Trabajo Práctico 1 - Regresión Lineal Múltiple (Métodos Computacionales)

Maximiliano Runnacles y Juan Ignacio Elosegui

Mayo 2025

Primera Parte

a) El espacio columna como subespacio vectorial

Queremos probar que:

$$\text{Col}(X) = \{b \in \mathbb{R}^n / b = X\beta, \text{ con } \beta \text{ variando en } \mathbb{R}^p\}$$

Demostremos que es un subespacio

1. ¿$\vec{0} \in \text{Col}(X)$?
   Si $\beta = \vec{0} \in \mathbb{R}^p$, ¿entonces $X\beta = \vec{0} \in \mathbb{R}^n$?

$$\vec{0} = X \cdot \vec{0} \Rightarrow$$

$$\therefore \vec{0} \in \text{Col}(X)$$

2. Si $\vec{u}, \vec{v} \in \text{Col}(X)$, ¿entonces $\vec{u} + \vec{v} \in \text{Col}(X)$?

$$\vec{u} = X\beta_1 \wedge \vec{v} = X\beta_2 \Rightarrow$$

$$\vec{u} + \vec{v} = X\beta_1 + X\beta_2 \Rightarrow$$

$$\vec{u} + \vec{v} = X(\beta_1 + \beta_2) \Rightarrow$$

$$\therefore \vec{u} + \vec{v} \in \text{Col}(X)$$

3. Si $\vec{u} \in \text{Col}(X)$, ¿$k\vec{u} \in \text{Col}(X)$?

$$\vec{u} = X\beta_1 \wedge k \in \mathbb{R} \Rightarrow$$

$$k\vec{u} = kX\beta_1 \Rightarrow \in \text{Col}(X)$$

$$k\vec{u} = X(k\beta_1) \Rightarrow$$

$$\therefore k\vec{u} \in Col(X)$$

b) Producto escalar como producto matricial

Queremos mostrar que el **producto escalar** entre dos vectores columna $u, v \in \mathbb{R}^n$ puede expresarse como:

$$u \cdot v = v^T \cdot u$$

Recordemos la definición del producto escalar:

$$u \cdot v = \sum_{i=1}^{n} u_i v_i$$

Si tratamos a u y v como vectores columna de dimensión $n \times 1$, entonces:

- v^T es un vector fila de dimensión $1 \times n$
- El producto matricial $v^T \cdot u$ da como resultado un escalar:

$$v^T \cdot u = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = \sum_{i=1}^{n} v_i \cdot u_i = u \cdot v$$

Por la conmutatividad del producto escalar, también se cumple:

$$u \cdot v = u^T \cdot v$$

c) Aplicación del teorema de proyección ortogonal

Queremos aplicar el **teorema de proyección ortogonal** para deducir una condición de optimalidad para la regresión lineal.

Buscamos el vector $\beta^* \in \mathbb{R}^p$ tal que:

$$\|y - X\beta^*\| = \min_{\beta \in \mathbb{R}^p} \|y - X\beta\|$$

El teorema establece que el mejor vector $b \in \text{Col}(X)$ que aproxima a $y \in \mathbb{R}^n$ es aquel tal que el error $y - b$ es ortogonal a todo vector del subespacio $\text{Col}(X)$.

Como en nuestro caso $b = X\beta^*$, esta condición se convierte en:

$$(y - X\beta^*) \cdot \vec{s} = 0, \forall \vec{s} \in \text{Col}(X)$$

Dado que cualquier vector $\vec{s} \in \text{Col}(X)$ puede escribirse como $\vec{s} = X\vec{z}$ con $\vec{z} \in \mathbb{R}^p$, podemos reescribir la condición como:

$$(y - X\beta^*) \cdot X\vec{z} = 0, \forall \vec{z} \in \mathbb{R}^p$$

Esta es la condición de ortogonalidad que nos proporciona una ecuación para hallar la solución óptima β^*.

d) Aplicación del producto escalar

A partir de la condición obtenida en el inciso anterior:

$$(y - X\beta^*) \cdot X\vec{z} = 0, \forall \vec{z} \in \mathbb{R}^p$$

Aplicamos la propiedad del producto escalar demostrada en el punto (b), que nos permite escribir:

$$(y - X\beta^*) \cdot X\vec{z} = (X\vec{z})^T (y - X\beta^*)$$

Usamos que $(X\vec{z})^T = \vec{z}^T X^T$, y obtenemos:

$$(X\vec{z})^T (y - X\beta^*) = \vec{z}^T X^T (y - X\beta^*) = 0, \forall \vec{z} \in \mathbb{R}^p$$

Como esto se cumple para cualquier z, concluimos que el vector $X^\top (y - X\beta^*)$ debe ser el vector nulo:

$$X^T (y - X\beta^*) = 0$$

Esta es una ecuación clave en la deducción de la solución óptima del problema de mínimos cuadrados.

e) Deducción de la ecuación normal

Partimos de la expresión obtenida en el punto anterior:

$$X^T (y - X\beta^*) = 0$$

Distribuyendo el producto matricial, obtenemos:

$$X^T y - X^T X\beta^* = 0$$

Reordenando los términos, llegamos a la llamada **ecuación normal**:

$$X^T X\beta^* = X^T y$$

Este sistema lineal es fundamental en la resolución del problema de regresión lineal por mínimos cuadrados. Además, el enunciado nos recuerda que, si un vector $\vec{u} \in \mathbb{R}^n$ es ortogonal a todos los vectores $\vec{v} \in \mathbb{R}^n$, entonces necesariamente $\vec{u} = \vec{0}$. Esto justifica que, como:

$$\vec{z}^T X^T (y - X\beta^*) = 0, \forall \vec{z} \in \mathbb{R}^p$$

entonces debe cumplirse que:

$$X^T (y - X\beta^*) = 0$$

y, por lo tanto, la ecuación normal es válida.

f) Fórmula explícita para la solución óptima

Partimos de la ecuación normal obtenida anteriormente:

$$X^T X\beta^* = X^T y$$

Si asumimos que las columnas de $X \in \mathbb{R}^{n\times p}$ son linealmente independientes, entonces la matriz $X^T X \in \mathbb{R}^{p\times p}$ es invertible.

Podemos entonces multiplicar ambos lados de la ecuación por $(X^T X)^{-1}$, obteniendo:

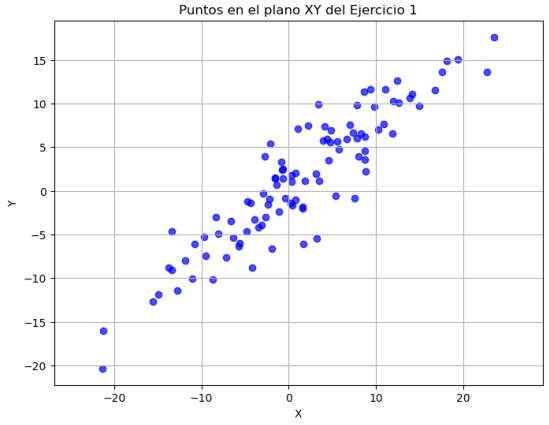
$$\beta^* = (X^T X)^{-1} X^T y$$

Con esta fórmula podemos calcular la solución óptima β^*.

Segunda Parte

Ejercicio 1

Primero graficamos todos los punto en el plano XY de los datos del archivo `ejercicio_1.csv`, y luego creamos una recta a partir de la fórmula que encontramos en la Primera Parte.

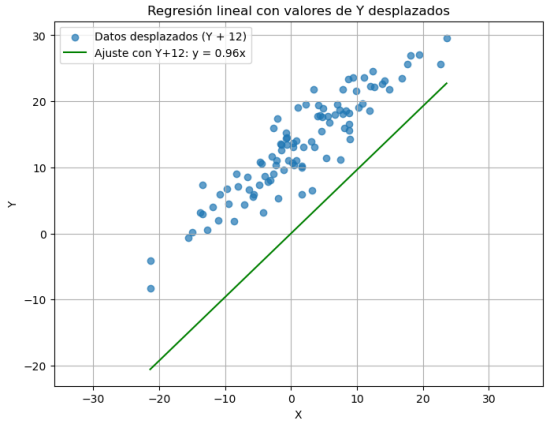


C) Regresión con valores de Y desplazados

En este paso modificamos cada valor de la variable dependiente Y sumándole 12 unidades:

$$Y_i^{\text{nuevo}} = Y_i + 12, \forall i = 1, \ldots, n$$

Luego volvemos a aplicar la misma fórmula para obtener los parámetros β_0 y β_1, esta vez sobre el nuevo conjunto de valores $Y_{\text{desplazado}}$:



$$\beta^* = (X^T X)^{-1} X^T Y_{\text{desplazado}}$$

Como la matriz X es la misma, el resto del procedimiento no cambia.

Podemos ver que el nuevo ajuste es una recta paralela a la anterior, pero con una ordenada al origen menor. Y también podemos ver un gran problema: la nueva recta no cruza ninguno de los datos. No es una buena aproximación.

D) ¿Cómo se puede extender el modelo para ajustarse a cualquier recta?

Para poder aproximar cualquier recta en el plano, el modelo debe incluir un término independiente β_0. Para esto armamos una columna de unos para la matriz X tal que:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

De esta manera, el modelo ajustado puede tener cualquier ordenada al origen, y no solo aquellas rectas que pasan por el punto $(0, 0)$, ya que la recta se desplazaria a la par de Y.

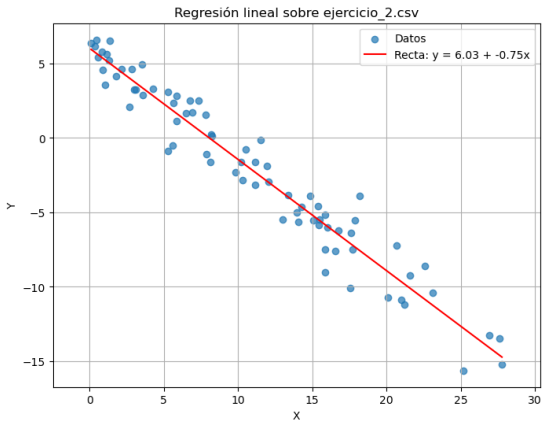
Si no hiciéramos esta modificación, el modelo no podría adapatarse a los valores desplazados de Y. La recta ajustada no cambia, por lo que tendría errores mayores con cualquier desplazamiento.

Ejercicio 2

En este ejercicio trabajamos con los datos contenidos en `ejercicio_2.csv`.

Aplicamos el mismo procedimiento que en el ejercicio anterior, construyendo la matriz X con una columna de unos y otra con los valores de la variable X, y utilizando la fórmula que encontramos para calcular el vector β^*:

$$\beta^* = (X^T X)^{-1} X^T Y$$

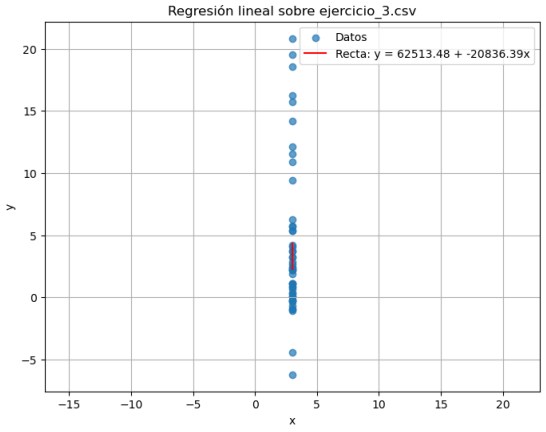


¿Es buena la aproximación? ¿Cual fue el problema?

Nuestra aproximación nos da una recta que intenta ajustar los datos, pero observamos que, si bien busca el promedio de los datos, varios quedaron alejados de la recta (sobre todo para los valores de Y mas negativos). No creemos que es una mala aproximación para este caso, pero ya nos revela algunos problemas que puede tener este modelo. Al usar una recta para la aproximación, estamos suponiendo que los datos tienen una relación lineal (siguen una tendencia lineal), lo cual lo hace poco flexible si los datos tuvieran una relación mas compleja. Si los datos formaran parte de mediciones, como la velocidad del viento a lo largo del tiempo, y quisieramos predecir su velocidad en el futuro, esta aproximación podría darnos predicciones erróneas. La velocidad del tiempo no va a seguir una tendencia lineal, por lo que nuestro modelo actual resultaría muy limitado.

Ejercicio 3: Problemas de singularidad en la regresión

En este ejercicio aplicamos la misma fórmula que venimos usando, ahora sobre los datos del `ejercicio_3.csv`. Sin embargo, tuvimos un resultado peculiar:



¿Qué ocurre y por qué? ¿Y los ejercicios anteriores?

Al intentar calcular la inversa de $X^\top X$, notamos que su determinante es extremadamente cercano a cero:

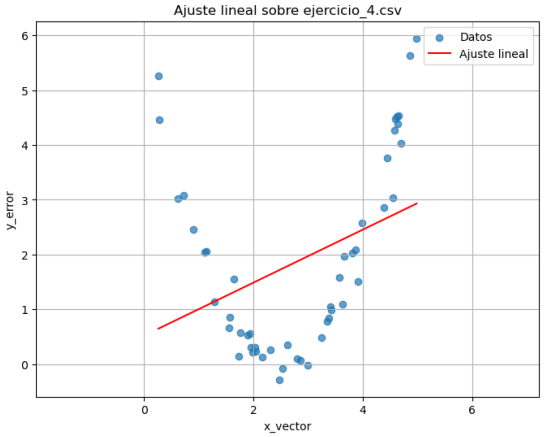
$$\det(X^T X) \approx 0$$

Lo cual nos indica que no podemos invertir de forma estable la matriz. Esto se debe a que los valores de X son casi todas idénticos (valen "3", salvo una que vale "2,999"). Entonces, a la hora de crear la matriz con su columna de unos y la columna de X, estas dos columnas se vuelven linealmente dependientes. Cualquier intento de invertir la matriz nos puede dar errores numéricos o resultados inestables, llevando al grafico que observamos ahora.

En casos anteriores, las variables X tenían mas variabilidad, por lo que podíamos calcular β^* y crear un modelo lineal dentro de todo confiable. Pero, en este caso, la falta de variabilidad impide que $X^T X$ sea invertible (no de una manera estable), complicando el modelo.

Ejercicio 4: Ajuste no lineal y sobreajuste

Repetimos los mismos pasos de los ejercicios anteriores y graficamos.



Podemos notar que estamos trabajando con datos cuyo comportamiento no es lineal: los puntos en el plano (x, y) forman una parábola. La aproximacion lineal que estábamos usando no nos servirá en este caso, por lo que debemos explorar otros modelos. Por ejemplo, podríamos intentar usar una función cuadrática del tipo:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

Con esta nueva aproximación, vemos que el nuevo modelo es mucho mejor y redujo el error con respecto al modelo lineal. Luego, acorde a lo que nos pide el ejercicio, decidimos observar que ocurre si repetimos el proceso con un polinomio de grado 10. (aca va el grafico) Observamos que la recta se va curvando para intentar lo mas cerca posible de los datos. Si bien no es por una enorme diferencia, notamos que esta nueva aproximación reduce el error (la recta está más cerca de más puntos) y es coherente con los datos que tenemos.

Tercera Parte

Ejercicio 1

Utilizamos las primeras 450 observaciones del conjunto de datos `student_performance.csv` para entrenar un modelo de regresión lineal múltiple, y calculamos los β, el $\hat{y}$ y el ECM.

a) Estimación de los β

Los calculamos con el modelo que habíamos descubierto antes:

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T y$$

Y agregando una columna de unos a la matriz X para incluir el término independiente.

Con esto obtuvimos:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} -34{,}4703 \\ 2{,}8671 \\ 1{,}0235 \\ 0{,}5578 \\ 0{,}4623 \\ 0{,}2069 \end{bmatrix}$$

b) Estimación de $\hat{y}$

Con los β estimados, calculamos las predicciones del modelo:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_5 x_{i5} \quad \text{para } i = 1, \ldots, 450$$

El resultado es un vector de 450 predicciones. A modo ilustrativo, mostramos los primeros cinco valores estimados de $\hat{y}$:

$$\hat{y}_1 = 91{,}85$$
$$\hat{y}_2 = 63{,}19$$
$$\hat{y}_3 = 44{,}87$$
$$\hat{y}_4 = 36{,}37$$
$$\hat{y}_5 = 67{,}10$$

c) Error cuadrático medio

Siguiendo la fórmula del ECM:

$$\text{ECM} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Para los 450 estudiantes (que pertenecen a los datos de entrenamiento), el ECM fue:

$$\text{ECM}_{\text{entrenamiento}} = 3{,}6009$$

Este valor muestra que, en promedio, las predicciones del modelo difieren en aproximadamente $\sqrt{3{,}6009} \approx 1{,}90$ unidades del valor real.

Ejercicio 2: A)

Ajustamos (en el Ejercicio 1C) el modelo de regresión lineal múltiple sobre los primeros 450 estudiantes, y obtuvimos un Error Cuadrático Medio (ECM) de:

$$\text{ECM}_{\text{entrenamiento}} = 3{,}6009$$

Luego, aplicamos el modelo sobre los 150 estudiantes restantes (conjunto de test), y obtuvimos:

$$\text{ECM}_{\text{test}} = 4{,}4364$$

La diferencia entre ambos es moderada. Como el modelo lo ajustamos para minimizar el error sobre el conjunto de entrenamiento, tiene sentido que el error aumente un poco (siendo en datos "no vistos"). Sin embargo, como no es una gran diferencia, podemos concluir que el modelo logra generalizar correctamente a nuevos casos.

Este resultado sugiere que las variables que usamos tienen una relación estable con la variable objetivo (Y).

B)

Luego de entrenar el modelo utilizando los 600 alumnos (y generar asi un nuevo $\hat{\beta}$), obtuvimos el siguiente ECM:

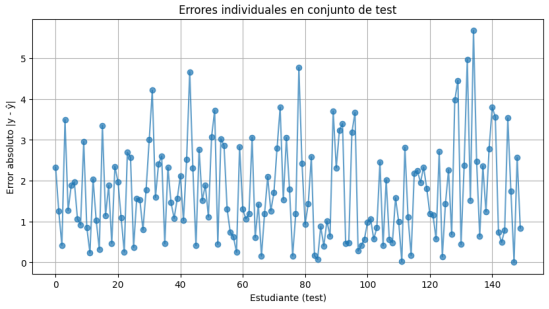
$$\text{ECM}_{\text{test con modelo de 600 datos}} = 4{,}3331$$

Este valor es ligeramente menor que el ECM que obtuvimos antes con el modelo entrenado sobre 450 alumnos (4,4364). Al incorporar más datos, el modelo está capturando mejor las relaciones entre las variables, mejorando así su capacidad predictiva.

Ejercicio 3

Graficamos el error absoluto cometido por cada estudiante (tomando en cuenta los "tests"solamente):

$$\text{error}_i = |\hat{y}_i - y_i|$$



La mayoría de los errores individuales se encuentran en un rango reducido (por debajo de 3 unidades), lo que indica que el modelo realiza buenas predicciones en general. Sin embargo, observamos que hay varios casos con errores mayores (entre ellos, varios entre 3 y 4, y algunos incluso más de 4).

Ejercicio 4

Agregar una nueva columna a los datos, como una que diga el año de egreso del estudiante, al modelo **nunca aumentará el ECM**: puede disminuir o quedarse igual. Al agregar un nuevo β, el modelo tiene más libertad para ajustarte a los datos.

Sin embargo, esto **no garantiza que el ECM disminuya**. Si esta nueva variable no tiene correlación con las otras variables, puede introducir **ruido o redundancia**, lo cual empeoraría la capacidad del modelo para generalizar.

La conclusión sería que no aumentaría el ECM, pero no necesariamente lo disminuiría. No tenemos una garantía de que una nueva columna mejore el modelo. Tendríamos que ajustar el modelo con esta nueva variable y comparar los ECM.