Regresión

Jocelyn Dunstan Escudero

jdunstan@uc.cl

Departamento de Ciencia de la Computación
& Instituto de Matemática Computacional
Pontificia Universidad Católica de Chile

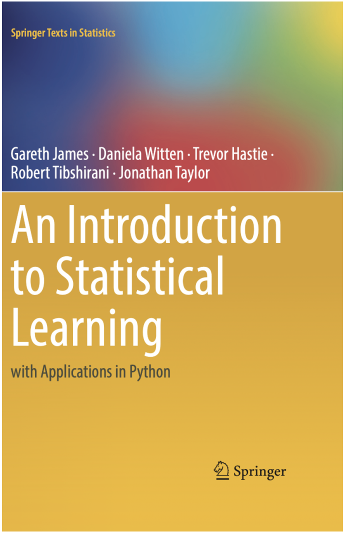
Santiago, Chile

20 de octubre de 2025

Objetivos

- Comprender el modelo de regresión lineal, en partícular el método de los mínimos cuadrados
- Evaluar el desempeño de una regresión lineal
- Entender como modelos vistos anteriormente se pueden extender para regresión
- *Hands-on* usando el código del libro Introduction to statistical learning.

Libro MEGA recomendado



`https://www.statlearning.com/`

Idea general

- Las variables pueden ser categóricas o continuas.

Idea general

- Las variables pueden ser categóricas o continuas.

- En la regresión nos interesa predecir un resultado cuantitativo

Idea general

- Las variables pueden ser categóricas o continuas.
- En la regresión nos interesa predecir un resultado cuantitativo
- La calidad de la predicción será una medida de la distancia entre el valor real y la predicción.

Idea general

- Las variables pueden ser categóricas o continuas.
- En la regresión nos interesa predecir un resultado cuantitativo
- La calidad de la predicción será una medida de la distancia entre el valor real y la predicción.
- La idea de los conjuntos de entrenamiento y prueba también es válida para la regresión.

Regresión Lineal

Regresión Lineal

- Es el método de regresión más sencillo y conocido.

Regresión Lineal

- Es el método de regresión más sencillo y conocido.
- Es un método paramétrico: Supone que los predictores y el resultado interactúan linealmente.

Regresión Lineal

- Es el método de regresión más sencillo y conocido.
- Es un método paramétrico: Supone que los predictores y el resultado interactúan linealmente.
- Es muy utilizado por su fácil interpretación.

https://www.statlearning.com/

Regresión lineal simple usando un predictor

$$\hat{Y} = \beta_0 + \beta_1 X + \epsilon$$

- β_0 es el intercepto y β_1 es la pendiente.

Regresión lineal simple usando un predictor

$$\hat{Y} = \beta_0 + \beta_1 X + \epsilon$$

- β_0 es el intercepto y β_1 es la pendiente.

- Ambos se obtienen del conjunto de entrenamiento minimización de la suma de cuadrados de los residuos. Los residuos son $y_i - \hat{y}_i$

Regresión lineal simple usando un predictor

$$\hat{Y} = \beta_0 + \beta_1 X + \epsilon$$

- β_0 es el intercepto y β_1 es la pendiente.
- Ambos se obtienen del conjunto de entrenamiento minimización de la suma de cuadrados de los residuos. Los residuos son $y_i - \hat{y}_i$
- El símbolo del sombrero significa "valor estimado"

Regresión lineal simple usando un predictor

$$\hat{Y} = \beta_0 + \beta_1 X + \epsilon$$

- β_0 es el intercepto y β_1 es la pendiente.
- Ambos se obtienen del conjunto de entrenamiento minimización de la suma de cuadrados de los residuos. Los residuos son $y_i - \hat{y}_i$
- El símbolo del sombrero significa "valor estimado"
- ϵ es un término de error.

Mínimos Cuadrados

- El método de mınimos cuadrados ordinarios (OLS en inglés) se usa para estimar los coeficientes minimizando la suma de los errores cuadráticos (SSE) de los datos observados.

Mínimos Cuadrados

- El método de mınimos cuadrados ordinarios (OLS en inglés) se usa para estimar los coeficientes minimizando la suma de los errores cuadráticos (SSE) de los datos observados.

- Supongamos que tenemos m observaciones de y y de x, calculamos la suma de los errores cuadráticos (SSE) o E de error de la siguiente forma:

$$E = \sum_{i=1}^{m} (y_i - h(x_i))^2 = \sum_{i=1}^{m} (y_i - \beta_0 - \beta_1 x_i)^2$$

Mínimos Cuadrados

- Para encontrar los parámetros que minimizan el error calculamos las derivadas parciales de SSE respecto a β_0 y β_1 . Luego igualamos las derivadas a cero y resolvemos la ecuación para despejar los parámetros.

$$\frac{\partial E}{\partial \beta_0} = -2\sum_{i=1}^{m}(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial E}{\partial \beta_1} = -2\sum_{i=1}^{m}(y_i - \beta_0 - \beta_1 x_i)x_i = 0$$

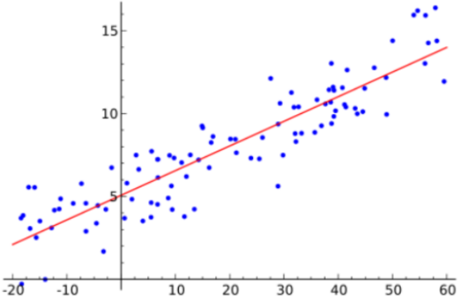
Se puede demostrar que:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

donde

$$\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{and} \quad \bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$$

- El modelo ajustado representa la recta del mínimo error cuadrático.



Supuestos de la RL

- Exogeneidad débil (las variables predictoras no tienen errores)
- Linealidad (la respuesta es una combinación lineal de los predictores)
- Varianza constante u homocedasticidad (los residuos son constantes para diferentes valores predichos)
- La colinealidad entre los predictores dificulta la interpretación de β.

¿Cuántas veces son estos supuestos válidos?

Supuestos de la RL

RESEARCH AND PRACTICE

Relationship of Soft Drink Consumption to Global Overweight, Obesity, and Diabetes: A Cross-National Analysis of 75 Countries

Sanjay Basu, MD, PhD, Martin McKee, MD, DSc, Gauden Galea, MD, and David Stuckler, PhD, MPH

$$
\begin{aligned}
\mathrm{OBESE}_i = {} & \mathrm{BSODA}_i + \mathrm{BCEREAL}_i \\
& + \mathrm{BFRUITVEG}_i + \mathrm{BMEAT}_i \\
& + \mathrm{BOIL}_i + \mathrm{BTOTAL}_i \\
& + \mathrm{BELDER}_i + \mathrm{BGDP}_i \\
& + \mathrm{BURBAN}_i + \mathrm{BWATER}_i + \varepsilon
\end{aligned}
$$

Exactitud de los modelos

Desempeño

- El error absoluto medio (MAE) y el error cuadrático medio (RMSE) se definen como sigue:

$$MSE = \frac{1}{n}\sum_{j=1}^{n}|y_i - \hat{y}_j|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_i - \hat{y}_j)^2}$$

- Se diferencian en la penalización del residuo y en el cálculo del valor absoluto.

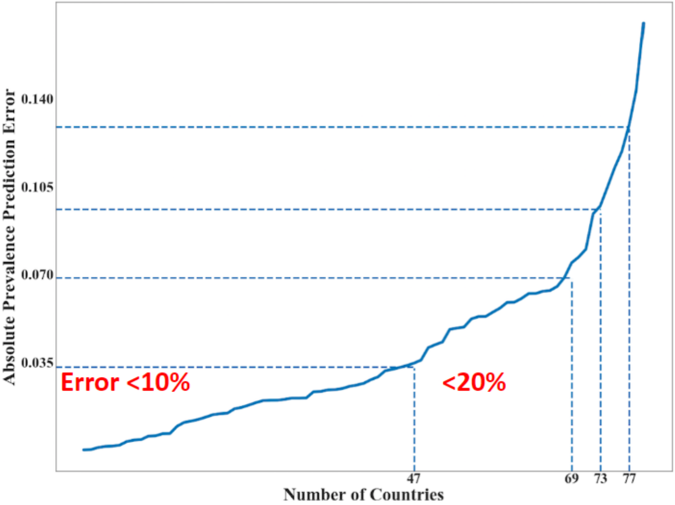
Desempeño

- La interpretación de los errores es más difícil en la regresión que en la clasificación. Entonces, puede ser útil evaluar los errores en comparación con el rango de valores.

Desempeño

- La interpretación de los errores es más difícil en la regresión que en la clasificación. Entonces, puede ser útil evaluar los errores en comparación con el rango de valores.

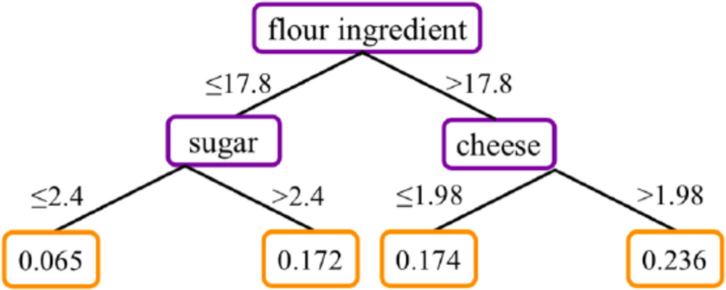
- En el siguiente ejemplo, el error de predicción de la prevalencia absoluta (APPEs) se evaluó mediante factores de 0,35, que es la prevalencia máxima



https://doi.org/10.1177/1460458219845959

Extensión de los modelos antes vistos al caso de regresión

Random Forest

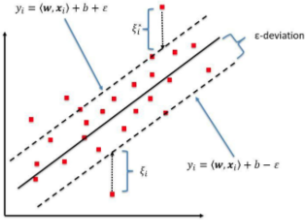


Support vector regression

- Si la regresión lineal funciona minimizando las tasas de error, SVR trata de ajustar el error en un determinado umbral ϵ.

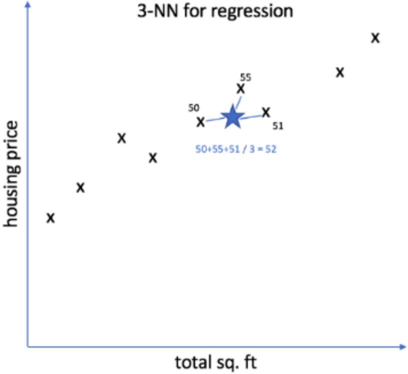
- Así, en lugar de tener la función g(x) que toma +1 o -1, la acotamos por ϵ.

$$\varepsilon \leq y - \boldsymbol{w}\,\boldsymbol{x} - w0 \leq \varepsilon$$



https://medium.com/coinmonks/support-vector-regression-or-svr-8eb3acf6d0ff

K-nearest neighbors



https://www.jeremyjordan.me/k-nearest-neighbors/

Actividad práctica

Colab