Asociaciones no lineales en el modelo Normal

Angie Rodríguez Duque & César Saavedra Vanegas

Octubre 30 de 2020

Introducción

Relación lineal

La correlación lineal y la regresión lineal simple son métodos estadísticos que estudian la relación lineal existente entre dos variables.

Diferencias:

- La correlación cuantifica como de relacionadas están dos variables, mientras que la regresión lineal genera un modelo, el cual se basa de la relación existente entre ambas variables, para predecir el valor de una a partir de la otra.

- El cálculo de la correlación entre dos variables mide únicamente la relación entre ambas sin considerar dependencias. En el caso de la regresión lineal, el modelo varía según qué variable se considere dependiente de la otra.

- Primero se analiza si ambas variables están correlacionadas y, en caso de estarlo, se procede a generar el modelo de regresión.

Correlación lineal

Para estudiar la relación lineal existente entre dos variables continuas es necesario disponer de parámetros que permitan cuantificar dicha relación. Uno de estos parámetros es la covarianza, que indica el grado de variación conjunta de dos variables aleatorias.

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

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 utiliza para estudiar la asociación entre un factor de estudio y una variable de respuesta cuantitativa, mide el grado de asociación entre dos variables tomando 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.

Las principales diferencias entre estos tres coeficientes de asociación 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 $(X_1, X_2, X_3, \dots)$.

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

Regresión polinómica

Antes de aplicar un modelo de regresión lineal simple, se hace necesario conocer si los datos se pueden ajustar a un modelo de regresión lineal, es decir conocer el grado de asociación entre la variable de respuesta y las variables predictoras y a su vez poder determinar la proporción de variabilidad existente entre la variable dependiente explicada por la variable independiente.

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:

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

Centramiento de las variables

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

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

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:

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

Ventajas

- Una ventaja adicional del centrado es que la estimación de la intersección β_0 ahora relaciona el valor de $\bar{y}$ con el valor de $\bar{x}$ en lugar del valor de y promedio cuando x es cero, lo que puede no ser significativo si x no puede ser cero. **Ejemplo:** El peso de una persona.

- Además, los parámetros de **"Slope"** ahora representan un cambio de una desviación estándar que es potencialmente más significativo que un cambio de una sola unidad que puede ser muy pequeño o grande.

- Por último, escalar por la desviación estándar facilita la comparación de la importancia de las variables.

Modelos lineales generalizados (GLM)

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 $x, x_2, x_3, \ldots, x_d$.

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

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

Recomendaciones:

- No se aconseja el uso de modelos polinómicos con grado mayor de 3 o 4 debido a un exceso de flexibilidad (overfitting), principalmente en los extremos del predictor X.
- La selección del grado de polinomio óptimo puede hacerse mediante cross validation.

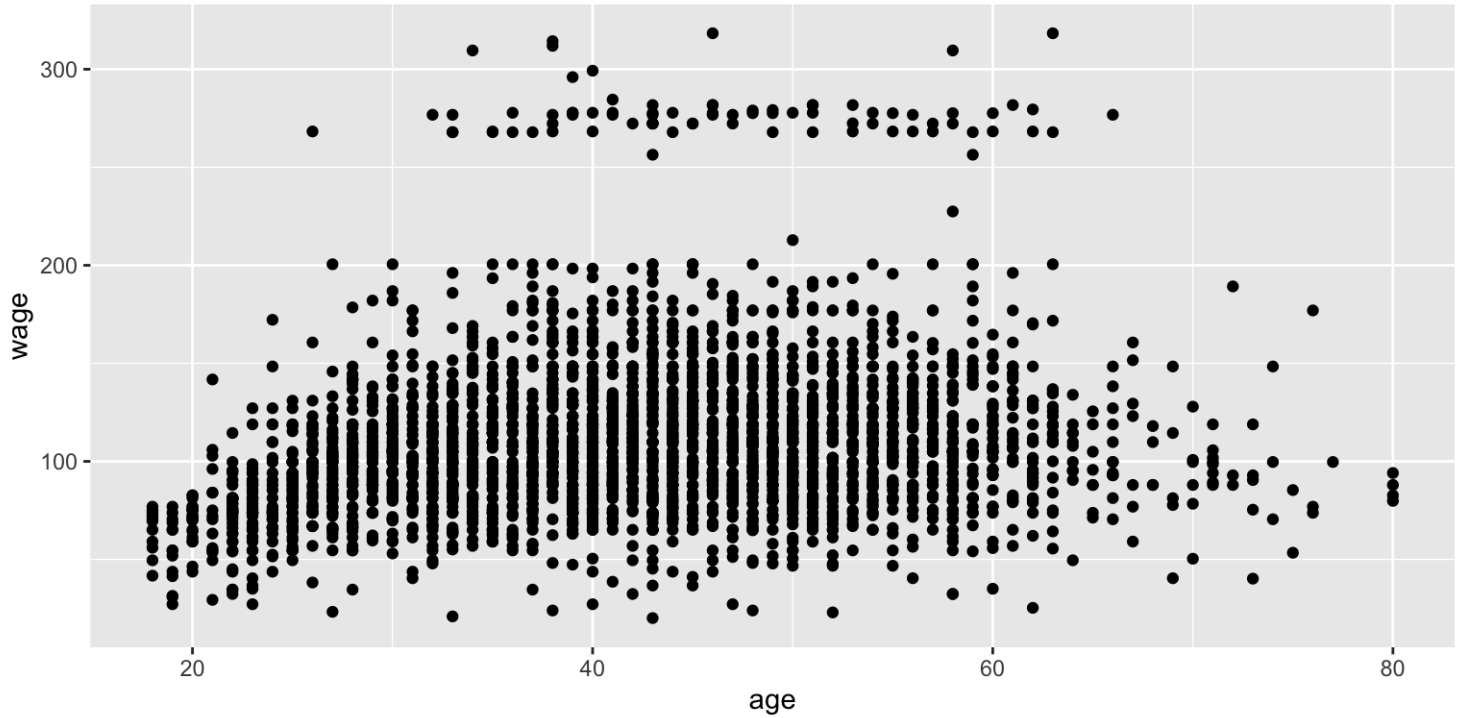
Ejemplo en R

Conjunto de datos

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")
```

Representación gráfica



Comparación de modelos por contraste de hipótesis ANOVA

```
modelo_1 <- lm(wage ~ age, data = Wage)
modelo_2 <- lm(wage ~ poly(age, 2), data = Wage)
modelo_3 <- lm(wage ~ poly(age, 3), data = Wage)
modelo_4 <- lm(wage ~ poly(age, 4), data = Wage)
```

```
anova(modelo_1, modelo_2, modelo_3, modelo_4)
```

```
## Analysis of Variance Table
##
## Model 1: wage ~ age
## Model 2: wage ~ poly(age, 2)
## Model 3: wage ~ poly(age, 3)
## Model 4: wage ~ poly(age, 4)
##   Res.Df     RSS Df Sum of Sq        F    Pr(>F)
## 1   2998 5022216
## 2   2997 4793430  1    228786 143.6025 < 2.2e-16 ***
## 3   2996 4777674  1     15756   9.8894  0.001679 **
## 4   2995 4771604  1      6070   3.8101  0.051039 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

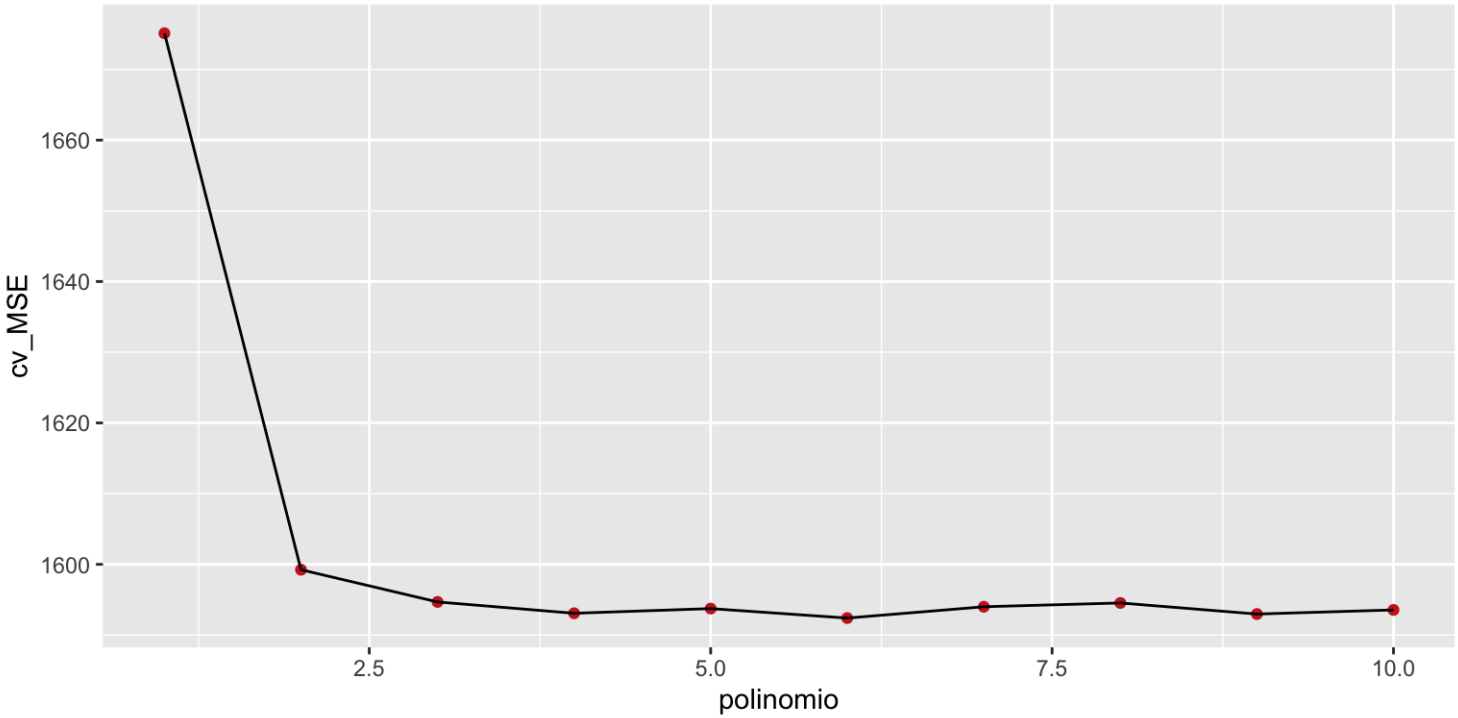
Validación cruzada

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.

```
cv_MSE_k10 <- rep(NA,10)

for (i in 1:10) {
  modelo <- glm(wage ~ poly(age, i), data = Wage)
  set.seed(17)
  cv_MSE_k10[i] <- cv.glm(data = Wage, glmfit = modelo, K = 10)$delta[1]
}
p4 <- ggplot(data = data.frame(polinomio = 1:10, cv_MSE = cv_MSE_k10),
             aes(x = polinomio, y = cv_MSE)) +
      geom_point(colour = c("firebrick3")) +
      geom_path()
```

Validación cruzada



Ejemplo 2. Regresión polinómica logística

Se genera la variable categórica

Se realiza la creación de una variable binaria para aquellos salarios > 250000 dolares para ajustar el modelo.

```
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”.

Ajuste del modelo logístico

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

Regresión polinómica logística grado 2

```
anova(modelo_logit)
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: wage_superior250
##
## Terms added sequentially (first to last)
##
##
##              Df Deviance Resid. Df Resid. Dev
## NULL                          2999     730.53
## poly(age, 2)  2   21.511      2997     709.02
```

Regresión polinómica logística grado 3

```
anova(modelo_logit1)
```

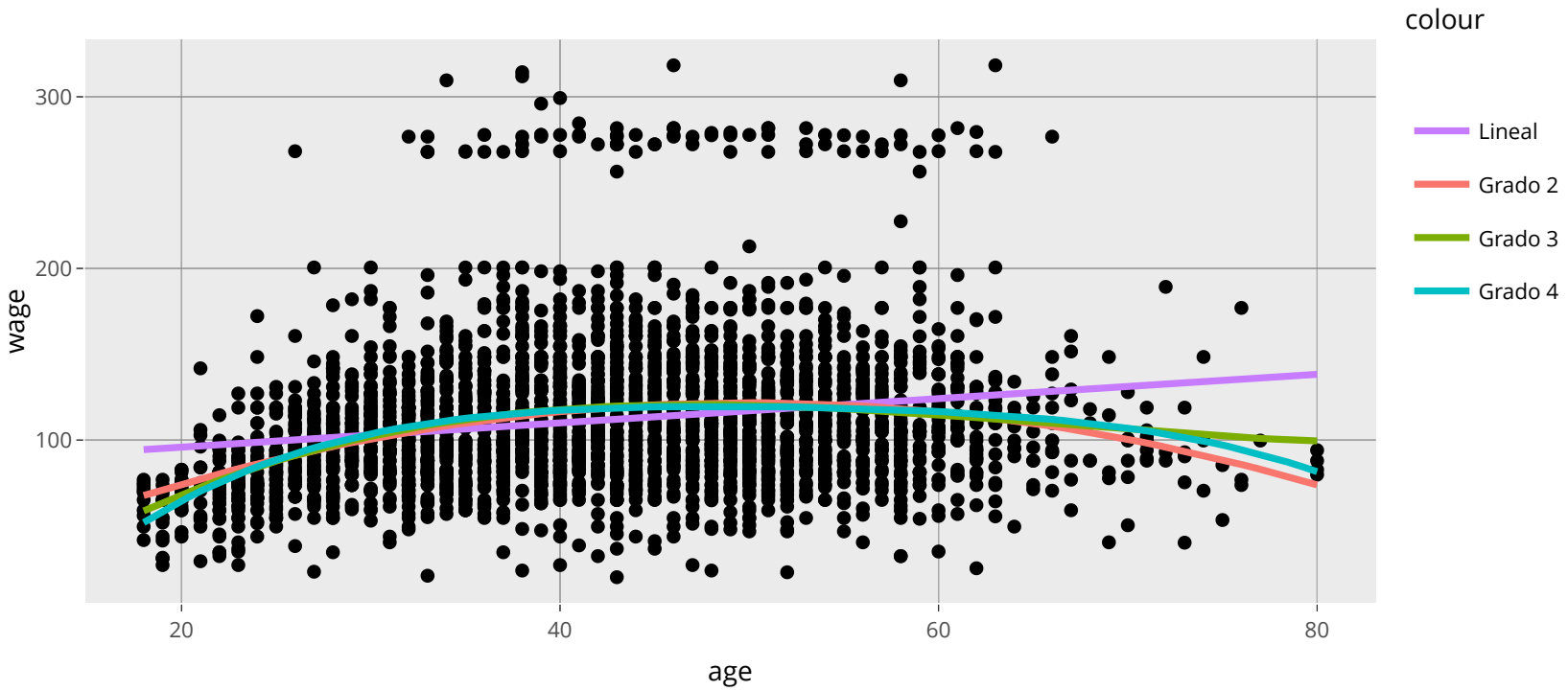
```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: wage_superior250
##
## Terms added sequentially (first to last)
##
##
##              Df Deviance Resid. Df Resid. Dev
## NULL                          2999     730.53
## poly(age, 3)  3   22.613      2996     707.92
```

Regresión polinómica logística grado 4

```
anova(modelo_logit2)
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: wage_superior250
##
## Terms added sequentially (first to last)
##
##
##              Df Deviance Resid. Df Resid. Dev
## NULL                          2999     730.53
## poly(age, 4)  4   29.315      2995     701.22
```

```
G4 <- ggplot(Wage, aes(x = age, y = wage)) + geom_point(colour = "black") +
  stat_smooth(method = 'lm', formula = y ~ poly(x, 1), aes(colour = 'Lineal'), se = FALSE) +
  stat_smooth(method = 'glm', formula = y ~ poly(x, 2), aes(colour = 'Grado 2'), se = FALSE) +
  stat_smooth(method = 'glm', formula = y ~ poly(x, 3), aes(colour = 'Grado 3'), se = FALSE) +
  stat_smooth(method = 'glm', formula = y ~ poly(x, 4), aes(colour = 'Grado 4'), se = FALSE)
```



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.

¡Gracias por tu atención!