Modelos lineales y aprendizaje estadístico © EDICIONES ROBLE, S.L.

Indice

Modelos lineales y aprendizaje estadístico

I. Introducción

El objetivo principal del aprendizaje estadístico es explicar la influencia de variables de los datos sobre otras, ayudar a entenderlas y generar modelos útiles en el ámbito científico o de negocio.

Los modelos estadísticos sirven para predecir los valores de una variable Y respecto a una variable X, así como para establecer reglas que ayuden a comprender los comportamientos de las mismas. Estas reglas tienen en cuenta la incertidumbre y el azar de los fenómenos estudiados.

Cuando creamos un modelo siempre simplificamos la visión de la realidad a través de un marco matemático cerrado. Este modelo debe ser capaz de recoger de los datos conocidos el patrón que está contenido en la expresión matemática elegida y fijar los parámetros pertinentes.

Una vez fijado el tipo de modelo que va a usarse, se realiza el proceso de aprendizaje estadístico conveniente, que suele ser un método numérico.

En los modelos lineales, se supone que la variación de la variable Y a tratar se puede explicar mediante una proporción de la variable predictora X:

$$Y = a X + b$$

En esta notación:

- a es el **coeficiente o pendiente (coeficient en R)**, indica la proporción de variación de X que pasa a Y a través del modelo.
- b es el **término independiente (intercept term en R)**, indica el estado base en el que se encuentra el valor de la variable Y, si la variable X es nula.

A su vez, a la variable X se le denominará **variable independiente** o **predictor** y a la variable Y se le denominará **variable objetivo.**

II. Objetivos específicos

- Conocer en profundidad el concepto de modelo estadístico.
- Ver características de los modelos de regresión lineal.
- Entender lo que es la correlación y las dependencias lineales entre variables.
- Comprender la dicotomía sesgo-varianza.
- Aprender las métricas estadísticas que se usan para comparar modelos.

III. Análisis multivariable: covarianza y correlación

La relación entre dos atributos de una misma población puede ser:

Dependencia

La variabilidad de los atributos sigue alguna ley de relación. El valor que toma un atributo tiene relación estadística respecto al que toma el otro.

Independencia

La variabilidad de los atributos no está relacionada. Los valores que toma cada uno no siguen ninguna ley común.

La medida de dependencia lineal entre atributos se define como la**covarianza/correlación** de los mismos.

Imagen 4.1. Relación lineal. *Fuente:* elaboración propia.

Para estudiar esta relación lineal entre dos variables, existen los siguientes coeficientes:

Coeficiente de correlación R

Coeficiente de correlación R: toma valores entre -1 y 1:

- **R=1**: relación lineal perfecta positiva, más de una variable implica más de la otra. Primer caso en la imagen que se muestra a continuación.
- **R=-1**: relación lineal perfecta negativa, más de una variable implica menos de la otra. Segundo caso en la imagen mostrada a continuación.
- **R=0:** no existe relación lineal, las variaciones de una variable no tienen influencia en la otra de manera proporcional directa. Tercer caso en la imagen.

Coeficiente de determinación lineal R^2

Coeficiente de determinación lineal R^2: toma valores entre 0 y 1. Representa la proporción de la variabilidad de la variable Y, que queda explicada por la variable X. Si vamos a construir modelos lineales, estamos interesados en que los predictores usados tengan el mayor R^2 con la variable objetivo Y.

$$R = \frac{Cov[X, Y]}{\sqrt{Var[X]Var[Y]}} \in [-1, 1], \qquad R^2 \in [0, 1]$$

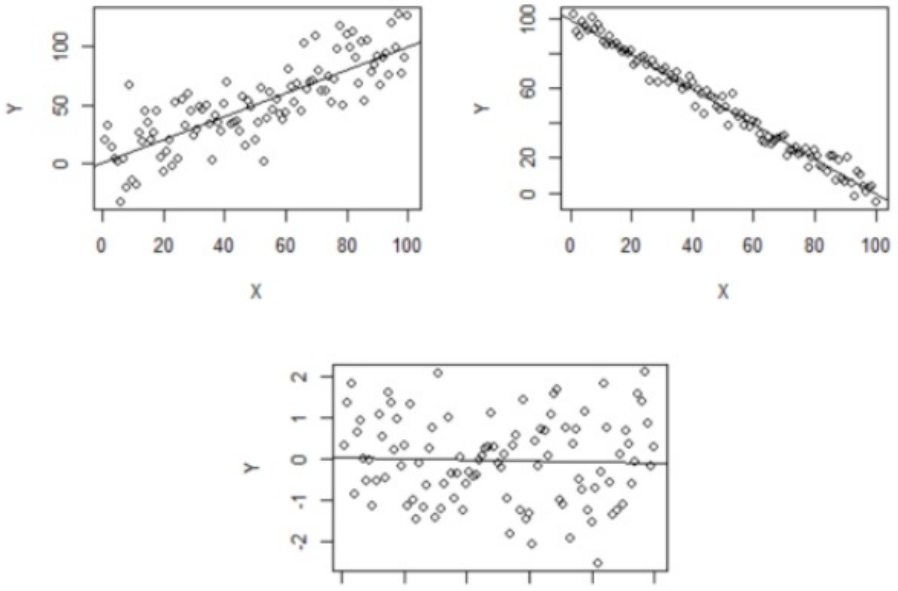

Imagen 4.2. Correlaciones. *Fuente:* elaboración propia con R.

Descarga: Consulta el notebook UD4 N01

Descárgate el archivo UD4_N01 y ejecútalo en R. También puedes verlo en.html

IV. Regresión lineal univariable

La regresión (o modelos lineales) es una técnica estadística que analiza la relación de dos o más variables. Se utiliza para inferir datos a partir de otros, hacer predicciones y entender las interacciones más básicas entre las variables.

Esta permite analizar los cambios de una variable respuesta respecto a otras variables explicativas o predictores.

Se pueden encontrar varios tipos de regresión:

Regresión lineal simple

Explicamos una variable continua respecto a otra linealmente.

Regresión lineal múltiple (varias variables)

Explicamos una variable continua respecto a varias linealmente.

Regresión logística

Explicamos una variable binaria respecto a una o varias con una combinación lineal y no lineal.

Algunas ecuaciones regresión lineal simple son:

- **Regresión lineal**: y = A + Bx
- **Regresión potencial**: y = aX^b
- **Regresión logarítmica**: y = A + B Ln(x)
- **Regresión exponencial**: y = A exp(bx)
- **Regresión cuadrática**: y = A + Bx +Cx^2

El tipo de regresión univariable a escoger depende del gráfico de dispersión que observemos entre ambas variables.

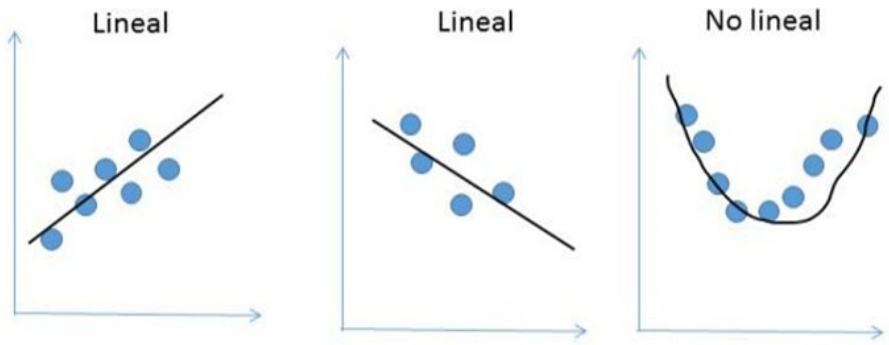

Imagen 4.3. Relaciones. *Fuente:* elaboración propia.

Si observamos que la relación es como en los dos primeros casos que aparecen en la figura de arriba, aplicaremos una regresión lineal y = A + Bx. Si se observa una nube de puntos como en el tercer caso, será más conveniente probar relaciones logarítmicas, exponenciales o cuadráticas.

Para elegir cuál es la mejor de las opciones, se transforma la variable X probando opciones como log(x), exp(x), x^2, etc. y se selecciona aquella que mejor R^2 da respecto a Y.

Salida del modelo de regresión

Cuando ejecutamos un modelo de regresión en R, la salida que obtenemos es la siguiente:

```
Call:
lm(formula = Y ~ X)

Coefficients:
(Intercept)            X
     1.3081       0.1156


Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7839 -2.6905 -0.9711  0.3180 11.3384

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.3081     4.2789   0.306    0.763
X             0.1156     0.1155   1.001    0.330

Residual standard error: 3.953 on 18 degrees of freedom
Multiple R-squared:  0.05276,   Adjusted R-squared:  0.0001333
F-statistic: 1.003 on 1 and 18 DF,  p-value: 0.33
```

Imagen 4.4. Salida de modelo de regresión.
Fuente: elaboración propia con R.

Describimos los detalles:

- **Y~X** indicado dentro de lm significa que queremos explicar linealmente la variable Y usando la variable X.
- **Residuals** indica los cuantiles de los residuos, así como su máximo y mínimo, siendo estos la diferencia entre el valor de la Y real y el valor que adjudica el modelo.
- **Coefficientes** indica los valores de los coeficientes (pendiente y término independiente), aportando su estimación, desviación típica, estadístico t de significación y un p-valor de un contraste t-student. Si el coeficiente se puede considerar no nulo este p-valor debe ser menor que 0.05.
- **Multiple R-squared** y **Adj R-squared** indican los valores de explicación de la variable objetivo que se consiguen en el modelo. Valores próximos a 0 apuntan a un modelo pobre y valores próximos a 1 a un buen modelo.
- **F-statistic** indica si el modelo es significativamente mejor que no tener ningún modelo. Si el p-valor es menor que 0.05 el modelo es usable.

Descarga: Consulta el notebook UD4 N02

Descárgate el archivo UD4_N02 en R y este csv. También puedes verlo en .html

V. Regresión lineal múltiple

En un modelo de regresión lineal múltiple, hay varios predictores para una única variable de respuesta:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon$$

La salida que proporciona R en su lm indica una serie de atributos que se asocian a los coeficientes: diversos contrastes de hipótesis sobre la significación de los coeficientes de las distintas variables.

Es parecida a la vista anteriormente. En realidad, la regresión múltiple se ha mostrado en el apartado anterior, ya que cuando se han creado nuevas variables en el caso polinómico, ya se estaba haciendo una regresión múltiple. La diferencia es que en el caso anterior se crearon variables basándose en un único predictor de los datos, y ahora disponemos de varios predictores en los datos. También se puede seguir realizando esa ingeniería de atributos, creando nuevos mediante las fórmulas de R al entrenar los modelos.

Para conocer la bondad del ajuste, se usa una serie de criterios objetivos que permiten valorar la capacidad de explicación de los datos.

La bondad del ajuste se mide con R^2 ajustado. Este es un coeficiente que tiene en cuenta la capacidad de explicación, pero también penaliza modelos muy complejos (con muchos parámetros o variables):

$$R^2_{ajustado} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-p)}{\sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)}$$

Para saber si se tiene un modelo significativamente distinto del constante, se hace un contraste F. Este es el que aparece al final de los sumarios de los modelos lineales:

$$\begin{aligned} H_0 &: \beta_{i_1} = \ldots = \beta_{i_r} = 0 \\ H_1 &: \text{existe algún j tal que } \beta_{i_j} \neq 0. \end{aligned}$$

Para saber si se puede prescindir por separado de cada variable predictora, tenemos los contrastes t-student con hipótesis alternativa que indica, para cada variable por separado, si el valor es significativamente no nulo:

$$H_1 : \beta_i \neq 0$$

Además de estos contrastes, se pueden analizar las distribuciones de los residuos de los modelos sobre el conjunto que modelan. Los residuos son los errores que comete el modelo al aproximar la variable objetivo en el conjunto donde se ha ajustado.

Idealmente, según se ha construido matemáticamente, el modelo debería tener una distribución de residuos centrada en 0 (es decir, el valor central es 0, que equivale a no equivocarse) y con valores de error repartidos simétricamente a ambos lados con decaimiento exponencial.

En la nomenclatura de las distribuciones de probabilidad que conocemos, corresponde a una distribución Normal(0, ε). El valor ε corresponde, en el argot matemático, a un valor pequeño desconocido.

De esta forma, si representamos un histograma de los residuos o una gráfica cuantil-cuantil o*Q-Q plot* enfrentada respecto a la normal relativa, los resultados de un modelo lineal correctamente diseñado deben ser parecidos a los siguientes:

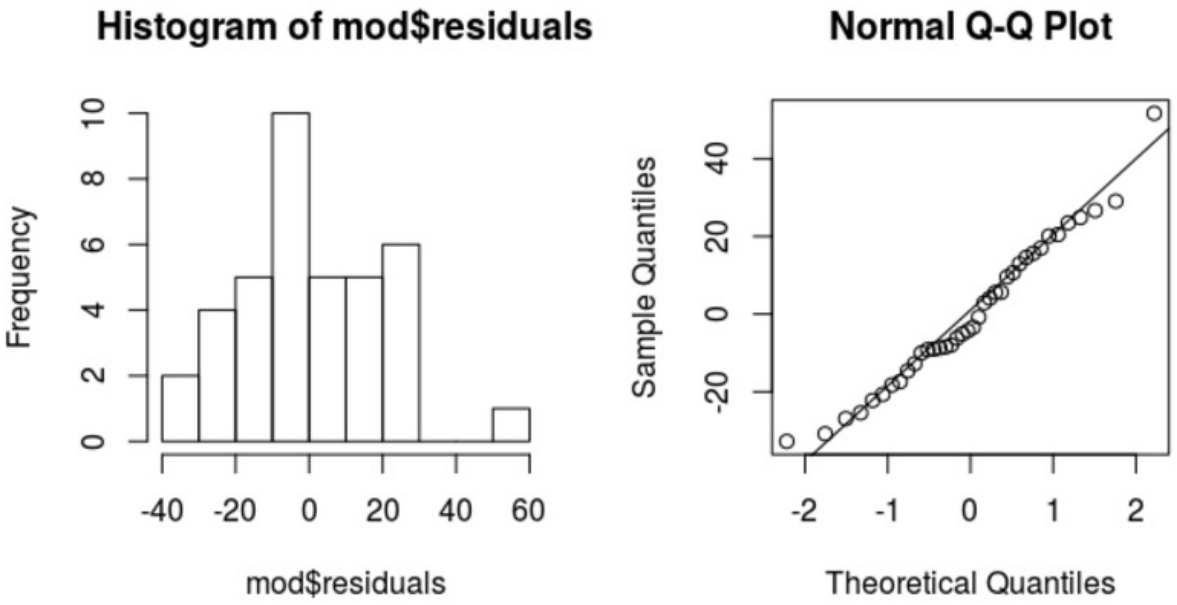

Imagen 4.5. Análisis de residuos. *Fuente:* elaboración propia con R.

Descarga: Consulta el notebook UD4 N03

Descárgate el archivo UD4_N03 y ejecútalo en R. También puedes verlo en.html

VI. Selección de modelos: equilibrio entre sesgo y varianza

Los modelos estadísticos que se ajustan sobre datos tienen dos características importantes:

Sesgo

Es el error de ajuste que tiene el modelo sobre los datos aportados. Está relacionado con los residuos. Cuantos más parámetros o variables tenga el modelo, este valor se acercará más a 0, es decir, mayor complejidad del modelo favorece el sesgo nulo, ya que se puede recoger más información de los datos en los parámetros.

Varianza

Es la variabilidad que el modelo tiene implícita. No depende de los datos sino exclusivamente de la estructura del mismo, fundamentalmente del número de parámetros. Cuanto más simple es un modelo (menos parámetros o variables), menor varianza tiene.

Estos conceptos se relacionan con la complejidad del modelo. Cuantos más atributos toma como predictores, el modelo lineal se hará más complejo. El nivel de complejidad se puede medir con el número de parámetros del modelo.

Lo que se pretende es tener el modelo perfecto con sesgo nulo y varianza nula. Sin embargo, la realidad es que son conceptos antagónicos, **mejorar el sesgo empeora la varianza y viceversa** De este modo, si se busca un modelo que explique "demasiado bien" los datos, reduciendo su sesgo al mínimo, el modelo también será más complejo, esto es, con mayor varianza.

En la imagen, se muestra un modelo con mucho sesgo y poca varianza y un modelo con sesgo nulo y mucha varianza:

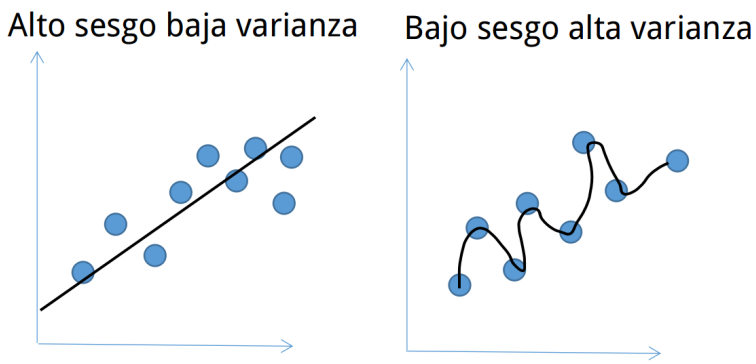

Imagen 4.6. Sesgo-varianza en modelo. *Fuente:* elaboración propia.

En la siguiente imagen (bias = sesgo y variance = varianza), la situación de arriba-izquierda es imposible y la de abajo-derecha es indeseable. La situación real es un balance entre la posición que aparece arriba-derecha y la de abajo-izquierda, esto es, entre modelos con sesgo nulo y varianza alta o sesgo alto y varianza nula:

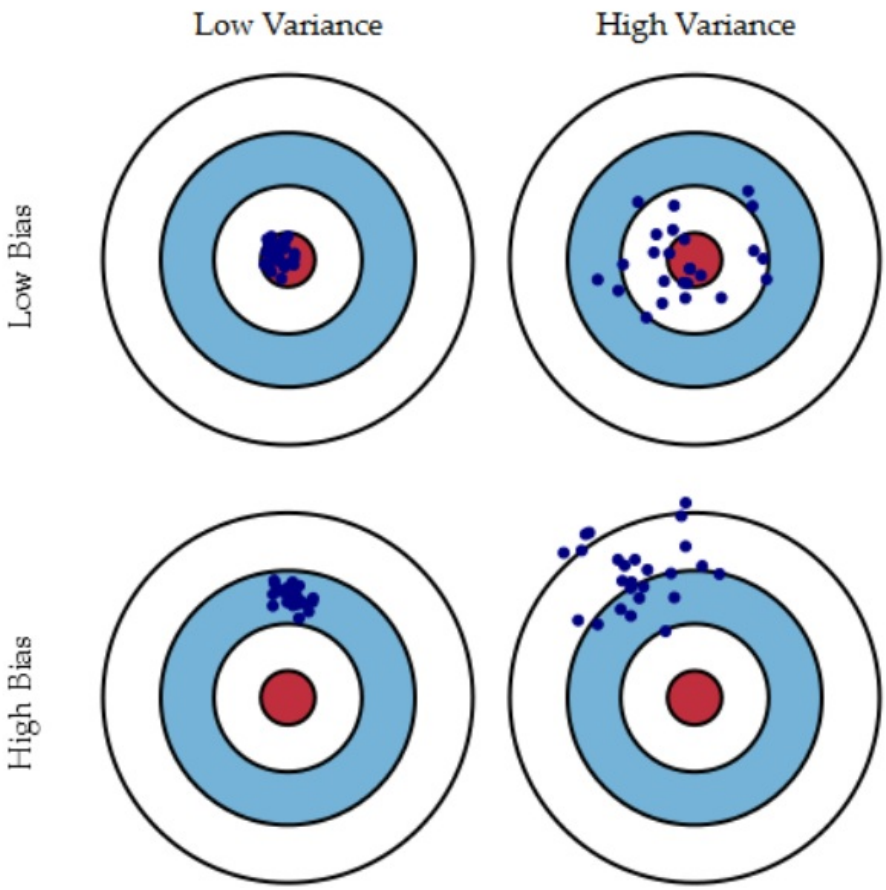

Imagen 4.7. Equilibrio entre sesgo y varianza. *Fuente:* Elements of Statistical Learning de Trevor Hastie.

VII. Criterios estadísticos de selección de modelos

A la hora de seleccionar modelos lineales que se ajusten lo mejor posible a la realidad de los datos que tenemos, es necesario establecer criterios estadísticos objetivos para decidir. Estos criterios tienen en cuenta el equilibrio mencionado en el apartado anterior: optimizar en función del mismo es equivalente a seleccionar el mejor modelo.

Existen criterios relativos y criterios absolutos.

Criterios relativos:

- **R² ajustado**: capacidad de explicación de la variabilidad de la variable respuesta respecto a los predictores usados, penalizando la complejidad del modelo. De entre todos los modelos entrenados en un conjunto de datos se debe seleccionar el que tenga mayor R² ajustado.
- **Contraste F**: comparamos con un ANOVA cuál de dos modelos es mejor.

Criterios absolutos:

Los criterios AIC y BIC están basados en la verosimilitud del modelo, penalizando, a su vez, la complejidad del mismo. Son criterios absolutos porque se usa como puntuación el valor de máxima verosimilitud:

- **AIC (Akaike Information Criterion)**: este criterio es el que mejor modela una realidad con alta dimensionalidad. Debe seleccionarse el que tenga menor valor.

- **BIC (Bayesian Information Criterion)**: el verdadero modelo. Debe seleccionarse el que tenga menor valor. BIC penaliza más que AIC la complejidad de parámetros.

Descarga: Consulta el notebook UD4 N04

Descárgate el archivo UD4_N04 y ejecútalo en R. También puedes verlo en.html

VIII. Resumen final

En esta unidad, se ha desarrollado el contenido central del aprendizaje estadístico exponiendo cómo se crean los modelos más sencillos: los basados en interacciones lineales. Estos suponen una primera aproximación —económica computacionalmente— comprensible y capaz de explicar los datos. Por lo tanto, es recomendable aplicar siempre este tipo de modelos en el inicio de un análisis.

Además, se ha mostrado que crear un modelo de mayor complejidad no implica necesariamente que sea mejor, ya que existe el equilibrio entre sesgo y varianza.

Finalmente, se han estudiado los criterios para seleccionar modelos que penalizan esa complejidad o cantidad de parámetros.

Ejercicios

Caso práctico

Como repaso del tema y preparación para el Caso práctico final, se presenta el siguiente caso práctico.

Descárgate el archivo ACTIVIDAD4_UD4 en R y la csv del caso. También puedes verlo en .html

Cuando lo hayas realizado, puedes descargar su solución y comprobar tus resultados.

Solución

En los siguientes archivos dispones de la solución de la actividad propuesta:

- Solución en .html
- Solución en R.

Recursos

Glosario.

- **Correlación**: Valor de dependencia lineal entre dos variables. Su rango es de -1 a 1.
- **Criterios de selección estadísticos**: Son métricas asociadas a un modelo y un conjunto de datos que evalúan la capacidad explicativa del modelo respecto a los mismos.
- **Modelo de regresión lineal**: Modelo que explica una variable objetivo a través de operaciones de multiplicación de coeficientes y suma de otras variables.
- **Sesgo**: Error que comete el modelo al realizar predicciones sobre los datos sobre los que ha entrenado.
- **Varianza**: Valor de variación que tiene un modelo relacionado con la complejidad o cantidad de parámetros del mismo.