**: Este método consiste en, partiendo del modelo que contiene todas las variables, ir eliminando **una a una** las variables que menos influyan en el modelo hasta el modelo no mejore con la eliminación de ninguna de ellas.

  Una vez que una variable se elimina, **no puede volver a entrar en el modelo**.
- **Stepwise o paso a paso**: Este método es una **mezcla** de los anteriores. El método es similar al forward, salvo por que **se pueden eliminar** las variables que han entrado en el modelo (ya que al entrar alguna otra pudiera hacer no significativo su aporte). La eliminación de las variables se hace de acuerdo al método backward.

  Por lo tanto, en cada paso se evaluan todas las posibles variables a eliminar y a introducir y se selecciona aquella acción que mayor mejora produzca en el modelo.

  Es recomendable incluir un número **máximo de iteraciones** para paliar la posible aparición de bucles de entrada-salida de una misma variable.

El modelo de regresión ***LASSO*** (Least Absolute Shrinkage and Selection Operator) se basa en limitar el número de parámetros del modelo de regresión clásico, por lo que puede verse como un método de **selección de variables**.

Para el modelo de regresión lineal, este consiste en obtener los valores de β tales que:

$$\min_{\beta} \sum_{i=1}^{n} [y - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m)]^2$$

$$\text{s.a.} \sum_{j=1}^{m} |\beta_j| = t$$

Siempre que se asigne al parámetro t un valor inferior a la suma de los parámetros minimo-cuadráticos, los betas se verán **reducidos**.

Nótese que la expresión anterior se puede expresar alternativamente como:

$$\min_{\beta} SSE + \lambda \sum_{j=1}^{m} |\beta_j|$$

La dificultad consiste en determinar el **valor óptimo de** t/λ. Lo habitual es probar con varios valores y seleccionar el que **mejores resultados** ofrezca.

El modelo de regresión logística LASSO es análogo al anterior pero, en lugar de minimizar la SSE, se maximiza la verosimilitud, siempre sujeto a la restricción de los parámetros.