Aprendizaje de máquinas
Regresión lineal

Felipe Tobar

Facultad de Ciencias Físicas y Matemáticas
Universidad de Chile

Otoño, 2021.

Problema de regresión

Para un conjunto de entrenamiento $\mathcal{D}$ que contiene $N \in \mathbb{N}$ observaciones de entrada y salida, respectivamente $\{x_i\}_{i=1}^N$ y $\{y_i\}_{i=1}^N$, de la forma

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^M \times \mathbb{R}$$

Problema de regresión

Para un conjunto de entrenamiento $\mathcal{D}$ que contiene $N \in \mathbb{N}$ observaciones de entrada y salida, respectivamente $\{x_i\}_{i=1}^N$ y $\{y_i\}_{i=1}^N$, de la forma

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^M \times \mathbb{R}$$

la regresión lineal busca encontrar un modelo lineal, es decir, una función f definida por

$$\begin{aligned} f &: \mathbb{R}^M \to \mathbb{R} \\ x &\mapsto f(x) = a^\top x + b, \quad a \in \mathbb{R}^M, b \in \mathbb{R} \end{aligned}$$

que *mejor represente* la forma en que la variable y depende de la variable x, en base a las las observaciones contenidas en el conjunto $\mathcal{D}$.

Regresión mediante mínimos cuadrados

Dada la formulación anterior, surge la pregunta natural acerca de qué criterio utilizar para poder elegir la *mejor* función que represente los datos. Un criterio ampliamente usado (y el que se utilizará en el resto del curso) corresponde al criterio de mínimos cuadrados (MC).

Regresión mediante mínimos cuadrados

Dada la formulación anterior, surge la pregunta natural acerca de qué criterio utilizar para poder elegir la *mejor* función que represente los datos. Un criterio ampliamente usado (y el que se utilizará en el resto del curso) corresponde al criterio de mínimos cuadrados (MC).

Para una función afín f, su error cuadrático de ajuste con respecto a $\mathcal{D}$ es:

$$J(\mathcal{D}, f) = \sum_{i=1}^{N} (y_i - f(x_i))^2$$

Regresión mediante mínimos cuadrados

Dada la formulación anterior, surge la pregunta natural acerca de qué criterio utilizar para poder elegir la *mejor* función que represente los datos. Un criterio ampliamente usado (y el que se utilizará en el resto del curso) corresponde al criterio de mínimos cuadrados (MC).

Para una función afín f, su error cuadrático de ajuste con respecto a $\mathcal{D}$ es:

$$J(\mathcal{D}, f) = \sum_{i=1}^{N} (y_i - f(x_i))^2$$

De esta forma, se elegirá el regresor óptimo $f^\star$ como la función afín que minimice dicho error, es decir:

$$f^\star = \underset{f \text{ es afín}}{\arg\min}\, J(\mathcal{D}, f)$$

Regresión mediante mínimos cuadrados

Dado que toda función afín se escribe de la forma $f(x) = a^\top x + b$, encontrar el funcional óptimo equivalente a encontrar los parámetros $a^\star, b^\star$ que minimicen el error cuadrático:

$$a^\star, b^\star = \arg\min_{a,b} \sum_{i=1}^{N} (y_i - a^\top x_i - b)^2$$

Regresión mediante mínimos cuadrados

Dado que toda función afín se escribe de la forma $f(x) = a^\top x + b$, encontrar el funcional óptimo equivalente a encontrar los parámetros $a^\star, b^\star$ que minimicen el error cuadrático:

$$a^\star, b^\star = \arg\min_{a,b} \sum_{i=1}^{N} (y_i - a^\top x_i - b)^2$$

Para optimizar sobre un único parámetro, es común usar el siguiente cambio de variable sobre los datos:

$$\tilde{x}_i = \begin{pmatrix} x_i \\ 1 \end{pmatrix} \in \mathbb{R}^{M+1}, \quad \theta = \begin{pmatrix} a \\ b \end{pmatrix} \in \mathbb{R}^{M+1} \implies J(\mathcal{D}, \theta) = \sum_{i=1}^{N} (y_i - \theta^\top \tilde{x}_i)^2$$

Regresión mediante mínimos cuadrados

Dado que toda función afín se escribe de la forma $f(x) = a^\top x + b$, encontrar el funcional óptimo equivalente a encontrar los parámetros $a^\star, b^\star$ que minimicen el error cuadrático:

$$a^\star, b^\star = \arg\min_{a,b} \sum_{i=1}^{N} (y_i - a^\top x_i - b)^2$$

Para optimizar sobre un único parámetro, es común usar el siguiente cambio de variable sobre los datos:

$$\tilde{x}_i = \begin{pmatrix} x_i \\ 1 \end{pmatrix} \in \mathbb{R}^{M+1}, \quad \theta = \begin{pmatrix} a \\ b \end{pmatrix} \in \mathbb{R}^{M+1} \implies J(\mathcal{D}, \theta) = \sum_{i=1}^{N} (y_i - \theta^\top \tilde{x}_i)^2$$

El funcional anterior puede ser simplificado utilizando una única matriz que contenga todos los datos:

$$\tilde{X} = \begin{pmatrix} \tilde{x}_1^\top \\ \vdots \\ \tilde{x}_N^\top \end{pmatrix} \in \mathbb{R}^{N \times (M+1)}, \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \in \mathbb{R}^N \implies J = \left\| Y - \tilde{X}\theta \right\|_2^2$$

Donde $\tilde{X}$ se denomina matriz de diseño o matriz de regresión.

Regresión mediante mínimos cuadrados

dado que el funcional $J = \left\| Y - \tilde{X}\theta \right\|_2^2$ es estrictamente convexo, tiene un único mínimo y puede ser encontrado utilizando la condición de primer orden:

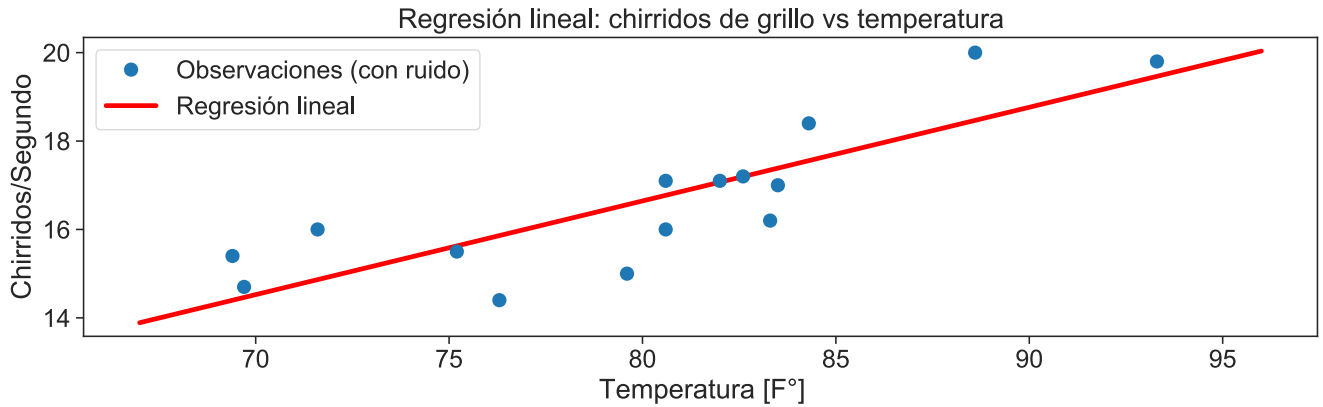
$$\begin{aligned} \nabla_\theta J = 2(Y - \tilde{X}\theta)^\top(-\tilde{X}) = 0 \\ \iff Y^\top \tilde{X} - \theta^\top \tilde{X}^\top \tilde{X} = 0 \\ \iff \theta^\top = Y^\top \tilde{X}(\tilde{X}^\top \tilde{X})^{-1} \\ \iff \theta = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top Y \end{aligned}$$

Regresión mediante mínimos cuadrados

dado que el funcional $J = \left\| Y - \tilde{X}\theta \right\|_2^2$ es estrictamente convexo, tiene un único mínimo y puede ser encontrado utilizando la condición de primer orden:

$$\begin{aligned} \nabla_\theta J = 2(Y - \tilde{X}\theta)^\top(-\tilde{X}) = 0 \\ \iff Y^\top \tilde{X} - \theta^\top \tilde{X}^\top \tilde{X} = 0 \\ \iff \theta^\top = Y^\top \tilde{X}(\tilde{X}^\top \tilde{X})^{-1} \\ \iff \theta = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top Y \end{aligned}$$

Ejemplo de regresión lineal mediante mínimos cuadrados:



Análisis de la solución de mínimos cuadrados

Se obtuvo que el parámetro óptimo $\theta = (a, b)^\top$ corresponde a:

$$\theta = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top Y$$

Análisis de la solución de mínimos cuadrados

Se obtuvo que el parámetro óptimo $\theta = (a, b)^\top$ corresponde a:

$$\theta = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top Y$$

- La expresión $\left(\tilde{X}^\top \tilde{X}\right)^{-1} \tilde{X}^\top$ corresponde a la pseudoinversa de Moore-Penrose de $\tilde{X}$.

Análisis de la solución de mínimos cuadrados

Se obtuvo que el parámetro óptimo $\theta = (a, b)^\top$ corresponde a:

$$\theta = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top Y$$

- La expresión $\left(\tilde{X}^\top \tilde{X}\right)^{-1} \tilde{X}^\top$ corresponde a la pseudoinversa de Moore-Penrose de $\tilde{X}$.
- Para que $\tilde{X}^\top \tilde{X}$ sea invertible, es necesario que $r(\tilde{X}) = M + 1$.

Análisis de la solución de mínimos cuadrados

Se obtuvo que el parámetro óptimo $\theta = (a, b)^\top$ corresponde a:

$$\theta = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top Y$$

- La expresión $\left(\tilde{X}^\top \tilde{X}\right)^{-1} \tilde{X}^\top$ corresponde a la pseudoinversa de Moore-Penrose de $\tilde{X}$.
- Para que $\tilde{X}^\top \tilde{X}$ sea invertible, es necesario que $r(\tilde{X}) = M + 1$.
- Dado que $\tilde{X} \in \mathbb{R}^{N \times (M+1)}$, entonces $r(\tilde{X}) \leq \text{mín}\{N, M + 1\}$ por lo que necesariamente se debe cumplir que $N \geq M + 1$.

Análisis de la solución de mínimos cuadrados

Se obtuvo que el parámetro óptimo $\theta = (a, b)^\top$ corresponde a:

$$\theta = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top Y$$

- La expresión $\left(\tilde{X}^\top \tilde{X}\right)^{-1} \tilde{X}^\top$ corresponde a la pseudoinversa de Moore-Penrose de $\tilde{X}$.
- Para que $\tilde{X}^\top \tilde{X}$ sea invertible, es necesario que $r(\tilde{X}) = M + 1$.
- Dado que $\tilde{X} \in \mathbb{R}^{N \times (M+1)}$, entonces $r(\tilde{X}) \leq \text{mín}\{N, M + 1\}$ por lo que necesariamente se debe cumplir que $N \geq M + 1$.
- Es decir, el número de muestras debe ser mayor que el número de dimensiones. Intuitivamente esto se debe a que se necesitan $M + 1$ puntos para fijar un hiperplano en $\mathbb{R}^{M+1}$

¿Por qué mínimos cuadrados?

Luego de haber encontrado el regresor óptimo de acuerdo al criterio de mínimos cuadrados, nace la pregunta de por qué utilizar ese criterio y no otro. Los principales argumentos son los siguiente:

¿Por qué mínimos cuadrados?

Luego de haber encontrado el regresor óptimo de acuerdo al criterio de mínimos cuadrados, nace la pregunta de por qué utilizar ese criterio y no otro. Los principales argumentos son los siguiente:

- Su solución tiene forma cerrada, por lo que se evitan algoritmos numéricos.

¿Por qué mínimos cuadrados?

Luego de haber encontrado el regresor óptimo de acuerdo al criterio de mínimos cuadrados, nace la pregunta de por qué utilizar ese criterio y no otro. Los principales argumentos son los siguiente:

- Su solución tiene forma cerrada, por lo que se evitan algoritmos numéricos.
- Desde un punto interpretativo, el exponente 2 busca penalizar mayormente grandes diferencias (mayores que 1) entre la predicción y el valor real, mientras que le quita importancia a errores de predicción pequeños (menores que 1).

¿Por qué mínimos cuadrados?

Luego de haber encontrado el regresor óptimo de acuerdo al criterio de mínimos cuadrados, nace la pregunta de por qué utilizar ese criterio y no otro. Los principales argumentos son los siguiente:

- Su solución tiene forma cerrada, por lo que se evitan algoritmos numéricos.
- Desde un punto interpretativo, el exponente 2 busca penalizar mayormente grandes diferencias (mayores que 1) entre la predicción y el valor real, mientras que le quita importancia a errores de predicción pequeños (menores que 1).
- La medida del error cuadrático representa la varianza muestral: si considerásemos que x_i e y_i son observaciones iid de variables aleatorias X e Y respectivamente, entonces el error cuadrático medio asociado a la función f definido por

$$ECM = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2$$

corresponde a la varianza muestral de la variable aleatoria $Y - f(X)$ (asumiendo que $\mathbb{E}(Y - f(X)) = 0$).

Interpretación geométrica

El problema de regresión lineal requiere encontrar una solución aproximada de un sistema lineal sobredeterminado definido por

$$\tilde{X}\theta = Y$$

donde la cantidad de incógnitas $(M + 1)$ es ampliamente superada por el número de ecuaciones (N), por lo que desde el punto de vista de un sistema lineal, la solución no necesariamente existe.

Interpretación geométrica

El problema de regresión lineal requiere encontrar una solución aproximada de un sistema lineal sobredeterminado definido por

$$\tilde{X}\theta = Y$$

donde la cantidad de incógnitas $(M + 1)$ es ampliamente superada por el número de ecuaciones (N), por lo que desde el punto de vista de un sistema lineal, la solución no necesariamente existe.

Por lo anterior, se puede proceder a encontrar la solución para θ que reporta *la menor discrepancia* entre ambos lados de la ecuación (8), para esto, se realiza lo siguiente:

Interpretación geométrica

El problema de regresión lineal requiere encontrar una solución aproximada de un sistema lineal sobredeterminado definido por

$$\tilde{X}\theta = Y$$

donde la cantidad de incógnitas $(M+1)$ es ampliamente superada por el número de ecuaciones (N), por lo que desde el punto de vista de un sistema lineal, la solución no necesariamente existe.

Por lo anterior, se puede proceder a encontrar la solución para θ que reporta *la menor discrepancia* entre ambos lados de la ecuación (8), para esto, se realiza lo siguiente:

- Identificar el espacio formado por todos los posibles valores que toma la combinación lineal $\tilde{X}\theta$, es decir, el *span* de todas las columnas de $\tilde{X}$:

$$\text{span}(\tilde{X}) = \{\tilde{X}\theta : \theta \in \mathbb{R}^{M+1}\} \leq \mathbb{R}^{N}$$

Interpretación geométrica

El problema de regresión lineal requiere encontrar una solución aproximada de un sistema lineal sobredeterminado definido por

$$\tilde{X}\theta = Y$$

donde la cantidad de incógnitas $(M+1)$ es ampliamente superada por el número de ecuaciones (N), por lo que desde el punto de vista de un sistema lineal, la solución no necesariamente existe.

Por lo anterior, se puede proceder a encontrar la solución para θ que reporta *la menor discrepancia* entre ambos lados de la ecuación (8), para esto, se realiza lo siguiente:

- Identificar el espacio formado por todos los posibles valores que toma la combinación lineal $\tilde{X}\theta$, es decir, el *span* de todas las columnas de $\tilde{X}$:

$$\text{span}(\tilde{X}) = \{\tilde{X}\theta : \theta \in \mathbb{R}^{M+1}\} \leq \mathbb{R}^N$$

- identificar el elemento de dicho espacio que está más cerca de $Y \in \mathbb{R}^N$ como la proyección del propio Y en $\text{span}(\tilde{X})$.

Interpretación geométrica

La condición para identificar dicha proyección es precisamente que el vector error $e = Y - \tilde{X}\theta$ sea ortogonal al espacio span$(\tilde{X})$ generado por los datos de entrada. Dado que las columnas de $\tilde{X}$ son una base de span$(\tilde{X})$:

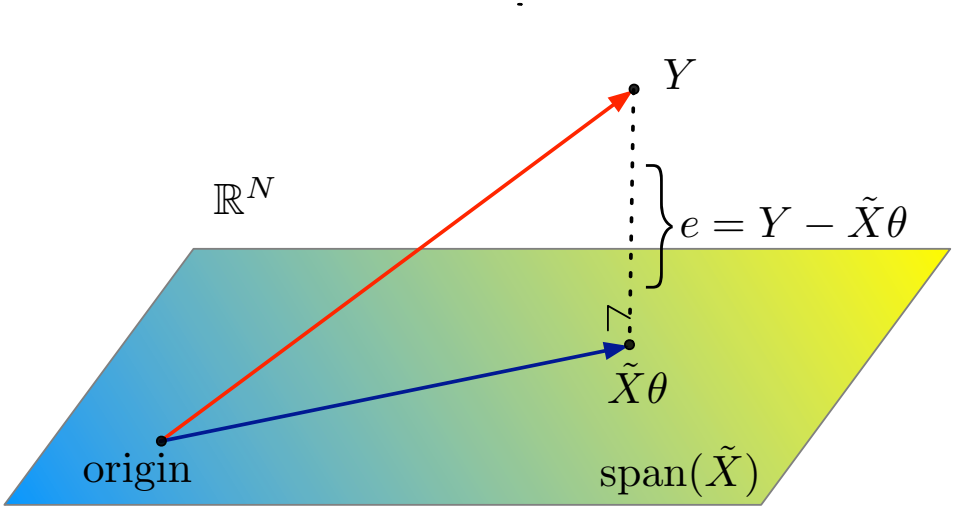
$$u \in \text{span}(\tilde{X}) \iff \exists \theta_u \in \mathbb{R}^{M+1} : u = \tilde{X}\theta_u,$$

Interpretación geométrica

La condición para identificar dicha proyección es precisamente que el vector error $e = Y - \tilde{X}\theta$ sea ortogonal al espacio span$(\tilde{X})$ generado por los datos de entrada. Dado que las columnas de $\tilde{X}$ son una base de span$(\tilde{X})$:

$$u \in \text{span}(\tilde{X}) \iff \exists \theta_u \in \mathbb{R}^{M+1} : u = \tilde{X}\theta_u,$$

por lo tanto:



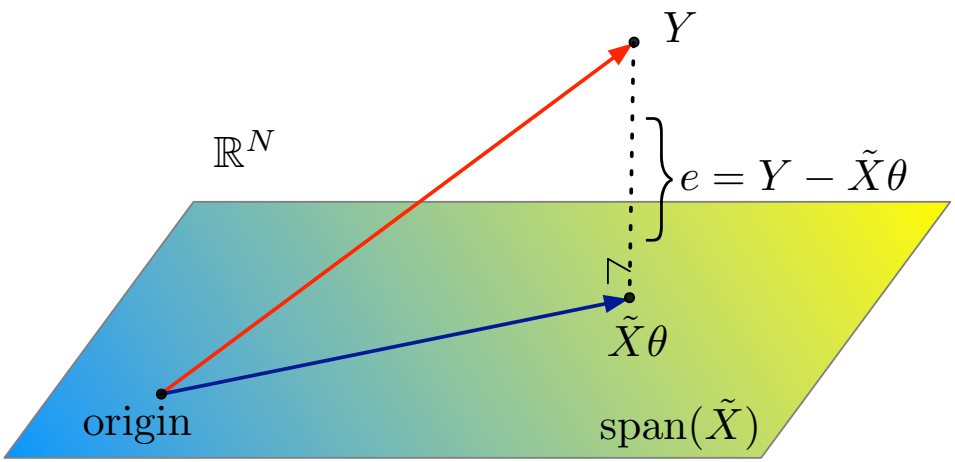
$$\begin{aligned} e \perp \text{span}(\tilde{X}) &\iff e \perp u, \forall u \in \text{span}(\tilde{X}) \\ &\iff (Y - \tilde{X}\theta)^\top \tilde{X}\theta_u = 0, \forall \theta_u \in \mathbb{R}^{M+1} \\ &\iff (Y - \tilde{X}\theta)^\top \tilde{X} = 0 \\ &\iff \theta = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top Y \end{aligned}$$

Interpretación geométrica

La condición para identificar dicha proyección es precisamente que el vector error $e = Y - \tilde{X}\theta$ sea ortogonal al espacio span$(\tilde{X})$ generado por los datos de entrada. Dado que las columnas de $\tilde{X}$ son una base de span$(\tilde{X})$:

$$u \in \text{span}(\tilde{X}) \iff \exists \theta_u \in \mathbb{R}^{M+1} : u = \tilde{X}\theta_u,$$

por lo tanto:

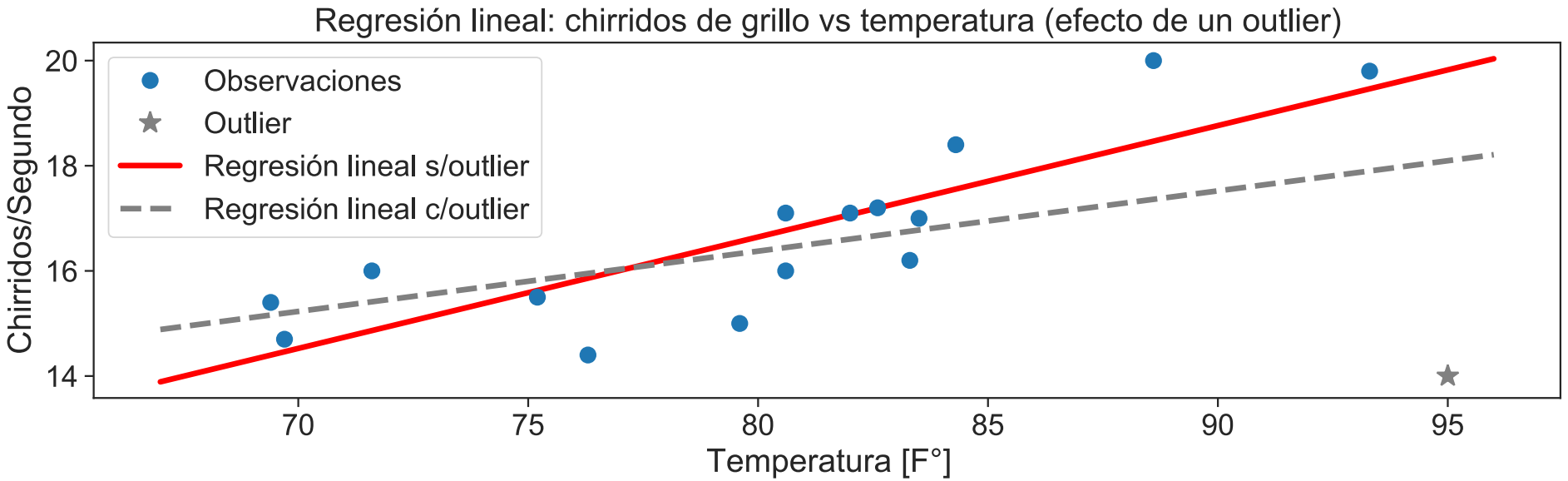


$$\begin{aligned} e \perp \text{span}(\tilde{X}) &\iff e \perp u, \forall u \in \text{span}(\tilde{X}) \\ &\iff (Y - \tilde{X}\theta)^\top \tilde{X}\theta_u = 0, \forall \theta_u \in \mathbb{R}^{M+1} \\ &\iff (Y - \tilde{X}\theta)^\top \tilde{X} = 0 \\ &\iff \theta = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top Y \end{aligned}$$

Es decir, la distancia se minimiza para el θ, correspondiente a la solución por mínimos cuadrados.

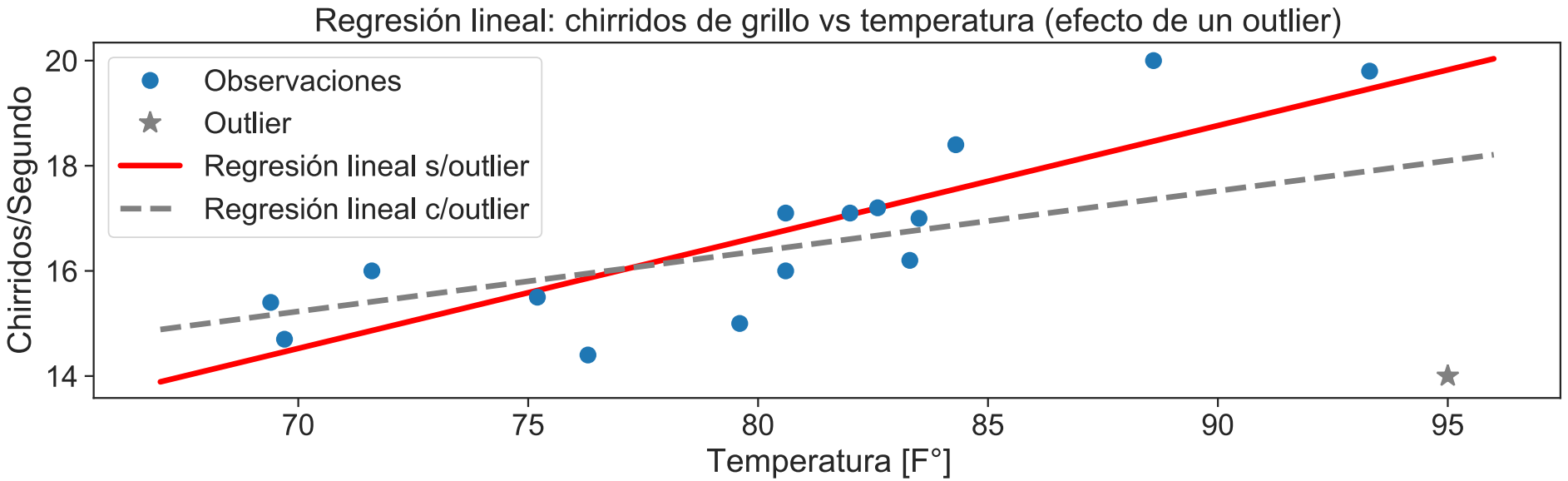
Problema de mínimos cuadrados

El principal problema del criterio de mínimos cuadrados es que es sensible a muestras atípicas (outliers) debido a la penalización cuadrática en su funcional de costos. Esto puede ser observado en la siguiente figura:



Problema de mínimos cuadrados

El principal problema del criterio de mínimos cuadrados es que es sensible a muestras atípicas (outliers) debido a la penalización cuadrática en su funcional de costos. Esto puede ser observado en la siguiente figura:



En conclusión, el criterio de mínimos cuadrados no es útil cuando existen muestras que se alejan de la tendencia buscada, por lo que se debe buscar algún método para corregir el problema. Una forma de realizar esto es *mediante mínimos cuadrados regularizados*.