Regresión lineal: mínimos cuadrados

Matías Carrasco

4 de septiembre de 2023

Índice

1. Modelo lineal univariado

Consideremos el siguiente setting:

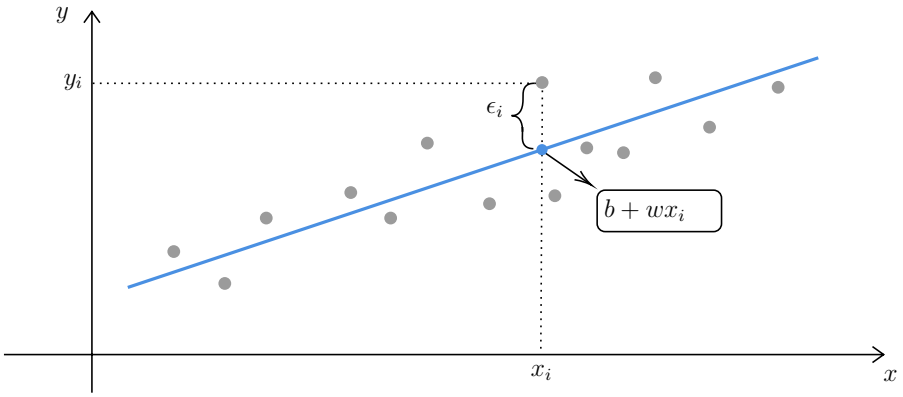
- Espacio de atributos $\mathcal{X} \subset \mathbb{R}$.

- Espacio de etiquetas $\mathcal{Y} \subset \mathbb{R}$.
- Espacio de observaciones $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.
- Conjunto de datos $S = \{(x_i, y_i)\}_{i=1}^N$.
- Función de pérdida cuadrática: $\text{Loss}(y, y') = (y - y')^2$

El modelo lineal consiste en aproximar la relación entre x e y mediante una recta

$$y = f(x; \boldsymbol{\theta}) = b + wx$$

en donde el vector de parámetros es $\boldsymbol{\theta} = [b, w]^\top \in \mathbb{R}^2$.



La pérdida promedio en todo el conjunto de datos S se llama MSE (Mean Square Error) y la usaremos para elegir los mejores parámetros:

$$L(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i; \boldsymbol{\theta}))^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - b - wx_i)^2$$

de modo que

$$\widehat{b}, \widehat{w} = \arg\min_{b,w} \{L(b, w)\}.$$

2. Cálculo de los coeficientes

Observación: dados números reales $A_1, \ldots, A_N$

$$\arg\min_{a} \left\{ \sum_{i=1}^{N} (A_i - a)^2 \right\} = \bar{A}$$

en donde denotamos $\bar{A}$ el promedio $\frac{1}{N}\sum_i A_i$.

Por la observación, si fijamos w, el mejor b es

$$\arg\min_b \left\{ \frac{1}{N} \sum_{i=1}^{N} (\underbrace{y_i - wx_i}_{A_i} - b)^2 \right\} = \bar{y} - w\bar{x}$$

Luego basta encontrar $\widehat{w}$

$$\begin{aligned}
L(w) &= L(\bar{y} - w\bar{x}, w) \\
&= \frac{1}{N} \sum_{i=1}^{N} (y_i - wx_i - \bar{y} + w\bar{x})^2 \\
&= \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{y} - w(x_i - \bar{x}))^2 \\
&= \underbrace{\frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{y})^2}_{s_y^2} - 2w \underbrace{\frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{y})(x_i - \bar{x})}_{s_{xy}} + w^2 \underbrace{\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2}_{s_x^2} \\
&= s_y^2 - 2s_{xy}w + s_x^2 w^2 \quad \text{(polinomio de grado 2 en } w\text{)}
\end{aligned}$$

Sus raíces son

$$\frac{2s_{xy} \pm \sqrt{4s_{xy-4s_x^2 s_y^2}^2}}{2s_x^2}$$

pero como $L(w) \geq 0$ a lo sumo tiene una raíz doble, de donde obtenemos la condición:

$$\frac{s_{xy}^2}{s_x^2 s_y^2} \leq 1 \Leftrightarrow \left| \frac{s_{xy}}{s_x s_y} \right| \leq 1$$

El número $r = \frac{s_{xy}}{s_x s_y}$ se llama coeficiente de correlación y acabamos de probar que $r \in [-1, 1]$. Más aún, probamos que $r = \pm 1$ si, y sólo si, $y_i = b + wx_i$ para todo i.

Volviendo a $L(w)$ y completando el cuadrado tenemos:

$$\begin{aligned}
L(w) &= s_y^2 - 2s_{xy}w + s_x^2 w^2 \\
&= s_y^2 - 2rs_y s_x w + s_x^2 w^2 \\
&= s_y^2 - r^2 s_y^2 + r^2 s_y^2 - 2rs_y s_x w + s_x^2 w^2 \\
&= s_y^2(1 - r^2) + (rs_y - s_x w)^2
\end{aligned}$$

De aquí vemos que el mínimo se alcanza en $\widehat{w} = r\frac{s_y}{s_x}$ y vale $L(\widehat{w}) = s_y(1 - r^2)$.

Si llamamos $\epsilon_i = y_i - (b + wx_i)$ llegamos a la siguiente fórmula para la MSE mínima:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}\epsilon_i^2 = s_y^2(1 - r^2)$$

En estadística se suele usar r^2 como medida de ajuste en lugar de la MSE.

3. Modelo lineal univariado en notación matricial

La función lineal puede escribirse como un producto escalar de vectores

$$b + wx_i = [1, x_i] \cdot \begin{bmatrix} b \\ w \end{bmatrix} = \boldsymbol{x}_i^\top \boldsymbol{\theta} = \boldsymbol{\theta}^\top \boldsymbol{x}_i$$

Considerando la matriz de diseño

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \\ 1 & x_N \end{bmatrix}$$

Podemos juntar todos los valores de $\{b + wx_i\}_{i=1}^N$ en un solo producto matricial

$$\underbrace{\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}}_{\boldsymbol{X}} \cdot \underbrace{\begin{bmatrix} b \\ w \end{bmatrix}}_{\boldsymbol{\theta}} = \underbrace{\begin{bmatrix} b + wx_1 \\ b + wx_2 \\ \vdots \\ b + wx_N \end{bmatrix}}_{\boldsymbol{X\theta}}$$

y compararlo con el vector de targets o etiquetas

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}.$$

De este modo la función de pérdida queda

$$L(\boldsymbol{\theta}) = \frac{1}{N} \underbrace{\sum_{i=1}^{N} (y_i - b - wx_i)^2}_{\text{norma}} = \frac{1}{N} \|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}\|^2$$

$$= \frac{1}{N} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})$$

Haciendo las distributivas obtenemos

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

Derivando respecto al vector $\boldsymbol{\theta}$ tenemos

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = -\boldsymbol{X}^\top \boldsymbol{y} - \boldsymbol{X}^\top \boldsymbol{y} + 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\theta} = -2\boldsymbol{X}^\top \boldsymbol{y} + 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{0}$$

de donde obtenemos

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

4. Ambos cálculos son equivalentes

Para obtener $\widehat{\boldsymbol{\theta}}$ a partir del producto matricial seguimos estos pasos:

1. Calculamos el producto transpuesto de la matriz de diseño $\boldsymbol{X}^\top \boldsymbol{X}$:

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

2. Calculamos la inversa de la matriz resultante:

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

3. Calculamos el producto de la matriz de diseño transpuesta $\boldsymbol{X}^\top$ y el vector de targets $\boldsymbol{y}$:

$$\boldsymbol{X}^\top \boldsymbol{y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_N \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^N y_i \\ \sum_{i=1}^N x_i y_i \end{bmatrix}$$

4. Finalmente, multiplicamos el resultado del paso 2 con el resultado del paso 3 para obtener los coeficientes $\widehat{\boldsymbol{\theta}}$:

$$\begin{aligned}
\widehat{\boldsymbol{\theta}} &= \frac{1}{N\sum_{i=1}^N x_i^2 - \left(\sum_{i=1}^N x_i\right)^2} \begin{bmatrix} \sum_{i=1}^N x_i^2 & -\sum_{i=1}^N x_i \\ -\sum_{i=1}^N x_i & N \end{bmatrix} \begin{bmatrix} \sum_{i=1}^N y_i \\ \sum_{i=1}^N x_i y_i \end{bmatrix} \\
&= \frac{1}{N^2\overline{x^2} - N^2\bar{x}^2} \begin{bmatrix} N\overline{x^2} & -N\bar{x} \\ -N\bar{x} & N \end{bmatrix} \begin{bmatrix} N\bar{y} \\ N\overline{xy} \end{bmatrix} \\
&= \frac{1}{\overline{x^2} - \bar{x}^2} \begin{bmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \begin{bmatrix} \bar{y} \\ \overline{xy} \end{bmatrix} \\
&= \frac{1}{\overline{x^2} - \bar{x}^2} \begin{bmatrix} \overline{x^2}\bar{y} - \bar{x}\overline{xy} \\ -\bar{x}\bar{y} + \overline{xy} \end{bmatrix} \\
&= \frac{1}{s_x^2} \begin{bmatrix} (s_x^2 + \bar{x}^2)\bar{y} - \bar{x}(s_{xy} + \bar{x}\bar{y}) \\ s_{xy} \end{bmatrix} = \begin{bmatrix} \bar{y} - \bar{x} r \frac{s_y}{s_x} \\ r\frac{s_y}{s_x} \end{bmatrix} = \begin{bmatrix} \widehat{b} \\ \widehat{w} \end{bmatrix}
\end{aligned}$$

5. Interpretación de la correlación

Una forma directa de interpretar la correlación es como $r^2 = \text{cor}(\boldsymbol{y}, \widehat{\boldsymbol{y}})$ siendo $\widehat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\theta}$ el vector de predicciones. Esto es así pues la correlación es invariante por transformaciones lineales.

En estadística aplicada se suele interpretar r^2 como porcentaje de la variabilidad de $\boldsymbol{y}$ explicada por $\boldsymbol{x}$. De hecho, la varianza de $\widehat{\boldsymbol{y}}$ es

$$s_{\widehat{y}}^2 = \text{var}\,(\widehat{\boldsymbol{y}}) = \text{var}\left(\widehat{b} + \widehat{w}\boldsymbol{x}\right) = \widehat{w}^2\,\text{var}(\boldsymbol{x}) = r^2\frac{s_y^2}{s_x^2}s_x^2 = r^2 s_y^2$$

de donde

$$\%\text{ de variación explicada} = \frac{\text{var}\,(\widehat{\boldsymbol{y}})}{\text{var}\,(\boldsymbol{y})} \times 100 = \frac{r^2 s_y^2}{s_y^2} \times 100 = r^2 \times 100$$

6. Modelo lineal multivariado

El setting para el modelo multivariado es análogo salvo que

- Espacio de atributos $\mathcal{X} \subset \mathbb{R}^D$.

pues disponemos de D atributos.

Ahora el modelo es

$$y = f(\boldsymbol{x}; \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \boldsymbol{x}$$

en donde los vectores $\boldsymbol{\theta}$ y $\boldsymbol{x}$ están dados por

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

y el producto representa

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

Considerando la matriz de diseño y el vector de targets

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

los cálculos son idénticos y la solución es la misma $\widehat{\boldsymbol{\theta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$.

Al mirar en detalle la función de pérdida

$$L(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \Bigg(\underbrace{y_i - \sum_{j=1}^{D} w_j x_i^{(j)}}_{A_i} - b \Bigg)^2$$

vemos que si fijamos $\boldsymbol{w}$ el mejor b es $\bar{y} - \sum_{j=1}^{D} w_j \bar{x}^{(j)}$. Es decir, nuevamente b cumple el rol de ajustar promedios.

Si escribimos la función de pérdida usando solamente $\boldsymbol{w}$ obtenemos

$$L(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N}\left((y_i - \bar{y}) - \sum_{j=1}^{D} w_j\left(x_i^{(j)} - \bar{x}^{(j)}\right)\right)^2$$

Si denotamos $u_i = y_i - \bar{y}$ y $z_i^{(j)} = x_i^{(j)} - \bar{x}^{(j)}$ podemos reescribir la pérdida

$$L(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N}\left(u_i - \sum_{j=1}^{D} w_j z_i^{(j)}\right)^2 = \frac{1}{N}\left(\boldsymbol{u} - \boldsymbol{Z}\boldsymbol{w}\right)^\top\left(\boldsymbol{u} - \boldsymbol{Z}\boldsymbol{w}\right)$$

en donde $\boldsymbol{u}$ es el vector de targets centrado y $\boldsymbol{Z}$ es la matriz de diseño cuyas columnas están centradas

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{z}_1^\top \\ \boldsymbol{z}_2^\top \\ \vdots \\ \boldsymbol{z}_N^\top \end{bmatrix} = \begin{bmatrix} z_1^{(1)} & \cdots & z_1^{(D)} \\ z_2^{(1)} & \cdots & z_2^{(D)} \\ \vdots & \vdots & \vdots \\ z_N^{(1)} & \cdots & z_N^{(D)} \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}$$

La solución óptima está dada por $\widehat{\boldsymbol{w}} = \left(\boldsymbol{Z}^\top\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^\top\boldsymbol{u}$.

7. Interpretación de $\boldsymbol{Z}^\top\boldsymbol{u}$

El producto es

$$\boldsymbol{Z}^\top\boldsymbol{u} = \left[\begin{array}{c|c|c|c} z_1^{(1)} & z_2^{(1)} & \cdots & z_N^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ z_1^{(D)} & z_2^{(D)} & \cdots & z_N^{(D)} \end{array}\right]\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix} = \sum_{i=1}^{N}\boldsymbol{z}_i u_i$$

Esto nos dará como resultado un vector columna de dimensiones $D \times 1$ donde las entradas son las sumas ponderadas de los u_i por las respectivas coordenadas de los $\boldsymbol{z}_i$:

$$\boldsymbol{Z}^\top\boldsymbol{u} = \begin{bmatrix} \sum_{i=1}^{N} z_i^{(1)} u_i \\ \vdots \\ \sum_{i=1}^{N} z_i^{(D)} u_i \end{bmatrix} = N\,\mathrm{cov}\left(\left[\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(D)}\right], \boldsymbol{y}\right)$$

8. Interpretación de $\boldsymbol{Z}^\top \boldsymbol{Z}$

Por otro lado el producto matricial $\boldsymbol{Z}^\top \boldsymbol{Z}$ es:

$$\boldsymbol{Z}^\top \boldsymbol{Z} = \begin{bmatrix} \boldsymbol{z}_1 & \boldsymbol{z}_2 & \cdots & \boldsymbol{z}_N \end{bmatrix} \begin{bmatrix} \boldsymbol{z}_1^\top \\ \boldsymbol{z}_2^\top \\ \vdots \\ \boldsymbol{z}_N^\top \end{bmatrix}$$

Es una matriz cuadrada de tamaño $D \times D$, donde cada elemento (j, k) es el resultado del producto escalar entre $\boldsymbol{z}^{(j)}$ y $\boldsymbol{z}^{(k)}$:

$$\boldsymbol{Z}^\top \boldsymbol{Z} = \begin{bmatrix} \boldsymbol{z}^{(1)} \cdot \boldsymbol{z}^{(1)} & \cdots & \boldsymbol{z}^{(1)} \cdot \boldsymbol{z}^{(D)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{z}^{(D)} \cdot \boldsymbol{z}^{(1)} & \cdots & \boldsymbol{z}^{(D)} \cdot \boldsymbol{z}^{(D)} \end{bmatrix}$$

Es decir, $\boldsymbol{Z}^\top \boldsymbol{Z} = N \operatorname{cov}\left(\left[\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(D)}\right]\right)$.

En resumen el vector de pesos óptimo se puede escribir como

$$\boxed{\widehat{\boldsymbol{w}} = \operatorname{cov}\left(\left[\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(D)}\right]\right)^{-1} \operatorname{cov}\left(\left[\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(D)}\right], \boldsymbol{y}\right)}$$

9. Efecto de la multicolinealidad

Supongamos que el atributo (columna de $\boldsymbol{Z}$)

$$\boldsymbol{x}^{(D)} = \sum_{j=1}^{D-1} \alpha_j \boldsymbol{x}^{(j)}$$

es combinación lineal de los otros. Esto afectara las columnas de la matriz $\boldsymbol{Z}^\top \boldsymbol{Z}$ pues

$$\begin{aligned} \boldsymbol{Z}^\top \boldsymbol{Z} &= \begin{bmatrix} \boldsymbol{z}^{(1)} \cdot \boldsymbol{z}^{(1)} & \cdots & \boldsymbol{z}^{(1)} \cdot \boldsymbol{z}^{(D)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{z}^{(D)} \cdot \boldsymbol{z}^{(1)} & \cdots & \boldsymbol{z}^{(D)} \cdot \boldsymbol{z}^{(D)} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{z}^{(1)} \cdot \boldsymbol{z}^{(1)} & \cdots & \sum_{j=1}^{D-1} \alpha_j \boldsymbol{z}^{(1)} \cdot \boldsymbol{z}^{(j)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{z}^{(D)} \cdot \boldsymbol{z}^{(1)} & \cdots & \sum_{j=1}^{D-1} \alpha_j \boldsymbol{z}^{(D)} \cdot \boldsymbol{z}^{(j)} \end{bmatrix} \end{aligned}$$

en donde también obtenemos una columna como combinación lineal de las demás. En particular el determinante $\det\left(\boldsymbol{Z}^\top \boldsymbol{Z}\right) = 0$ y la matriz no es invertible. Esto hace que el vector de pesos $\boldsymbol{w}$ no quede unívocamente determinado y dificulte su aproximación.

Veamos con más detalle el caso bi-variado, es decir con $D = 2$ atributos. Supongamos que $\boldsymbol{x}^{(1)}$ y $\boldsymbol{x}^{(2)}$ han sido estandarizados y denotemos $r = \operatorname{cor}\left(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}\right)$ su coeficiente de correlación. Si la y está estandarizada también, vemos que $\widehat{b} = 0$, y los pesos están dados por:

$$\widehat{\boldsymbol{w}} = \operatorname{cor}\left([\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}]\right)^{-1} \operatorname{cor}\left([\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}], \boldsymbol{y}\right)$$

La matriz de correlación entre $\boldsymbol{x}^{(1)}$ y $\boldsymbol{x}^{(2)}$ es

$$\begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$$

por lo que su inversa es

$$\operatorname{cor}\left([\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}]\right)^{-1} = \frac{1}{1-r^2} \begin{bmatrix} 1 & -r \\ -r & 1 \end{bmatrix}$$

Si denotamos el vector de correlaciones entre $\boldsymbol{x}^{(1)}$ y $\boldsymbol{x}^{(2)}$ con $\boldsymbol{y}$ como

$$\operatorname{cor}\left([\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}], \boldsymbol{y}\right) = \begin{bmatrix} r_1 \\ r_2 \end{bmatrix}$$

tenemos que

$$\widehat{\boldsymbol{w}} = \frac{1}{1-r^2} \begin{bmatrix} 1 & -r \\ -r & 1 \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} = \frac{1}{1-r^2} \begin{bmatrix} r_1 - r r_2 \\ r_2 - r r_1 \end{bmatrix}$$

Las correlaciones r, r_1, y r_2 no pueden elegirse arbitrariamente pero basta que satisfagan la relación

$$1 - r_1^2 - r_2^2 + 2 r_1 r_2 r - r^2 \geq 0$$

Esto equivale a que r pertenezca al intervalo

$$\frac{2 r_1 r_2 \pm \sqrt{4 r_1^2 r_2^2 + 4(1 - r_1^2 - r_2^2)}}{2} = r_1 r_2 \pm \sqrt{(1 - r_1^2)(1 - r_2^2)}$$

Se puede ver que es válido elegir $r_1 = \rho$, $r_2 = -\rho$ y $r = 1 - 2\rho^2$. En ese caso los pesos son

$$w_1 = \frac{1}{2\rho} \quad w_2 = -\frac{1}{2\rho}$$

y haciendo $\rho \to 0$ vemos que pueden ser arbitrariamente grandes. Este ejemplo muestra cómo la multicolinealidad puede dar origen a pesos muy grandes y motiva la consideración de la siguiente sección.

10. Regresión lineal con funciones base

La regresión con funciones de base es una técnica que permite modelar relaciones no lineales entre el target y y uno o más atributos, supongamos en este caso, un sólo atributo x.

Supongamos que deseamos modelar y no simplemente como una función lineal de x, sino utilizando una serie de funciones $h_1(x), h_2(x), \ldots, h_D(x)$. Entonces, nuestro modelo se ve así:

$$y = f(x; \boldsymbol{\theta}) = b + w_1 h_1(x) + w_2 h_2(x) + \ldots + w_D h_D(x)$$

Aquí $h_j(x)$ son las funciones de base y pueden ser cualquier cosa. Casos particulares son:

- Regresión polinomial: cuando $h_j(x) = x^j$.
- Regresión trigonométrica: cuando $h_j(x) = \cos(jx)$ o $h_j(x) = \sin(jx)$.

Lo que es "lineal" acerca de este modelo es la relación entre y y los coeficientes $\boldsymbol{\theta} = (b, \boldsymbol{w})$. Por eso aún se llama regresión lineal, a pesar de que la relación entre y y x puede ser altamente no lineal.

Considerando la matriz de diseño

$$\boldsymbol{X} = \begin{bmatrix} 1 & h_1(x_1) & \cdots & h_D(x_1) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & h_1(x_N) & \cdots & h_D(x_N) \end{bmatrix}$$

el procedimiento es exactamente igual que en el caso multivariado.

11. Regularización Ridge

La regularización Ridge (inglés significa cresta) consiste en agregar un termino a la función de pérdida que de algún modo penalice el tamaño del vector de pesos $\boldsymbol{w}$:

$$L_\alpha(\boldsymbol{w}) = L(\boldsymbol{w}) + \alpha \|\boldsymbol{w}\|^2 = L(\boldsymbol{w}) + \alpha \sum_{j=1}^{D} w_j^2$$

El objetivo es evitar los problemas de la multicolinealidad.

De hecho, podemos calcular explícitamente el valor del parámetro $\widehat{w}_\alpha$ del mismo modo que hicimos antes. Escribiendo la función de pérdida como

$$L_\alpha(\boldsymbol{w}) = \frac{1}{N} \boldsymbol{w}^\top \boldsymbol{Z}^\top \boldsymbol{Z} \boldsymbol{w} - \frac{2}{N} \boldsymbol{w}^\top \boldsymbol{Z}^\top \boldsymbol{u} + \frac{1}{N} \boldsymbol{u}^\top \boldsymbol{u} + \alpha \boldsymbol{w}^\top \boldsymbol{w}$$

derivamos e igualamos a cero:

$$\frac{\partial L_\alpha}{\partial \boldsymbol{w}} = \frac{2}{N}\boldsymbol{Z}^\top \boldsymbol{Z}\boldsymbol{w} - \frac{2}{N}\boldsymbol{Z}^\top \boldsymbol{u} + 2\alpha \boldsymbol{w} = 0$$

$$\left(\frac{1}{N}\boldsymbol{Z}^\top \boldsymbol{Z} + 2\alpha I\right)\boldsymbol{w} = \frac{1}{N}\boldsymbol{Z}^\top \boldsymbol{u}$$

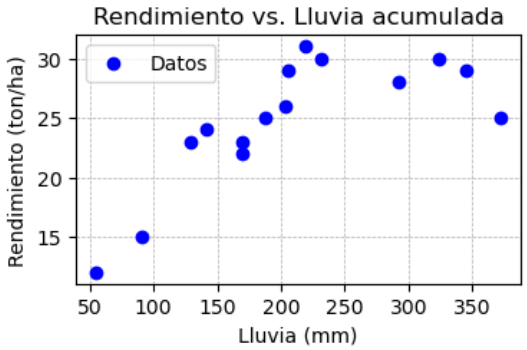
$$\widehat{\boldsymbol{w}}_\alpha = \left(\operatorname{cov}\left(\left[\boldsymbol{x}^{(j)}\right]\right) + \alpha I\right)^{-1} \operatorname{cov}\left(\left[\boldsymbol{x}^{(j)}\right], \boldsymbol{y}\right)$$

Notar el efecto de α en acercar proporcionalmente la matriz de covarianzas a la identidad, lo cual mejora su invertibilidad.

12. Multicolinealidad en regresión polinomial

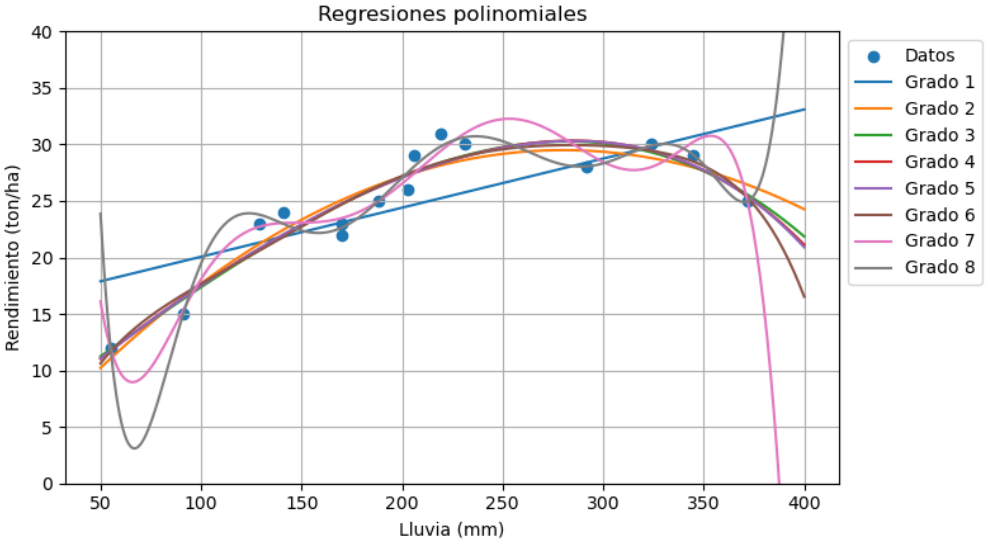
A modo de ilustración, considerar el siguiente conjunto de datos del rendimiento de un cultivo de papas en función de la lluvia acumulada:

x = Lluvia (mm)	y = Rendimiento (ton/ha)
206	29
188	25
219	31
372	25
345	29
231	30
203	26
170	23
55	12
91	15
292	28
141	24
129	23
170	22
324	30



Al mirar el gráfico vemos que la relación entre x e y no es lineal y parece razonable intentar con una regresión polinomial. Inmediatamente surge la pregunta de elegir el grado del polinomio.

El siguiente gráfico muestra varios polinomios, con grados que van desde 1 a 8, ajustados a estos datos:



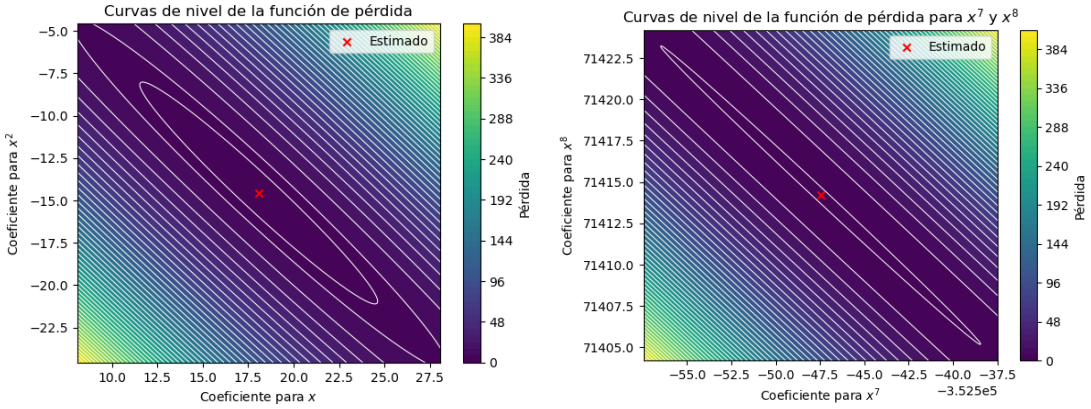
Para ajustar dichos polinomios siempre es conveniente estandarizar la matriz de diseño.

Observar la matriz de correlaciones para grado 8:

Cor	x^1	x^2	x^3	x^4	x^5	x^6	x^7	x^8
x^1	1	0.977	0.936	0.897	0.863	0.834	0.808	0.784
x^2	0.977	1	0.989	0.969	0.946	0.924	0.902	0.881
x^3	0.936	0.989	1	0.994	0.982	0.967	0.951	0.934
x^4	0.897	0.969	0.994	1	0.996	0.988	0.977	0.964
x^5	0.863	0.946	0.982	0.996	1	0.998	0.991	0.982
x^6	0.834	0.924	0.967	0.988	0.998	1	0.998	0.993
x^7	0.808	0.902	0.951	0.977	0.991	0.998	1	0.998
x^8	0.784	0.881	0.934	0.964	0.982	0.993	0.998	1

Más aún, el determinante de dicha matriz es $5{,}14 \times 10^{-39}$, y por lo tanto estamos ante la presencia de una marcada multicolinealidad, y por ende a riesgo de coeficientes grandes.

La multicolinealidad se puede visualizar al graficar las curvas de nivel de la función de pérdida. A modo de ejemplo:



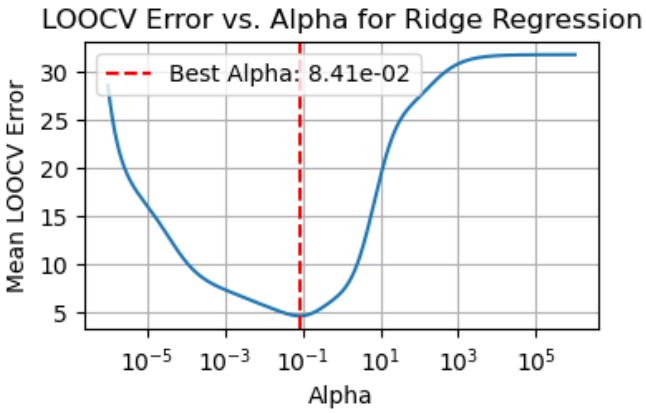
Las elipses estiradas (con los datos estandarizados) indican alta correlación entre los atributos.

Y de hecho los coeficientes para los distintos grados tienden a crecer rápidamente:

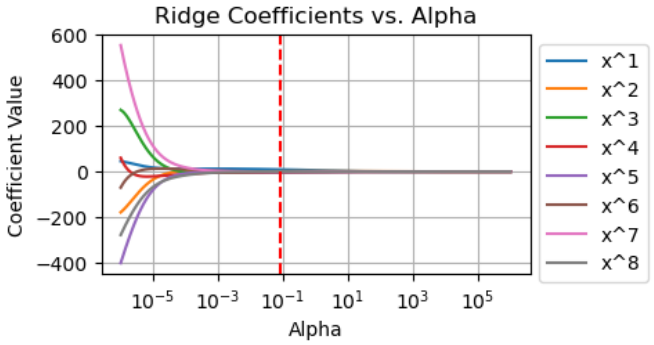
Coeficiente	Grado 1	Grado 2	Grado 3	Grado 4	Grado 5	Grado 6	Grado 7	Grado 8
w_1	3.85	18.08	10.79	15.02	12.71	76.59	-1533.75	-5491.59
w_2		-14.56	3.56	-13.02	-0.23	-453.11	13365.95	52990.47
w_3			-11.13	10.75	-16.30	1318.39	-50622.38	-231805.88
w_4				-9.49	15.63	-1950.95	102868.06	571182.95
w_5					-8.56	1419.35	-116379.46	-841240.02
w_6						-406.93	69088.87	735492.68
w_7							-16787.07	-352547.47
w_8								71414.21

Para controlar el tamaño de los coeficientes podemos correr una regresión polinomial de grado 8 regularizada. El problema se traslada ahora en elegir el valor de α.

El gráfico a continuación muestra la curva de error para varios valores de α, el error calculado usando la técnica de Leave One Out Cross Validation (LOOCV):



También podemos ver la evolución de los coeficientes en función de α:



Por último, podemos ver el efecto de la regularización en los coeficientes:

Coef	**Original**	**Ridge**
x^1	-5491.591405	9.290364281
x^2	52990.4687	1.239116016
x^3	-231805.8798	-2.013422183
x^4	571182.9466	-2.683297203
x^5	-841240.0155	-2.179430752
x^6	735492.6833	-1.218635966
x^7	-352547.4682	-0.16433541
x^8	71414.20573	0.79483848

Aquí tenemos el gráfico de la regresión regularizada: