Regresión lineal: el enfoque estadístico con máxima verosimilitud

Matías Carrasco

4 de septiembre de 2023

Índice

1. Modelo lineal gaussiano

El i-ésimo target y_i se genera de la forma $\boldsymbol{\theta}^\top \boldsymbol{x}_i$ más una cantidad aleatoria ϵ_i:

$$y_i = \boldsymbol{\theta}^\top \boldsymbol{x}_i + \epsilon_i$$

Recordar que la ecuación $y_i = \boldsymbol{\theta}^\top \boldsymbol{x}_i + \epsilon_i$ en forma matricial con los vectores $\boldsymbol{\theta}$ y $\boldsymbol{x}_i$ dados por

$$\boldsymbol{\theta} = \begin{bmatrix} b \\ w_1 \\ \vdots \\ w_D \end{bmatrix} \quad \boldsymbol{x}_i = \begin{bmatrix} 1 \\ x_i^{(1)} \\ \vdots \\ x_i^{(D)} \end{bmatrix}$$

representa

$$y_i = \begin{bmatrix} b & w_1 & \dots & w_D \end{bmatrix} \begin{bmatrix} 1 \\ x_i^{(1)} \\ \vdots \\ x_i^{(D)} \end{bmatrix} + \epsilon_i = b + \sum_{j=1}^{D} w_j x_i^{(j)} + \epsilon_i$$

La diferencia entre el modelo y el valor real del target, la ϵ_i, es una variable aleatoria continua. Las asumimos independientes entre sí:

$$p\left(\epsilon_1, \dots, \epsilon_i\right) = \prod_{i=1}^{N} p\left(\epsilon_i\right)$$

Para la forma de $p\left(\epsilon_i\right)$ asumiremos que se trata de una distribución gaussiana (o normal) con media cero y varianza σ^2. Esto permite que ϵ_i sea tanto positivo como negativo (permite que los datos se encuentren tanto por encima como por debajo de la línea $\boldsymbol{\theta}^\top \boldsymbol{x}$) y tiene interesantes propiedades de modelado que lo vinculan a la función de pérdida cuadrática.

Nuestro modelo consta entonces de dos componentes:

1. Un componente determinístico $\boldsymbol{\theta}^\top \boldsymbol{x}_i$.
2. Un componente aleatorio ϵ_i, a veces referido como ruido.

Existen alternativas al ruido gaussiano y también al ruido aditivo. Por ejemplo, para algunas aplicaciones, un término multiplicativo podría ser más apropiado (en cuyo caso, $y = f(\boldsymbol{x}; \boldsymbol{\theta})\epsilon$). Sin embargo, el ruido aditivo gaussiano nos permite obtener expresiones exactas para el valor óptimo del parámetro $\widehat{\boldsymbol{\theta}}$.

2. Verosimilitud

Nuestro modelo tiene la siguiente forma:

$$y_i = f\left(\boldsymbol{x}_i; \boldsymbol{\theta}\right) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$$

Necesitamos encontrar el valor óptimo $\widehat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$. También tenemos un parámetro adicional σ^2.

La salida del modelo, y, es ahora en sí misma una variable aleatoria. En particular, no hay un único valor de y_i para un valor particular de $\boldsymbol{x}_i$. La variable aleatoria y_i tiene la siguiente función de densidad:

$$p\left(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}, \sigma^2\right) = \mathcal{N}\left(\boldsymbol{\theta}^\top \boldsymbol{x}_i, \sigma^2\right)$$

Notar la condicional en el lado izquierdo - la densidad de y_i depende de valores particulares de $\boldsymbol{x}_i$ y $\boldsymbol{\theta}$ (determinan la media) y σ^2 (la varianza).

3. Verosimilitud del conjunto de datos

Si tenemos N datos, estamos interesados en la densidad conjunta condicional:

$$p\left(y_1, \ldots, y_N \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{\theta}, \sigma^2\right)$$

Esta es una densidad conjunta sobre todos los targets en nuestro conjunto de datos. Lo escribiremos de manera compacta (usando notación vectorial y la matriz de diseño $\boldsymbol{x}$) como

$$p\left(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}, \sigma^2\right).$$

Evaluar esta densidad en los datos observados da un valor de su verosimilitud, el cual podemos optimizar variando $\boldsymbol{\theta}$ y σ^2.

La suposición de independencia de los ruidos nos permite factorizar esta densidad en N términos separados, uno para cada dato:

$$L = p\left(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}, \sigma^2\right) = \prod_{i=1}^{N} p\left(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}, \sigma^2\right) = \prod_{i=1}^{N} \mathcal{N}\left(\boldsymbol{\theta}^\top \boldsymbol{x}_i, \sigma^2\right)$$

Esto no quiere decir que los valores y_i son ellos mismos completamente independientes. Este no es el caso - los valores y_i, en promedio, varían con el $\boldsymbol{x}_i$, sugiriendo una clara dependencia entre ellos.

> De hecho, son condicionalmente independientes - dado un valor para $\boldsymbol{\theta}$ (la parte determinista del modelo), los y_i son independientes.

Esta dependencia está encapsulada en el parámetro $\boldsymbol{\theta}$. Si conocemos $\boldsymbol{\theta}$, todo lo que queda son los errores entre los datos observados y $\boldsymbol{\theta}^\top \boldsymbol{x}_i$. Estos errores se asumen independientes. Por lo tanto, condicionado a $\boldsymbol{\theta}$, las observaciones son independientes.

Ahora mostraremos cómo podemos encontrar los valores de $\boldsymbol{\theta}$ y σ^2 que maximizan la verosimilitud.

4. Máxima verosimilitud

La verosimilitud nos da un único valor que nos indica cuán probable es nuestro conjunto de datos, dado el modelo actual (por modelo, nos referimos a la elección de $\boldsymbol{\theta}$ y σ^2). Como nuestro conjunto de datos es fijo, variar el modelo resultará en diferentes valores de verosimilitud. Una elección sensata de modelo sería aquella que maximice la verosimilitud. En otras palabras, seleccionaremos los parámetros del modelo que hagan nuestras observaciones más probables.

Sustituyendo la expresión para la función de densidad gaussiana:

$$\begin{aligned}
\ln L &= \sum_{i=1}^{N} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(y_i - f\left(\boldsymbol{x}_i;\boldsymbol{\theta}\right)\right)^2\right\}\right) \\
&= \sum_{i=1}^{N}\left(-\frac{1}{2}\ln(2\pi) - \ln\sigma - \frac{1}{2\sigma^2}\left(y_i - f\left(\boldsymbol{x}_i,\boldsymbol{\theta}\right)\right)^2\right) \\
&= -\frac{N}{2}\ln 2\pi - N\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - f\left(\boldsymbol{x}_i;\boldsymbol{\theta}\right)\right)^2
\end{aligned}$$

Sustituyendo $f\left(\boldsymbol{x}_i;\boldsymbol{\theta}\right) = \boldsymbol{\theta}^\top \boldsymbol{x}_i$:

$$\ln L = -\frac{N}{2}\ln 2\pi - N\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - \boldsymbol{\theta}^\top \boldsymbol{x}_i\right)^2$$

5. Derivada respecto a $\boldsymbol{\theta}$

Podemos encontrar los parámetros óptimos tomando derivadas e igualándolas a cero. Para $\boldsymbol{\theta}$:

$$\frac{\partial \ln L}{\partial \boldsymbol{\theta}} = \frac{1}{\sigma^2}\frac{\partial}{\partial \boldsymbol{\theta}}\left[\sum_{i=1}^{N}\left(y_i - \boldsymbol{\theta}^\top \boldsymbol{x}_i\right)^2\right] = \mathbf{0}$$

Notar que $\frac{\partial \ln L}{\partial \boldsymbol{\theta}}$ es un vector, por lo que lo igualamos a $\mathbf{0}$, un vector de ceros del mismo tamaño.

La solución de máxima verosimilitud para $\boldsymbol{\theta}$ coincide por lo tanto con la solución que ya hemos derivado para el caso de mínimos cuadrados. Minimizar la pérdida cuadrática es equivalente a la solución de máxima verosimilitud si se asume que el ruido es gaussiano.:

$$\widehat{\boldsymbol{\theta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top \boldsymbol{y}$$

Recordar la notación matricial que utilizamos antes:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_N^\top \end{bmatrix} = \begin{bmatrix} 1 & x_1^{(1)} & \cdots & x_1^{(D)} \\ 1 & x_2^{(1)} & \cdots & x_2^{(D)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_N^{(1)} & \cdots & x_N^{(D)} \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

6. Derivada respecto a σ

Para obtener una expresión para σ^2 (asumiendo que $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$) podemos seguir el mismo procedimiento. Tomando derivadas parciales e igualando a cero obtenemos:

$$\frac{\partial L}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{N} \left(y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\theta}} \right)^2 = 0.$$

Reorganizando, obtenemos $\widehat{\sigma}^2$, la estimación de máxima verosimilitud para σ^2:

$$\widehat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\theta}} \right)^2$$

La varianza es simplemente el error cuadrático medio.

En notación matricial, usando el hecho de que

$$\sum_{i=1}^{N} \left(y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\theta}} \right)^2 = (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}})^\top (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}})$$

tenemos:

$$\begin{aligned} \widehat{\sigma}^2 &= \frac{1}{N} (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}})^\top (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}) \\ &= \frac{1}{N} \left(\boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{y}^\top \boldsymbol{X}\widehat{\boldsymbol{\theta}} + \widehat{\boldsymbol{\theta}}^\top \boldsymbol{X}^\top \boldsymbol{X}\widehat{\boldsymbol{\theta}} \right) \end{aligned}$$

Sustituyendo $\widehat{\boldsymbol{\theta}} = \left(\boldsymbol{X}^\top \boldsymbol{X} \right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ y observando que

$$\widehat{\boldsymbol{\theta}}^\top = \boldsymbol{y}^\top \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X} \right)^{-1}$$

pues $\left(\boldsymbol{X}^\top \boldsymbol{X} \right)^{-1}$ es simétrica:

$$\begin{aligned} \widehat{\sigma}^2 &= \frac{1}{N} \left(\boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{y}^\top \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X} \right)^{-1} \boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{y}^\top \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X} \right)^{-1} \boldsymbol{X}^\top \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X} \right)^{-1} \boldsymbol{X}^\top \boldsymbol{y} \right) \\ &= \frac{1}{N} \left(\boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X} \right)^{-1} \boldsymbol{X}^\top \boldsymbol{y} \right) \\ &= \frac{1}{N} \left(\boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{X}\widehat{\boldsymbol{\theta}} \right) \end{aligned}$$

7. Modelos más complejos y sobreajuste

Sustituyendo la expresión para $\widehat{\sigma}^2$ en la expresión de la log-verosimilitud nos da el valor de la log-verosimilitud en el máximo:

$$\begin{aligned}\ln L &= -\frac{N}{2}\ln 2\pi - \frac{N}{2}\ln \widehat{\sigma}^2 - \frac{1}{2\widehat{\sigma}^2}N\widehat{\sigma}^2 \\ &= -\frac{N}{2}(1+\ln 2\pi) - \frac{N}{2}\ln \widehat{\sigma}^2.\end{aligned}$$

Esto nos indica que el valor máximo de L seguirá aumentando a medida que disminuyamos $\widehat{\sigma^2}$.

La varianza del ruido incorporado en el modelo para capturar efectos que la parte determinista de nuestro modelo (es decir, $f(\boldsymbol{x};\boldsymbol{\theta})$) no puede. Una forma de disminuir σ^2 es modificar $f(\boldsymbol{x};\boldsymbol{\theta})$ de manera que pueda capturar más de la variabilidad en los datos, es decir, hacerlo más flexible. El riesgo es caer en sobreajuste.

8. Incertidumbre en las estimaciones de los parámetros

El valor que obtenemos para $\widehat{\boldsymbol{\theta}}$ está fuertemente influenciado por los valores particulares del ruido en los datos. Queremos saber cuánta incertidumbre hay en $\widehat{\boldsymbol{\theta}}$.

Recordar lo que $\boldsymbol{\theta}$ y $\widehat{\boldsymbol{\theta}}$ significan. Hemos hipotetizado un modelo responsable de generar los datos:

$$y_i = \boldsymbol{\theta}^\top \boldsymbol{x}_i + \epsilon_i$$

donde $\boldsymbol{\theta}$ representa el verdadero valor de los parámetros y ϵ_i es una variable aleatoria con distribución normal. Entonces la distribución generadora de targets

$$p\left(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\theta}, \sigma^2\right)$$

es un producto de densidades normales:

$$p\left(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\theta}, \sigma^2\right) = \prod_{i=1}^{N} p\left(y_i \mid \boldsymbol{x}_i\boldsymbol{\theta}, \sigma^2\right) = \prod_{i=1}^{N} \mathcal{N}\left(\boldsymbol{\theta}^\top \boldsymbol{x}_i, \sigma^2\right)$$

Un producto de densidades Gaussianas univariadas puede escribirse como una densidad Gaussiana multivariada con una covarianza diagonal:

$$p\left(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\theta}, \sigma^2\right) = \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\theta}, \sigma^2\mathbf{I}\right).$$

Ahora, $\widehat{\boldsymbol{\theta}}$ es una estimación del verdadero valor del parámetro $\boldsymbol{\theta}$. Calculando la esperanza de $\widehat{\boldsymbol{\theta}}$ con respecto a la distribución generadora nos dirá qué esperamos que sea $\widehat{\boldsymbol{\theta}}$, en promedio:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\{\widehat{\boldsymbol{\theta}}\} &= \left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\{\boldsymbol{y}\} \\ &= \left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\theta} \\ &= \boldsymbol{\theta}\end{aligned}$$

Este resultado nos indica que el valor esperado de nuestra aproximación $\widehat{\boldsymbol{\theta}}$ es el verdadero valor del parámetro. Esto significa que nuestro estimador no tiene sesgo.

La variabilidad en la estimación de $\widehat{\boldsymbol{\theta}}$ se obtiende de su matriz de covarianza. Esta matriz de covarianza nos proporciona dos piezas de información útiles:

- Las entradas diagonales son las varianzas de los coeficientes individuales de $\widehat{\boldsymbol{\theta}}$, nos indican cuánta variabilidad podríamos esperar en los parámetros individuales, es decir, qué tan bien están definidos por los datos.
- Las entradas fuera de la diagonal nos indican cómo covarían los parámetros. Si los valores son altos y positivos, nos dice que aumentar uno requerirá un aumento en el otro para mantener un buen modelo. Valores negativos grandes nos dicen lo contrario: aumentar uno causará una disminución en el otro. Valores cercanos a cero nos indican que los parámetros no dependen uno del otro.

La matriz de covarianza es:

$$\begin{aligned}\mathrm{cov}\{\widehat{\boldsymbol{\theta}}\} &= \mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\left\{\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top\right\} - \mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\{\widehat{\boldsymbol{\theta}}\}\mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\{\widehat{\boldsymbol{\theta}}\}^\top \\ &= \mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\left\{\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top\right\} - \boldsymbol{\theta}\boldsymbol{\theta}^\top\end{aligned}$$

donde hemos usado la esperanza de $\widehat{\boldsymbol{\theta}}$ que derivamos anteriormente. El primer término es

$$\begin{aligned}\mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\left\{\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top\right\} &= \mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\left\{\left(\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y}\right)\left(\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y}\right)^\top\right\} \\ &= \left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\left\{\boldsymbol{y}\boldsymbol{y}^\top\right\}\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\end{aligned}$$

La covarianza de $\boldsymbol{y}$ es, por definición, $\sigma^2\mathbf{I}$ y su media es $\mathbf{X}\boldsymbol{\theta}$. Luego tenemos:

$$\mathrm{cov}\{\boldsymbol{y}\} = \sigma^2\mathbf{I} = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\left\{\mathbf{y}\mathbf{y}^\top\right\} - \mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\{\boldsymbol{y}\}\mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\{\boldsymbol{y}\}^\top$$

Por lo tanto, podemos reorganizar esta expresión para obtener:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\left\{\boldsymbol{y}\boldsymbol{y}^\top\right\} &= \mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\{\boldsymbol{y}\}\mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\{\boldsymbol{y}\}^\top + \sigma^2\mathbf{I} \\ &= \boldsymbol{X}\boldsymbol{\theta}(\boldsymbol{X}\boldsymbol{\theta})^\top + \sigma^2\mathbf{I} \\ &= \boldsymbol{X}\boldsymbol{\theta}\boldsymbol{\theta}^\top\boldsymbol{X}^\top + \sigma^2\mathbf{I}\end{aligned}$$

Finalmente obtenemos:

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\left\{\boldsymbol{\theta}\boldsymbol{\theta}^\top\right\} = \boldsymbol{\theta}\boldsymbol{\theta}^\top + \sigma^2\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}$$

Obtenemos así la expresión para la covarianza de $\widehat{\boldsymbol{\theta}}$:

$$\begin{aligned}\mathrm{cov}\{\widehat{\boldsymbol{\theta}}\} &= \boldsymbol{\theta}\boldsymbol{\theta}^\top + \sigma^2\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1} - \boldsymbol{\theta}\boldsymbol{\theta}^\top \\ &= \sigma^2\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\end{aligned}$$

Para $\widehat{\sigma}^2$ los cálculos son un poco más engorrosos, pero el resultado es:

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\left\{\widehat{\sigma^2}\right\} = \sigma^2\left(1 - \frac{D}{N}\right)$$

donde D es el número de atributos (el número de columnas de $\boldsymbol{X}$).

Asumiendo que $D < N$, es decir, el número de atributos que medimos para cada observación es menor que el número de datos, entonces nuestra estimación de la varianza será, en promedio, menor que la varianza real:

$$\mathrm{E}_{\boldsymbol{y}|\boldsymbol{X}}\left\{\widehat{\sigma^2}\right\} < \sigma^2.$$

A diferencia de $\widehat{\boldsymbol{\theta}}$, este estimador es sesgado.

9. Variabilidad en las predicciones

Además de obtener una indicación de la variabilidad de nuestra estimación del parámetro, $\widehat{\boldsymbol{\theta}}$, tiene interés proporcionar indicaciones de cualquier variabilidad o incertidumbre en nuestras predicciones.

Supongamos que observamos un nuevo conjunto de atributos, $\boldsymbol{x}_{\text{nuevo}}$. Nos gustaría predecir la salida y_{nuevo} y además, la variabilidad asociada con esta salida, σ^2_{nuevo}.

Para predecir y_{nuevo}, multiplicamos $\boldsymbol{x}_{\text{nuevo}}$ por $\widehat{\boldsymbol{\theta}}$: $y_{\text{nuevo}} = \widehat{\boldsymbol{\theta}}^\top\boldsymbol{x}_{\text{nuevo}}$.

Para empezar podemos calcular su esperanza:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\left\{y_{\text{nuevo}}\right\} &= \mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\{\widehat{\boldsymbol{\theta}}\}^{\top}\boldsymbol{x}_{\text{nuevo}} \\ &= \boldsymbol{\theta}^{\top}\boldsymbol{x}_{\text{nuevo}}.\end{aligned}$$

El valor esperado de nuestra predicción es el nuevo atributo de datos multiplicado por el verdadero $\boldsymbol{\theta}$.

La varianza es:

$$\sigma^2_{\text{nuevo}} = \text{var}\left\{y_{\text{nuevo}}\right\} = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\left\{y^2_{\text{nuevo}}\right\} - \left(\mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\left\{y_{\text{nuevo}}\right\}\right)^2$$

Al sustituir $y_{\text{nuevo}} = \widehat{\boldsymbol{\theta}}^{\top}\boldsymbol{x}_{\text{nuevo}}$:

$$\begin{aligned}\text{var}\left\{y_{\text{nuevo}}\right\} &= \mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\left\{\left(\widehat{\boldsymbol{\theta}}^{\top}\boldsymbol{x}_{\text{nuevo}}\right)^2\right\} - \left(\boldsymbol{\theta}^{\top}\boldsymbol{x}_{\text{nuevo}}\right)^2 \\ &= \mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\left\{\boldsymbol{x}^{\top}_{\text{nuevo}}\widehat{\boldsymbol{\theta}}^{\top}\widehat{\boldsymbol{\theta}}\boldsymbol{x}_{\text{nuevo}}\right\} - \boldsymbol{x}^{\top}_{\text{nuevo}}\boldsymbol{\theta}\boldsymbol{\theta}^{\top}\boldsymbol{x}_{\text{nuevo}}.\end{aligned}$$

Sustituyendo la expresión para $\widehat{\boldsymbol{\theta}}$:

$$\text{var}\left\{t_{\text{nuevo}}\right\} = \sigma^2\boldsymbol{x}^{\top}_{\text{nuevo}}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_{\text{nuevo}}$$

Notar que al sustituir nuestra expresión para $\text{cov}\{\widehat{\boldsymbol{\theta}}\}$ la varianza puede reescribirse como

$$\sigma^2_{\text{nuevo}} = \boldsymbol{x}^{\top}_{\text{nuevo}}\,\text{cov}\{\widehat{\boldsymbol{\theta}}\}\boldsymbol{x}_{\text{nuevo}}.$$

Para resumir, nuestra predicción y la varianza asociada:

$$\begin{aligned}y_{\text{nuevo}} &= \boldsymbol{x}^{\top}_{\text{nuevo}}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\boldsymbol{x}^{\top}\boldsymbol{y} = \boldsymbol{x}^{\top}_{\text{nuevo}}\widehat{\boldsymbol{\theta}} \\ \sigma^2_{\text{nuevo}} &= \sigma^2\boldsymbol{x}^{\top}_{\text{nuevo}}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_{\text{nuevo}}.\end{aligned}$$

σ^2 es la verdadera varianza del ruido del conjunto de datos. En su lugar, podemos usar nuestra estimación, $\widehat{\sigma^2}$.

10. Varianza de y_{nuevo} en el caso univariado

Recordar que en el caso univariado la matriz $\boldsymbol{X}^{\top}\boldsymbol{X}$ es:

$$\boldsymbol{X}^{\top}\boldsymbol{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_N \end{bmatrix}\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} = \begin{bmatrix} N & \sum_{i=1}^N x_i \\ \sum_{i=1}^N x_i & \sum_{i=1}^N x_i^2 \end{bmatrix}$$

y la inversa de la matriz es:

$$\begin{aligned}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} &= \frac{1}{N\sum_{i=1}^N x_i^2 - \left(\sum_{i=1}^N x_i\right)^2}\begin{bmatrix} \sum_{i=1}^N x_i^2 & -\sum_{i=1}^N x_i \\ -\sum_{i=1}^N x_i & N \end{bmatrix} \\ &= \frac{1}{\overline{x^2} - \bar{x}^2}\begin{bmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}\end{aligned}$$

Luego, la varianza de la predicción:

$$\begin{aligned}\sigma^2_{\text{nuevo}} &= \frac{\sigma^2}{s_x^2}[1, x_{\text{nuevo}}]\begin{bmatrix} s_x^2 + \bar{x}^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}\begin{bmatrix} 1 \\ x_{\text{nuevo}} \end{bmatrix} \\ &= \frac{\sigma^2}{s_x^2}[1, x_{\text{nuevo}}]\begin{bmatrix} s_x^2 + \bar{x}^2 - \bar{x}x_{\text{nuevo}} \\ x_{\text{nuevo}} - \bar{x} \end{bmatrix} \\ &= \frac{\sigma^2}{s_x^2}\left(s_x^2 + (x_{\text{nuevo}} - \bar{x})^2\right) \\ &= \sigma^2\left(1 + \frac{(x_{\text{nuevo}} - \bar{x})^2}{s_x^2}\right)\end{aligned}$$