Uso de los MLG en los casos de asociaciones no linealeales entre las variables

Angie Rodríguez Duque & César Saavedra Vanegas

Octubre de 2020

Introducción

Para comprender las asociaciones no lineales resulta necesario retomar los conceptos de correlación y de regresión lineal simple, los cuales corresponden a métodos estadísticos que estudian la relación lineal existente entre dos variables. De esta manera, es posible identificar diferencias entre si:

- En primer lugar, la correlación cuantifica qué tan relacionadas se encuentran dos variables, mientras que la regresión lineal se encargar de generar un modelo basado en la relación existente entre ambas variables con el objetivo de predecir el valor de una a partir de la otra.
- El cálculo de la correlación entre dos variables mide tan solo la relación entre ambas sin considerar dependencias. Por otro lado, en el caso de la regresión lineal, el modelo varía según qué variable se considere dependiente de la otra.
- Finalmente, se debe tener en cuenta que primero se analiza si ambas variables están correlacionadas y, en caso de estarlo, se procede a generar el modelo de regresión.

Correlación lineal

La correlación lineal analiza como su nombre lo indica la relación lineal existente entre dos variables continuas. Para ello, es necesario contar con determinados parámetros que permitirán cuantificar dicha relación. Entre ellos se tiene: La covarianza, que indica el grado de variación conjunta de dos variables aleatorias, su expresión está dada por:

$$Covarianza\ \ muestral = Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

Donde:

- $\bar{x}$ e $\bar{y}$: Son la media de cada variable
- x_i e y_i : Son el valor de las variables para la observación i.

Coeficiente de correlación de Pearson:

Se emplea para estudiar la asociación entre un factor de estudio y una variable de respuesta cuantitativa. Así, mide el grado de asociación entre dos variables tomando rango de valores entre -1 y 1.

- **Población:**

$$\rho = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

- **Muestra:**

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Coeficiente de correlación de Spearman:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Siendo d_i la distancia entre los rangos de cada observación $(x_i - y_i)$ y n el número de observaciones.

Coeficiente Tau de Kendall:

$$\tau = \frac{C - D}{\frac{1}{2}n(n-1)}$$

Siendo C el número de pares concordantes, aquellos en los que el rango de la segunda variable es mayor que el rango de la primera variable. D el número de pares discordantes, cuando el rango de la segunda es igual o menor que el rango de la primera variable.

Después de exponer los tres coeficientes de asociación es importante identificar las principales diferencias entre ellos, estas son:

- La correlación de Pearson funciona bien con variables cuantitativas que tienen una distribución normal. Es más sensible a los valores extremos que las otras dos alternativas.
- La correlación de Spearman se emplea cuando los datos son ordinales, de intervalo, o bien cuando no se satisface la condición de normalidad para variables continuas y los datos se pueden transformar a rangos. Es un método no paramétrico.
- La correlación de Kendall es otra alternativa no paramétrica para el estudio de la correlación que trabaja con rangos. Se emplea cuando se dispone de pocos datos y muchos de ellos ocupan la misma posición en el rango, es decir, cuando hay muchas ligaduras.

Regresión

Regresión lineal simple: Consiste en generar un modelo de regresión (ecuación de una recta) que permita explicar la relación lineal que existe entre dos variables.

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

- β_0 : La ordenada en el origen.
- β_1 : La pendiente.
- ϵ : El error aleatorio.

Regresión lineal múltiple:

Es una extensión de la regresión lineal simple. Permite generar un modelo lineal en el que el valor de la variable dependiente o respuesta (Y) se determina a partir de un conjunto de variables independientes llamadas predictores $(X_1, X_2, X_3, ...)$

$$Y_i = (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_n X_{ni}) + \epsilon_i$$

Regresión polinómica

La Regresión Polinomial permite describir relaciones no lineales, La forma más sencilla de hacerlo es incorporar flexibilidad a un modelo lineal introduciendo nuevos predictores obtenidos al elevar a distintas potencias el predictor original.

Partiendo del modelo lineal:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Se obtiene un modelo polinómico de grado d a partir de la ecuación:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + ... + \beta_d x_i^d + \epsilon_i$$

Estandarizacion de las variables: Una asociación en forma de U se puede modelar agregando una versión cuadrática de la variable y un parámetro β adicional:

$$E(Y_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 \qquad\qquad i = 1, ..., N.$$

En la práctica, cuando se utilizan transformaciones como la cuadrática, que pueden crear valores grandes de x_i, puede resultar útil centrar las variables explicativas utilizando su media ($\bar{x}$) y escalar utilizando su desviación estándar (sd):

$$\tilde{x}_i = \frac{(x_i - \bar{x}_i)}{sd}$$

Y se ajusta al modelo:

$$E(Y_i) = \beta_0 + \beta_1 \tilde{x}_i + \beta_2 \tilde{x}_i^2 \qquad\qquad i = 1, ..., N.$$

Modelos lineales generalizados (GLM)

Los modelos polinómicos se pueden ajustar mediante regresión lineal por mínimos cuadrados ya que, aunque generan modelos no lineales, su ecuación no deja de ser una ecuación lineal con predictores $x, x_2, x_3, ..., x_d$.

Por esta misma razón, las funciones polinómicas pueden emplearse en regresión logística para predecir respuestas binarias. Solo es necesario realizar una transformación logit.

$$P(y_i > Y | x_i = X) = \frac{exp\left(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + ... + \beta_d x_i^d\right)}{1 + exp\left(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + ... + \beta_d x_i^d\right)}$$

Modelos Aditivos Generalizados (GAM)

Una forma natural de extender el modelo de regresión lineal múltiple para permitir relaciones no lineales entre cada característica y la respuesta es reemplazar cada componente lineal $\beta_j x_{ij}$ con una función no lineal (suave) $f_j(x_{ij})$. Lo que darìa como resultado el siguiente modelo

$$y_i = \beta_0 + \sum_{i=1}^{n} f_j(x_{ij}) + \epsilon$$

Se denomina modelo aditivo GAM porque calculamos una f_j para cada X_j y luego sumamos todas sus contribuciones.

Los modelos GAM también se pueden usar en situaciones donde Y es una variable cualitativa. A partir del modelo logístico:

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

podemos recordar que la transformación logit corresponde al logaritmo de odds de $P(Y = 1|X)$ versus $P(Y = 0|X)$. Incorporando funciones no lineales $f_p(x_p)$ entre variables obtenemos el modelo el modelo de regresión logística GAM.

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + f_1(X_1) + f_2(X_2) + ... + f_p(X_p)$$

- Los GAM nos permiten ajustar un f_j no lineal a cada X_j, de modo que podamos modelar automáticamente relaciones no lineales que la regresión lineal estándar perderá. Esto significa que no necesitamos probar manualmente muchas transformaciones diferentes en cada variable individualmente.
- Los ajustes no lineales pueden potencialmente hacer predicciones más precisas para la respuesta Y.
- Debido a que el modelo es aditivo, aún podemos examinar el efecto de cada X_j en Y individualmente mientras se mantienen fijas todas las demás variables. Por lo tanto, si estamos interesados en la inferencia, los GAM proporcionan una representación útil.
- La suavidad de la función f_j para la variable X_j se puede resumir mediante grados de libertad.
- La principal limitación de los GAM es que el modelo está restringido a ser aditivo.

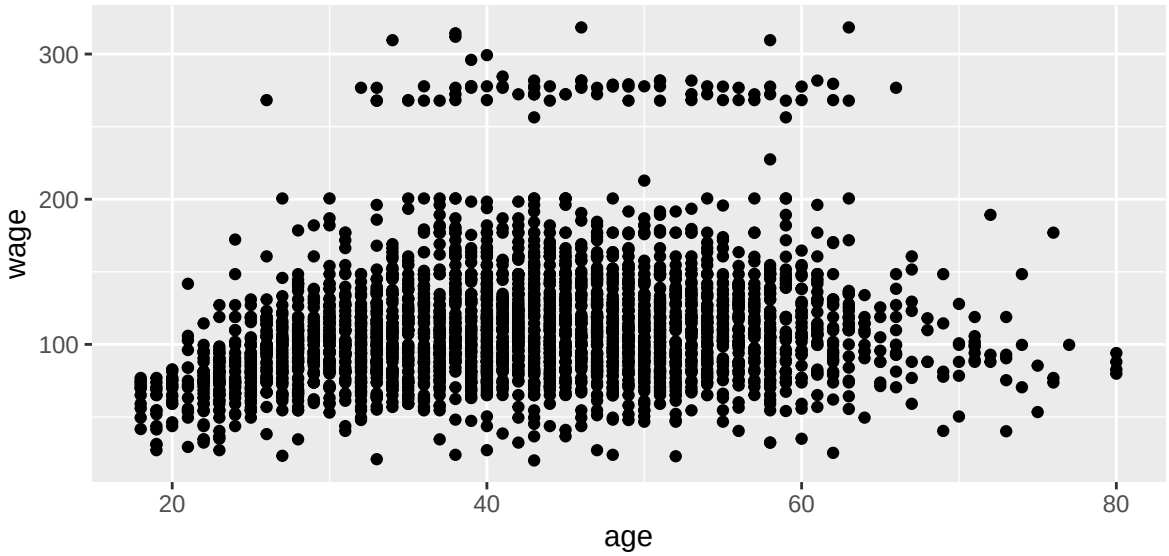
Ejemplo de aplicación

El set de datos *Wage* del paquete ISRL contiene información sobre 3000 trabajadores. Entre las 12 variables registradas se encuentra el salario (wage) y la edad (age). Dada la relación no lineal existente entre estas dos variables, se recurre a un modelo polinómico de grado 4 que permita predecir el salario en función de la edad.

```
suppressMessages(library(ISLR))
suppressMessages(library(boot))
suppressMessages(library(plotly))
data("Wage")
```

Inicialmente se realiza la representación gráfica de los datos a los cuales se desea ajustar el modelo, en la figura se observa como el ajuste de un modelo lineal simple no es la mejor opción, para lo cual se recurre a un modelo lineal generalizado para asociaciones no lineales con una variable de respuesta binaria como se presentara a continuación.

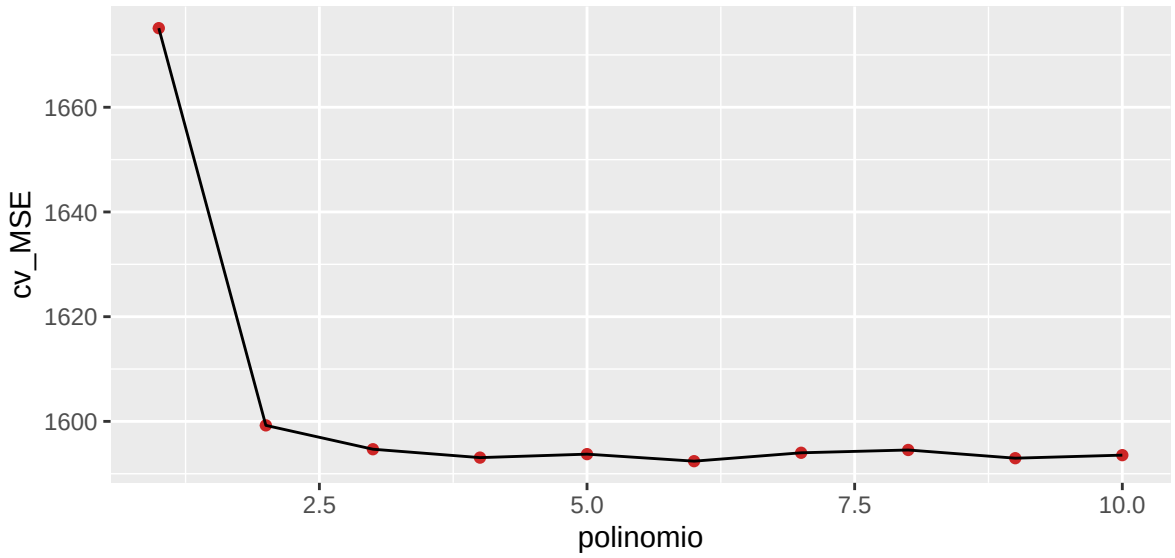
```
ggplot(Wage, aes(x = age, y = wage)) + geom_point(colour = "black")
```



Validación cruzada

Mediante cross-validation se identifica con que polinomio se consigue el mejor modelo. El proceso consiste en ajustar un modelo para cada grado de polinomio y estimar su test error (Mean Square Error). El mejor modelo es aquel a partir del cual ya no hay una reducción sustancial del test error, esto se evidencia a partir del polinomio de grado 4 tal y como se presenta en la figura donde se observa como se estabiliza el error cuadratico medio.

```
p4
```



Variable binaria

Posteriormente se realiza la creación de una variable binaria para aquellos salarios > 250000 dolares para de esta forma ajustar el modelo con respuesta binaria.

```
Wage$wage_superior250 <- I(Wage$wage > 250)
table(Wage$wage_superior250)
```

```
##
## FALSE  TRUE
##  2921    79
```

Donde aquellas personas que cumplan con tener un salario mayor a $250000 dólares serán clasificados como "True" y los que no como "False".

Se ajustan tres modelos Logit, esto teniendo en cuenta el resultado obtenido mediante validacion cruzada que arrojo que el polinomio con mejor ajuste es el de grado 3.

```
modelo_logit <- glm(wage_superior250  ~ poly(age, 2), family = "binomial", data = Wage)

modelo_logit1 <- glm(wage_superior250  ~ poly(age, 3), family = "binomial", data = Wage)

modelo_logit2 <- glm(wage_superior250  ~ poly(age, 4), family = "binomial", data = Wage)
```

Ajuste del modelo logístico

Al emplear la función predict() con un modelo logístico, es importante tener en cuenta que por defecto se devuelve el logaritmo de ODDs (Log_ODDs). Para transformarlos en probabilidad se invierte la función logística:

$$P\left(Y = 1|X = \frac{e^{LogODDs}}{1 + e^{LogODDs}}\right)$$

Es posible obtener directamente la probabilidad de las predicciones seleccionando el argumento type = "response".

A pesar de que esta forma es más directa, si junto al valor predicho se quiere obtener su intervalo de confianza y que este caiga dentro de [0, 1], se tienen que realizar los cálculos con los Log ODDs para finalmente transformarlos a probabilidad.

Comparación de modelos logit

Grado (p)	Df	Deviance	Resid. Df	Resid. Dev
Nulo			2999	730.53
2	2	21.511	2997	709.02
3	3	22.613	2996	707.92
4	4	29.315	2995	701.22

Es posible observar como en la tabla anterior se muestran los resultados obtenidos para los distintos modelos ajustados, esto es, modelos de grado 1, 2, 3 y 4 para los cuales se calcula la devianza, la cual es la encargada de explicar la varianza de cada uno de los modelos ajustados. Teniendo en cuenta lo anteriormente mencionado se tiene que a medida que se incrementa el grado del polinomio, el ajuste que se realiza a los datos es mayor y por lo tanto se tiene una mayor explicación de la varianza, mediante la validación cruzada se obtuvo que el grado 4 seria el grado de polinomio que mejor ajuste los datos y esto se evidencia en que este es quien logra la devianza mas alta, esto es, 29.31% de la variabilidad total de los datos, lo cual indica que se tiene un porcentaje de explicación alto para los datos esto en comparación a si se ajustara un modelo lineal simple.

Bibliografía

- Dobson, A. J., & Barnett, A. G. (2018). An introduction to generalized linear models. CRC press.

- Faraway, J. J. (2014). Linear models with R. CRC press.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2015). An Introduction to Statistical Learning with Applications in R, Edn. 6th.