2 El problema de regresión y la regresión lineal

El primer problema que estudiaremos es el *problema de regresión*. La regresión es uno de los dos problemas principales que abordamos en estas notas (el otro es la clasificación). El primer método que encontraremos es la *regresión lineal*, que es una (de muchas) soluciones al problema de regresión. A pesar de la relativa simplicidad de la regresión lineal, es sorprendentemente útil y constituye además un bloque fundamental para métodos más avanzados (como el aprendizaje profundo, en el Capítulo 7).

2.1 El problema de regresión

La regresión se refiere al problema de aprender la relación entre algunas variables de entrada $\mathbf{x} = [x_1 \; x_2 \; \ldots \; x_p]^\top$ (cualitativas o cuantitativas[1]) y una variable de salida cuantitativa y. En términos matemáticos, la regresión consiste en aprender un modelo f tal que

$$y = f(\mathbf{x}) + \varepsilon, \tag{2.1}$$

donde ε es un término de ruido/error que describe todo aquello que no puede ser capturado por el modelo. Desde nuestra perspectiva estadística, consideramos que ε es una variable aleatoria independiente de $\mathbf{x}$ y con valor esperado igual a cero.

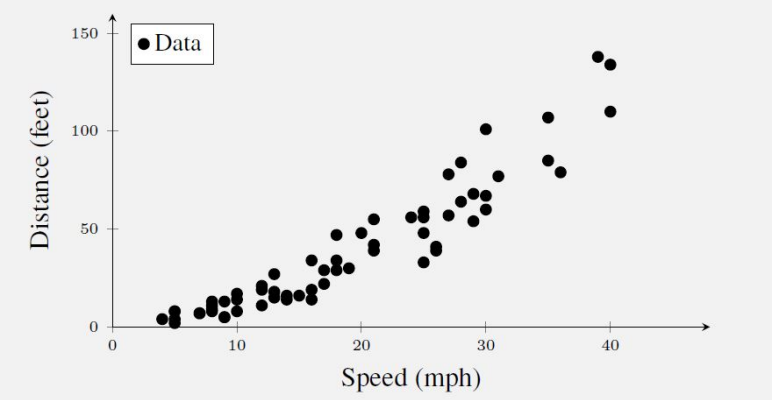
A lo largo de este capítulo, utilizaremos el conjunto de datos introducido en el Ejemplo 2.1 sobre las distancias de frenado de automóviles para ilustrar la regresión. En una frase, el problema consiste en aprender un modelo de regresión que pueda predecir qué distancia se necesita para que un automóvil se detenga por completo, dada su velocidad actual.

Ejemplo 2.1: Distancias de frenado de automóviles

Ezekiel y Fox (1959) presentan un conjunto de datos con 62 observaciones sobre la distancia necesaria para que distintos automóviles, partiendo de distintas velocidades iniciales, logren detenerse por completo.[a] El conjunto de datos contiene las siguientes dos variables:

- **Speed**: La velocidad del automóvil cuando se da la señal de frenar.
- **Distance**: La distancia recorrida desde que se da la señal hasta que el automóvil se detiene por completo.

Decidimos interpretar **Speed** como la **variable de entrada x**, y **Distance** como la **variable de salida** y.



Nuestro objetivo es usar regresión lineal para estimar (es decir, *predecir*) cuánto tiempo tomaría la distancia de frenado si la velocidad inicial fuera de 33 mph o 45 mph (dos velocidades para las cuales no se han registrado datos).

[a] El conjunto de datos es algo antiguo, por lo que las conclusiones quizá no sean aplicables a automóviles modernos. Sin embargo, creemos que el lector puede suponer que los datos provienen de su ejemplo favorito.

2.2 El modelo de regresión lineal

El modelo de regresión lineal describe la variable de salida y (un escalar) como una combinación afín de las variables de entrada $x_1, x_2, \ldots, x_p$ (cada una escalar) más un término de error ε:

$$y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}_{f(\mathbf{x};\boldsymbol{\beta})} + \varepsilon. \tag{2.2}$$

Nos referimos a los coeficientes $\beta_0, \beta_1, \ldots, \beta_p$ como los *parámetros* del modelo, y a veces nos referimos específicamente a β_0 como el término independiente. El término de error ε representa los errores no sistemáticos, es decir, aleatorios, entre los datos y el modelo. Se asume que el error tiene media cero y es independiente de $\mathbf{x}$.

El aprendizaje automático trata sobre entrenar, o aprender, modelos a partir de datos. Por lo tanto, gran parte de este capítulo se dedicará a aprender los parámetros $\beta_0, \beta_1, \ldots, \beta_p$ a partir de un conjunto de entrenamiento $\mathcal{T} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$.

Antes de profundizar en los detalles en la Sección 2.3, comencemos discutiendo brevemente el propósito de usar regresión lineal. Este modelo puede usarse, al menos, para dos propósitos diferentes:

- Para **describir** las relaciones en los datos interpretando los parámetros $\boldsymbol{\beta} = [\beta_0 \ \beta_1 \ \ldots \ \beta_p]^\top$. - Para **predecir** salidas futuras para entradas que aún no hemos visto.

Observación 2.1

Es posible formular el modelo también para múltiples salidas $y_1, y_2, \ldots$, ver los ejercicios. Esto se conoce comúnmente como *regresión lineal multivariada*.

2.2.1 Describir relaciones — estadística clásica

Una pregunta común en ciencias como la medicina o la sociología es determinar si existe correlación entre algunas variables, por ejemplo: "¿vives más si solo comes mariscos?".

Estas preguntas pueden abordarse estudiando los parámetros $\boldsymbol{\beta}$ del modelo de regresión lineal, una vez que el modelo ha sido aprendido a partir de los datos. La pregunta más común es si se puede indicar que existe *alguna* correlación entre dos variables x_1 e y, lo cual se puede hacer con el siguiente razonamiento:

Si $\beta_1 = 0$, indicaría que no hay correlación entre y y x_1 (a menos que las demás variables también dependan de x_1). Estimando β_1 junto con un intervalo de confianza (que describe la incertidumbre de la estimación), se puede descartar (con cierto nivel de significancia) que x_1 e y no están correlacionadas si 0 no está dentro del intervalo de confianza para β_1.

La conclusión entonces es que probablemente existe alguna correlación entre x_1 e y. Este tipo de razonamiento se conoce como *prueba de hipótesis* y constituye una parte importante de la estadística clásica. Sin embargo, aquí nos concentraremos en otro propósito del modelo de regresión lineal: hacer predicciones.

2.2.2 Predecir salidas futuras — aprendizaje automático

En el aprendizaje automático, el énfasis está en predecir alguna salida (aún no observada) $\hat{y}_*$ para una nueva entrada $\mathbf{x}_* = [x_{*1} \ x_{*2} \ \ldots \ x_{*p}]^\top$.

Para hacer una predicción para la entrada de prueba $\mathbf{x}_*$, la insertamos en el modelo (2.2). Como ε (por hipótesis) tiene valor esperado cero, tomamos la **predicción** como

$$\hat{y}_* = \beta_0 + \beta_1 x_{*1} + \beta_2 x_{*2} + \cdots + \beta_p x_{*p}. \tag{2.3}$$

Usamos el símbolo $\hat{}$ sobre y_* para indicar que es una predicción, nuestra mejor suposición. Si de alguna manera pudiéramos observar el valor real de salida para $\mathbf{x}_*$, lo denotaríamos como y_* (sin sombrero).

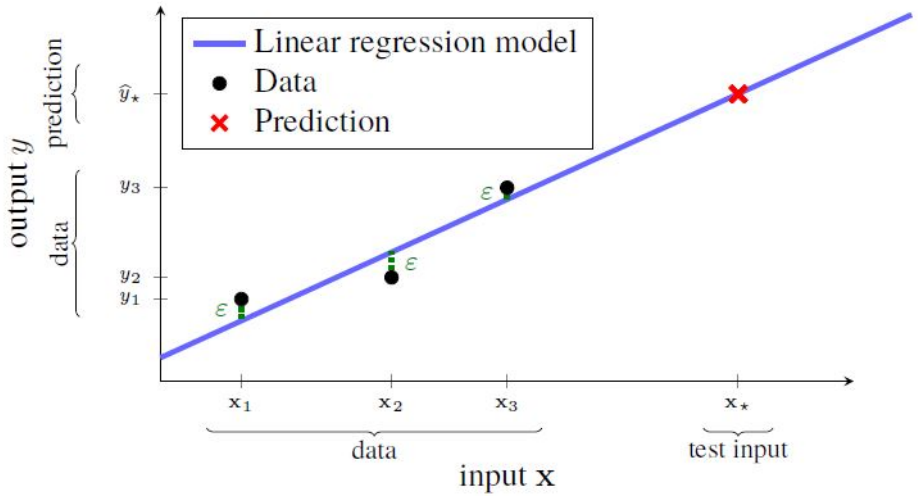


Figura 1: **Regresión lineal con $p = 1$:** los puntos negros representan $n = 3$ puntos de datos, a partir de los cuales se aprende un modelo de regresión lineal (línea azul). El modelo no se ajusta perfectamente a los datos, pero queda un error/ruido ε (en verde). El modelo puede usarse para *predecir* (cruz roja) la salida $\hat{y}_*$ para un punto de entrada de prueba $\mathbf{x}_*$.

2.3 Aprender el modelo a partir de los datos de entrenamiento

Para usar el modelo de regresión lineal, primero necesitamos aprender los parámetros desconocidos $\beta_0, \beta_1, \ldots, \beta_p$ a partir de un conjunto de datos de entrenamiento $\mathcal{T}$. Este conjunto de entrenamiento consiste en n muestras de la variable de salida y, que denotamos como y_i (con $i = 1, \ldots, n$), y las correspondientes n muestras de entrada $\mathbf{x}_i$ (también con $i = 1, \ldots, n$), donde cada $\mathbf{x}_i$ es un vector columna.

Escribimos el conjunto de datos en forma matricial como:

$$\mathbf{X} = \begin{bmatrix} 1 & \mathbf{x}_1^\top \\ 1 & \mathbf{x}_2^\top \\ \vdots & \vdots \\ 1 & \mathbf{x}_n^\top \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \text{donde cada } \mathbf{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix}. \tag{2.4}$$

Observa que $\mathbf{X}$ es una matriz de tamaño $n \times (p+1)$, y $\mathbf{y}$ es un vector n-dimensional. La primera columna de $\mathbf{X}$, compuesta solo por unos, corresponde al término independiente β_0 en el modelo de regresión lineal (2.2).

Si además agrupamos los parámetros desconocidos $\beta_0, \beta_1, \ldots, \beta_p$ en un vector de dimensión $(p+1)$:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \tag{2.5}$$

entonces podemos expresar el modelo de regresión lineal como una multiplicación matricial:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.6}$$

donde $\boldsymbol{\varepsilon}$ es un vector de errores o ruido.

Aprender los parámetros desconocidos $\boldsymbol{\beta}$ equivale a encontrar aquellos valores tales que *el modelo se ajuste bien a los datos*. Hay múltiples formas de definir qué significa exactamente .ªjustarse bien". Nosotros adoptaremos una perspectiva estadística y elegiremos el valor de $\boldsymbol{\beta}$ que haga que los datos observados de entrenamiento $\mathbf{y}$ sean lo más probables posible bajo el modelo —lo que se conoce como la **solución de máxima verosimilitud** (*maximum likelihood*).

Ejemplo 2.2: Distancias de frenado de automóviles

Continuamos con el Ejemplo 2.1 y formamos las matrices $\mathbf{X}$ y $\mathbf{y}$. Como solo tenemos una variable de entrada y una de salida, tanto x_i como y_i son escalares. Obtenemos:

$$\mathbf{X} = \begin{bmatrix} 1 & 4 \\ 1 & 5 \\ 1 & 5 \\ 1 & 5 \\ 1 & 7 \\ 1 & 7 \\ 1 & 8 \\ \vdots & \vdots \\ 1 & 39 \\ 1 & 39 \\ 1 & 40 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 4 \\ 2 \\ 4 \\ 8 \\ 8 \\ 7 \\ 7 \\ 8 \\ \vdots \\ 138 \\ 110 \\ 134 \end{bmatrix}. \tag{2.7}$$

2.3.1 Máxima verosimilitud

Nuestra estrategia para aprender los parámetros desconocidos $\boldsymbol{\beta}$ a partir del conjunto de entrenamiento $\mathcal{T}$ será el método de *máxima verosimilitud*. El término "verosimilitud" hace referencia al concepto estadístico de la función de verosimilitud, y maximizar dicha función equivale a encontrar el valor de $\boldsymbol{\beta}$ que haga que observar $\mathbf{y}$ sea lo más probable posible bajo el modelo.

Es decir, queremos resolver:

$$\max_{\boldsymbol{\beta}} \; p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}), \tag{2.8}$$

donde $p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta})$ es la densidad de probabilidad de los datos $\mathbf{y}$ dada una cierta elección de los parámetros $\boldsymbol{\beta}$. Denotamos la solución de este problema —los parámetros aprendidos— como $\hat{\boldsymbol{\beta}} = [\hat{\beta}_0 \; \hat{\beta}_1 \; \cdots \; \hat{\beta}_p]^\top$. De forma más compacta, escribimos esto como:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}). \tag{2.9}$$

Para tener una noción matemática de qué significa "lo más probable", necesitamos hacer suposiciones sobre el término de error ε. Una suposición común es que ε sigue una distribución gaussiana con media cero y varianza σ_ε^2:

$$\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2). \tag{2.10}$$

Esto implica que la función de densidad de probabilidad condicional de la salida y para un valor dado de la entrada $\mathbf{x}$ está dada por:

$$p(y \mid \mathbf{x}, \boldsymbol{\beta}) = \mathcal{N}(y \mid \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \; \sigma_\varepsilon^2). \tag{2.11}$$

Además, se asume que los n puntos observados de entrenamiento son realizaciones independientes según este modelo estadístico. Esto implica que la verosimilitud de los datos de entrenamiento se factoriza como:

$$p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}) = \prod_{i=1}^{n} p(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}). \tag{2.12}$$

Combinando (2.11) y (2.12) obtenemos:

$$p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}) = \frac{1}{(2\pi\sigma_\varepsilon^2)^{n/2}} \exp\left(-\frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^{n} (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} - y_i)^2 \right). \tag{2.13}$$

Recordando (2.8), queremos maximizar la verosimilitud con respecto a $\boldsymbol{\beta}$. Sin embargo, dado que (2.13) depende solo de $\boldsymbol{\beta}$ a través de la suma en el exponente, y la exponencial es una función monótonamente creciente, maximizar (2.13) es equivalente a minimizar:

$$\sum_{i=1}^{n}(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} - y_i)^2. \tag{2.14}$$

Esta es la suma de los cuadrados de las diferencias entre cada dato de salida y_i y la predicción del modelo para esa salida, $\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$. Por esta razón, minimizar (2.14) suele denominarse *mínimos cuadrados*.

Volveremos más adelante a cómo calcular los valores $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$. Solo mencionamos aquí que también es posible —y a veces muy conveniente— suponer que la distribución de ε no es gaussiana. Por ejemplo, podríamos asumir que ε sigue una distribución de Laplace, lo que llevaría a la siguiente función de coste:

$$\sum_{i=1}^{n}|\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} - y_i|. \tag{2.15}$$

Contiene la suma de los valores absolutos de todas las diferencias (en lugar de sus cuadrados). El principal beneficio de la suposición gaussiana (2.10) es que existe una solución en forma cerrada para los parámetros $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, mientras que otras suposiciones sobre ε usualmente requieren métodos computacionalmente más costosos.

Observación 2.2

Con la terminología que introduciremos en el próximo capítulo, podríamos referirnos a (2.13) como la *función de verosimilitud*, que denotaremos por $\ell(\boldsymbol{\beta})$.

Observación 2.3

No es raro en la literatura omitir la motivación por máxima verosimilitud y simplemente presentar (2.14) como una *función de coste* (algo arbitraria) para la optimización.

2.3.2 Mínimos cuadrados y las ecuaciones normales

Si asumimos que el ruido/error ε tiene una distribución gaussiana como en (2.10), los parámetros de máxima verosimilitud $\hat{\boldsymbol{\beta}}$ son la solución del problema de optimización (2.14). Ilustramos esto mediante la Figura 2.2, y escribimos el problema de mínimos cuadrados usando notación matricial compacta (2.6) como:

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2, \tag{2.16}$$

donde $\|\cdot\|_2$ denota la norma euclídea usual, y $\|\cdot\|_2^2$ su cuadrado. Desde el punto de vista del álgebra lineal, esto puede verse como el problema de encontrar el vector más cercano (en sentido euclídeo) a $\mathbf{y}$ en el subespacio de $\mathbb{R}^n$ generado por las columnas de $\mathbf{X}$. La solución a este problema es la proyección ortogonal de $\mathbf{y}$ sobre ese subespacio, y la correspondiente $\hat{\boldsymbol{\beta}}$ satisface (como se muestra en la Sección 2.A):

$$\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^\top\mathbf{y}. \tag{2.17}$$

La ecuación (2.17) se conoce comúnmente como las *ecuaciones normales*, y proporciona la solución al problema de mínimos cuadrados (2.14, 2.16). Si $\mathbf{X}^\top\mathbf{X}$ es invertible, lo cual sucede con frecuencia, $\hat{\boldsymbol{\beta}}$ tiene una solución en forma cerrada:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}. \tag{2.18}$$

El hecho de que exista una solución en forma cerrada es importante, y quizás explica por qué el método de mínimos cuadrados se ha vuelto tan popular y ampliamente utilizado. Como se mencionó, otras suposiciones sobre ε, distintas a la gaussianidad, conducen a otros problemas, como (2.15), en los cuales no existe solución cerrada.

Momento para reflexionar 2.1

¿Qué significa en la práctica que $\mathbf{X}^\top\mathbf{X}$ no sea invertible?

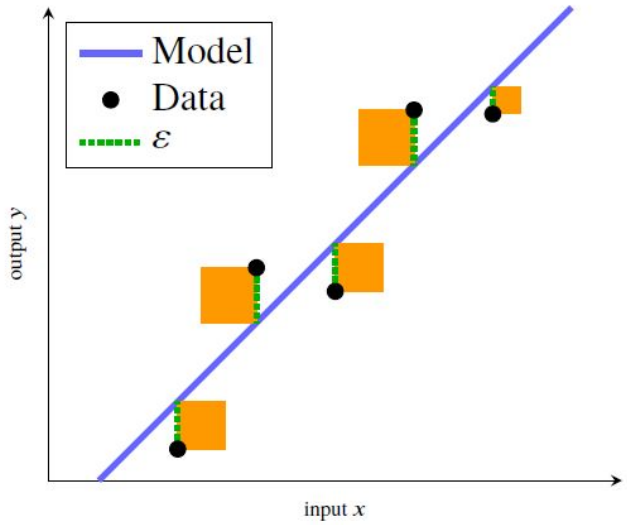


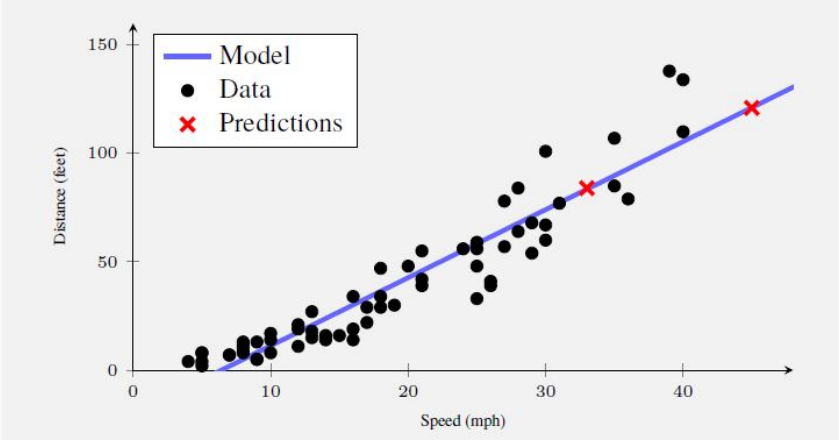
Figura 2: **Explicación gráfica del criterio de mínimos cuadrados:** el objetivo es elegir el modelo (línea azul) de modo que se minimice la suma de los cuadrados (en naranja) de cada error ε (en verde). Es decir, la línea azul debe seleccionarse para minimizar la cantidad de color naranja. Esto motiva el nombre *mínimos cuadrados*.

Momento para reflexionar 2.2

Si las columnas de $\mathbf{X}$ *son linealmente independientes y* $p = n - 1$*, entonces* $\mathbf{X}$ *abarca todo* $\mathbb{R}^n$*. Esto significa que existe una solución única tal que* $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$ *exactamente, es decir, el modelo ajusta perfectamente los datos de entrenamiento. En ese caso, (2.17) se reduce a* $\boldsymbol{\beta} = \mathbf{X}^{-1}\mathbf{y}$ *y el modelo ajusta perfectamente los datos. ¿Por qué no se desea esto?*

Ejemplo 2.3: Distancias de frenado de automóviles

Insertando las matrices de (2.7) del Ejemplo 2.2 en las ecuaciones normales (2.6), obtenemos $\hat{\beta}_0 = -20{,}1$ y $\hat{\beta}_1 = 3{,}1$. Si graficamos el modelo resultante, se ve así:

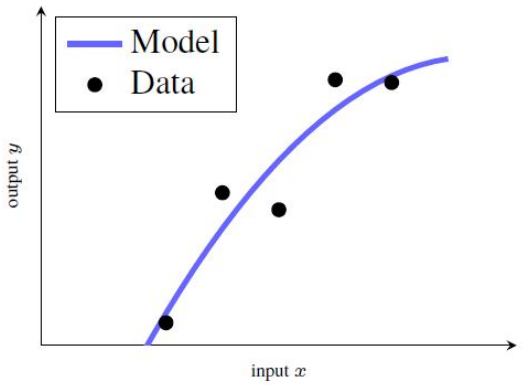


Con este modelo, la distancia de frenado predicha para $\mathbf{x}_* = 33$ mph es $\hat{y}_* = 84$ pies, y para $\mathbf{x}_* = 45$ mph es $\hat{y}_* = 121$ pies.

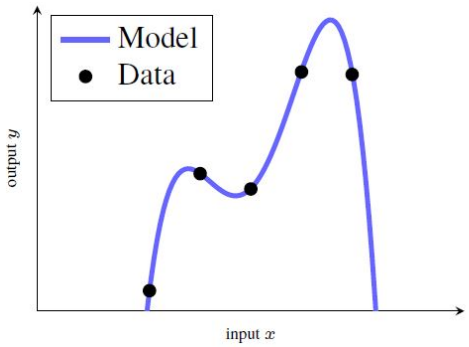
2.4 Transformaciones no lineales de las entradas — creando más características

La razón del término "lineal.[en] regresión lineal.[es] que la salida se modela como una *combinación lineal* de las entradas.[2] Sin embargo, no hemos definido con claridad qué se considera una entrada: si la velocidad es una entrada, ¿por qué no podría serlo también la energía cinética —es decir, su cuadrado?

La respuesta es que sí, puede serlo. De hecho, podemos hacer uso de transformaciones *no lineales* arbitrarias de las variables de entrada "originales" como nuevas variables de entrada dentro del modelo de regresión lineal.



(a) La solución de máxima verosimilitud con un polinomio de segundo orden en el modelo de regresión lineal. Como se comentó, la línea ya no es recta (ver Figura 2.1). Sin embargo, esto es solo un artefacto de la gráfica: en una representación tridimensional con cada característica (aquí, x y x^2) en un eje separado, seguiría siendo un conjunto afín.



(b) La solución de máxima verosimilitud con un polinomio de cuarto orden en el modelo de regresión lineal. Nótese que un polinomio de orden 4 contiene 5 coeficientes desconocidos, lo que sugiere que podemos esperar que el modelo aprendido se ajuste exactamente a 5 puntos de datos (cf. Observación 2.2, $p = n-1$).

Si, por ejemplo, tenemos solo una variable de entrada unidimensional para una entrada x, el modelo de regresión lineal estándar es:

$$y = \beta_0 + \beta_1 x + \varepsilon. \tag{2.19}$$

Sin embargo, también podemos extender el modelo incluyendo, por ejemplo, $x^2, x^3, \ldots, x^p$ como variables de entrada, obteniendo así un modelo de regresión lineal que es un polinomio en x:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + \varepsilon. \tag{2.20}$$

Obsérvese que esto sigue siendo un modelo de regresión lineal *ya que los parámetros desconocidos aparecen de forma lineal* con x, x^2, ..., x^p como nuevas entradas. Los parámetros $\hat{\boldsymbol{\beta}}$ se aprenden del mismo modo, pero la matriz $\mathbf{X}$ es diferente para los modelos (2.19) y (2.20). Nos referiremos a las entradas transformadas como *características*.

En escenarios más complejos, la distinción entre la entrada original y las características transformadas puede no ser tan clara, y los términos *característica* y *entrada* pueden usarse indistintamente.

Momento para reflexionar 2.3

La Figura 2.3 muestra un ejemplo de dos modelos de regresión lineal con entradas transformadas (polinomiales). Al estudiar la figura uno puede preguntarse cómo un modelo de regresión lineal puede dar lugar a una línea curva. ¿No están restringidos los modelos de regresión lineal a líneas rectas (o afines)? La respuesta es que depende del gráfico: la Figura 2.3(a) muestra una gráfica bidimensional con x, y (las entradas 'originales'), pero una gráfica tridimensional con x, x^2, y (cada característica en un eje separado) seguiría siendo afín. Lo mismo se aplica para la Figura 2.3(b), aunque en ese caso necesitaríamos una gráfica de 5 dimensiones.

Aunque el modelo de la Figura 2.3(b) puede ajustarse exactamente a todos los puntos de datos, esto también sugiere que los polinomios de orden superior no siempre son muy útiles: el comportamiento del modelo entre y fuera de los puntos de datos es bastante peculiar y no está muy bien motivado por los datos. Por esta razón, los polinomios de alto orden rara vez se usan en la práctica del aprendizaje automático.

Una alternativa —y mucho más común— es el llamado *núcleo de base radial* (RBF):

$$K_c(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{c}\|_2^2}{\ell}\right), \tag{2.21}$$

es decir, una campana de Gauss centrada en $\mathbf{c}$. Este núcleo puede usarse, en lugar de los polinomios, en el modelo de regresión lineal como:

$$y = \beta_0 + \beta_1 K_{c_1}(\mathbf{x}) + \beta_2 K_{c_2}(\mathbf{x}) + \cdots + \beta_p K_{c_p}(\mathbf{x}) + \varepsilon. \tag{2.22}$$

Este modelo puede interpretarse como p "bultos" ubicados en $c_1, c_2, \ldots, c_p$, respectivamente. Nótese que tanto las ubicaciones $c_1, c_2, \ldots, c_p$ como la escala de longitud ℓ deben ser decididas por el usuario, y solo los parámetros $\beta_0, \beta_1, \ldots, \beta_p$ se aprenden a partir de los datos en la regresión lineal. Esto se ilustra en la Figura 2.4.

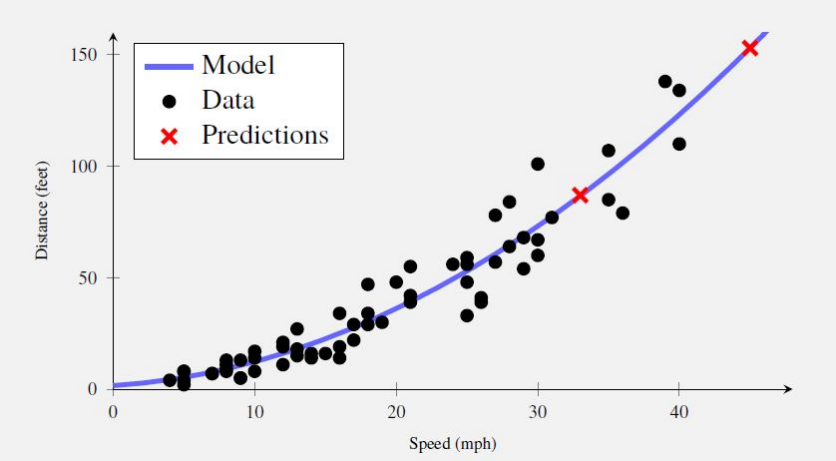
Los núcleos RBF son en general preferidos sobre los polinomios porque tienen propiedades "locales", lo que significa que un pequeño cambio en un parámetro afecta al modelo solo localmente alrededor de ese núcleo, mientras que un pequeño cambio en un parámetro en un modelo polinómico afecta al modelo en todas partes.

Ejemplo 2.4: Distancias de frenado de automóviles

Retomamos el Ejemplo 2.1, pero esta vez también añadimos la velocidad al cuadrado como una característica, es decir, las características ahora son x y x^2. Esto da las nuevas matrices (cf. (2.7)):

$$\mathbf{X} = \begin{bmatrix} 1 & 4 & 16 \\ 1 & 5 & 25 \\ 1 & 5 & 25 \\ \vdots & \vdots & \vdots \\ 1 & 39 & 1521 \\ 1 & 40 & 1600 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 4 \\ 2 \\ 4 \\ \vdots \\ 110 \\ 134 \end{bmatrix}. \tag{2.23}$$

Al insertar estas matrices en las ecuaciones normales (2.17), las nuevas estimaciones de los parámetros son $\hat{\beta}_0 = 1{,}58$, $\hat{\beta}_1 = 0{,}42$ y $\hat{\beta}_2 = 0{,}07$. (Nótese que $\hat{\beta}_0$ y $\hat{\beta}_1$ cambian en comparación con el Ejemplo 2.3). Este nuevo modelo se ve así:



Con este modelo, la distancia de frenado predicha ahora es $\hat{y}_* = 87$ pies para $\mathbf{x}_* = 33$ mph, y $\hat{y}_* = 153$ para $\mathbf{x}_* = 45$ mph.
Esto se puede comparar con el Ejemplo 2.3, que da predicciones distintas. Basándonos solo en los datos, no podemos afirmar que este sea el "modelo verdadero", pero al comparar visualmente este modelo con el del Ejemplo 2.3, parece que el modelo con más características sigue mejor los datos. Un método sistemático para seleccionar entre diferentes características (más allá de solo comparar gráficamente) es la validación cruzada, ver Capítulo 5.

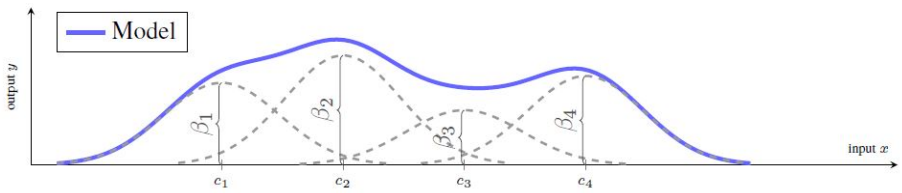


Figura 3: Un modelo de regresión lineal que utiliza núcleos RBF (ecuación 2.22) como características. Cada núcleo (líneas grises discontinuas) está ubicado en c_1, c_2, c_3 y c_4, respectivamente. Cuando el modelo se aprende a partir de los datos, los parámetros $\beta_0, \beta_1, \ldots, \beta_p$ se eligen de modo que la suma de todos los núcleos (línea azul sólida) se ajuste a los datos, por ejemplo, en el sentido de mínimos cuadrados.

Los polinomios y los núcleos RBF son solo dos casos especiales, pero podemos considerar cualquier transformación no lineal de las entradas. Para distinguir entre las entradas 'originales' y las 'nuevas' entradas transformadas, a estas últimas se las denomina *características* (*features*). Para decidir qué características utilizar, un enfoque consiste en comparar modelos competidores (con diferentes características) usando validación cruzada; véase el Capítulo 5.

2.5 Variables de entrada cualitativas

El problema de regresión se caracteriza por una salida cuantitativa y, pero la naturaleza de las entradas x es arbitraria. Hasta ahora solo hemos discutido el caso de entradas cuantitativas x, pero también son perfectamente posibles entradas cualitativas.

Supongamos que tenemos una variable de entrada cualitativa que solo toma dos valores diferentes (o niveles o clases), que llamamos tipo A y tipo B. Podemos entonces crear una *variable ficticia* x como

$$x = \begin{cases} 0 & \text{si es tipo A} \\ 1 & \text{si es tipo B} \end{cases} \tag{2.24}$$

y usar esta variable en el modelo de regresión lineal. Esto nos da efectivamente un modelo de regresión lineal que se ve así:

$$y = \beta_0 + \beta_1 x + \varepsilon = \begin{cases} \beta_0 + \varepsilon & \text{si es tipo A} \\ \beta_0 + \beta_1 + \varepsilon & \text{si es tipo B} \end{cases} \tag{2.25}$$

El criterio de elección es algo arbitrario, y los tipos A y B, por supuesto, se pueden intercambiar. Otras elecciones, como $x = 1$ o $x = -1$, también son posibles. Este enfoque se puede generalizar a variables de entrada cualitativas que toman más de dos valores; digamos, tipos A, B, C y D. Con cuatro valores diferentes, creamos $3 = 4 - 1$ variables ficticias como

$$x_1 = \begin{cases} 1 & \text{si el tipo es B} \\ 0 & \text{si no es tipo B} \end{cases}, \quad x_2 = \begin{cases} 1 & \text{si el tipo es C} \\ 0 & \text{si no es tipo C} \end{cases}, \quad x_3 = \begin{cases} 1 & \text{si el tipo es D} \\ 0 & \text{si no es tipo D} \end{cases} \tag{2.26}$$

lo que, en conjunto, da el modelo de regresión lineal

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon = \begin{cases} \beta_0 + \varepsilon & \text{si el tipo es A} \\ \beta_0 + \beta_1 + \varepsilon & \text{si el tipo es B} \\ \beta_0 + \beta_2 + \varepsilon & \text{si el tipo es C} \\ \beta_0 + \beta_3 + \varepsilon & \text{si el tipo es D} \end{cases} \tag{2.27}$$

Las entradas cualitativas pueden manejarse de manera similar en otros problemas y métodos, como regresión logística, k-NN, aprendizaje profundo, etc.

2.6 Regularización

Aunque el modelo de regresión lineal, a primera vista (cf. Figura 2.1), puede parecer un modelo bastante rígido y poco flexible, no es necesariamente así. Si se obtienen más características ampliando el modelo con transformaciones no lineales como en las Figuras 2.3 o 2.4, o si el número de entradas p es grande y el número de puntos de datos n es pequeño, puede experimentarse sobreajuste (overfitting).

Si consideramos los datos como formados por "señal" (la información real) y "ruido" (errores de medición, efectos irrelevantes, etc.), el término sobreajuste indica que el modelo se ajusta no solo a la "señal" sino también al "ruido". Un ejemplo de sobreajuste se da en el Ejemplo 2.5, donde un modelo de regresión lineal con $p = 8$ núcleos RBF se aprende a partir de $n = 9$ puntos de datos. Aunque el modelo sigue muy bien todos los puntos de datos, podemos juzgar intuitivamente que el modelo no es particularmente útil: ni la interpolación (entre los puntos de datos) ni la extrapolación (fuera del rango de datos) parecen razonables. Nótese que usar $p = n - 1$ es un caso extremo, pero el problema conceptual del sobreajuste suele estar presente incluso en situaciones menos extremas. El sobreajuste se discutirá a fondo más adelante, en el Capítulo 5.

Un enfoque útil para manejar el sobreajuste es la regularización. La regularización puede motivarse por "mantener los parámetros β pequeños a menos que los datos realmente nos convenzan de lo contrario", o alternativamente, "si un modelo con valores pequeños de los parámetros β se ajusta a los datos casi tan bien como un modelo con valores más grandes de los parámetros, debe preferirse el modelo con valores de parámetros pequeños". Hay varias formas de implementar esto matemáticamente, lo que conduce a soluciones ligeramente diferentes. Nos centraremos en la regresión ridge y el LASSO.

Para la regresión lineal, otra motivación para usar regularización es cuando $\mathbf{X}^\mathsf{T}\mathbf{X}$ no es invertible, lo que significa que (2.16) no tiene una solución única $\hat{\beta}$. En tales casos, la regularización puede introducirse para hacer que $\mathbf{X}^\mathsf{T}\mathbf{X}$ sea invertible y dar a (2.16) una solución única. Sin embargo, el concepto de regularización se extiende mucho más allá de la regresión lineal y también puede usarse al trabajar con otros tipos de problemas y modelos.

2.6.1 Regresión ridge

En la regresión ridge (también conocida como regularización de Tikhonov, regularización L2 o decaimiento de pesos), el criterio de mínimos cuadrados (2.16) se reemplaza por el problema de minimización modificado

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{minimizar}} \quad \|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \gamma\|\beta\|_2^2. \tag{2.28}$$

El valor $\gamma \geq 0$ se denomina parámetro de regularización y debe ser elegido por el usuario. Para $\gamma = 0$ recuperamos el problema original de mínimos cuadrados (2.16), mientras que si dejamos que $\gamma \to \infty$ forzaremos que todos los parámetros β_j tiendan a cero. Una buena elección de γ suele encontrarse entre ambos extremos y depende del problema en cuestión. Puede encontrarse mediante ajuste manual o de forma sistemática usando validación cruzada.

Es posible derivar una versión de las ecuaciones normales (2.17) para (2.28), a saber:

$$(\mathbf{X}^\mathsf{T}\mathbf{X} + \gamma I_{p+1})\hat{\beta} = \mathbf{X}^\mathsf{T}\mathbf{y}, \tag{2.29}$$

donde I_{p+1} es la matriz identidad de tamaño $(p+1) \times (p+1)$. Si $\gamma > 0$, la matriz $\mathbf{X}^\mathsf{T}\mathbf{X} + \gamma I_{p+1}$ siempre es invertible, y tenemos la solución en forma cerrada

$$\hat{\beta} = (\mathbf{X}^\mathsf{T}\mathbf{X} + \gamma I_{p+1})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}. \tag{2.30}$$

2.6.2 LASSO

Con LASSO (abreviatura de Least Absolute Shrinkage and Selection Operator), o de forma equivalente regularización L1, el criterio de mínimos cuadrados (2.16) se reemplaza por

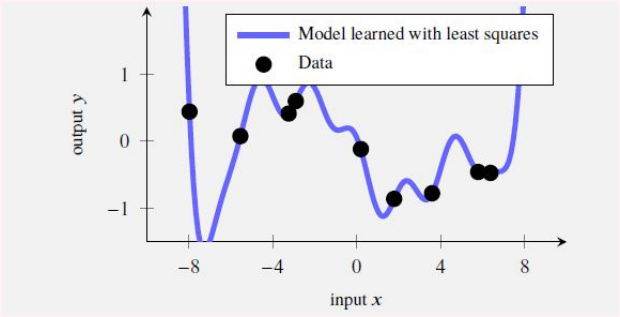
$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{minimizar}} \quad \|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \gamma\|\beta\|_1, \tag{2.31}$$

donde $\|\cdot\|_1$ es la norma Manhattan. A diferencia de la regresión ridge, no existe una solución en forma cerrada para (2.31). Sin embargo, es un problema convexo que puede resolverse de forma eficiente mediante optimización numérica.

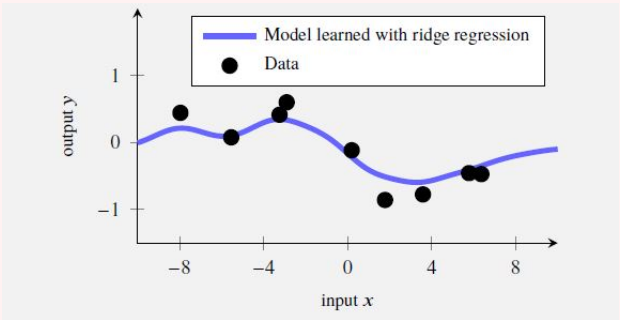
Al igual que en la regresión ridge, el parámetro de regularización γ debe ser elegido por el usuario en LASSO: $\gamma = 0$ da el problema de mínimos cuadrados y $\gamma \to \infty$ produce $\beta = 0$. Entre estos extremos, LASSO y la regresión ridge producen soluciones diferentes: mientras que la regresión ridge empuja todos los parámetros $\beta_0, \beta_1, \ldots, \beta_p$ hacia valores pequeños, LASSO tiende a favorecer soluciones dispersas, donde solo unos pocos parámetros son distintos de cero y el resto son exactamente cero. Por tanto, la solución LASSO puede apagar algunas entradas (o variables de entrada) poniendo sus parámetros correspondientes a cero, y puede utilizarse así como un método de selección de variables (o features).

Ejemplo 2.5: Regularización en un modelo RBF de regresión lineal

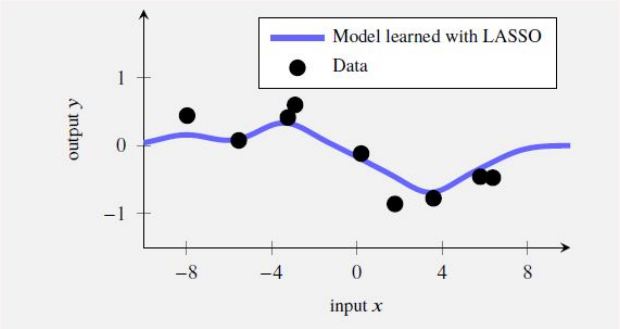
Consideramos el problema de aprender un modelo de regresión lineal (línea azul) con $p = 8$ kernels de base radial (RBF) como características a partir de $n = 9$ puntos de datos (puntos negros). Dado que tenemos $p = n - 1$, podemos esperar que el modelo se ajuste perfectamente a los datos. Sin embargo, como se ve en (a) a la derecha, el modelo presenta sobreajuste, lo que significa que se adapta demasiado a los datos y tiene un comportamiento 'extraño' entre los puntos. Para remediarlo, podemos usar regresión ridge (b) o LASSO (c). Aunque los modelos finales con ridge y LASSO parezcan similares, sus parámetros $\hat{\beta}$ son diferentes: LASSO usa de forma efectiva solo 5 (de 8) funciones de base radial. Esto se denomina una solución esparsa. La elección del método depende, por supuesto, del problema específico.



(a) El modelo aprendido con **mínimos cuadrados** (2.16). Aunque el modelo sigue exactamente los datos, normalmente no deberíamos estar satisfechos: el comportamiento entre puntos y fuera del rango no es plausible y es solo un efecto del sobreajuste, ya que el modelo se adapta 'demasiado bien'. Los valores de $\hat{\beta}$ rondan entre 30 y -30.



(b) El mismo modelo, esta vez aprendido con **ridge regression** (2.28) y un cierto valor de γ. Sin adaptarse perfectamente a los datos, este modelo ofrece un compromiso más razonable entre ajuste y sobreajuste, y es más útil que (a) en la mayoría de situaciones. Los valores de $\hat{\beta}$ están distribuidos en el rango $-0{,}5$ a $0{,}5$.



(c) El mismo modelo, esta vez aprendido con **LASSO** (2.31) y un cierto valor de γ. Nuevamente, el modelo no se adapta perfectamente a los datos, pero ofrece un compromiso razonable y es más útil que (a). A diferencia de (b), 3 de los 9 parámetros se ajustan exactamente a 0, y el resto está en el rango -1 a 1.

2.6.3 Regularización general de la función de costo

La *Ridge Regression* y el *LASSO* son dos casos especiales muy populares de regularización para la regresión lineal. Ambos tienen en común que modifican la función de costo, u objetivo de optimización, de (2.16). Pueden verse como dos instancias de un esquema de regularización más general:

$$\min_{\beta} \quad \underbrace{V(\beta, \mathbf{X}, \mathbf{y})}_{\text{ajuste a los datos}} + \gamma \quad \underbrace{R(\beta)}_{\substack{\text{penalización de la} \\ \text{flexibilidad del modelo}}}$$

Obsérvese que (2.32) contiene tres elementos importantes: (i) un término que describe qué tan bien se ajusta el modelo a los datos, (ii) un término que penaliza la complejidad del modelo (valores grandes de los parámetros), y (iii) un parámetro de equilibrio γ entre ambos.

2.7 Lecturas recomendadas

La regresión lineal se usa desde hace más de 200 años. Fue introducida de manera independiente por Adrien–Marie Legendre en 1805 y Carl Friedrich Gauss en 1809, cuando descubrieron el método de los mínimos cuadrados. La importancia de la regresión lineal se refleja en numerosos textos de estadística y aprendizaje automático, como Bishop (2006), Gelman *et al.* (2013), Hastie, Tibshirani y Friedman (2009), y Murphy (2012). Aunque la técnica básica de mínimos cuadrados existe desde hace mucho, sus versiones regularizadas son bastante más recientes. La regresión ridge se introdujo de forma independiente en estadística por Hoerl y Kennard (1970) y en análisis numérico bajo el nombre de regularización de Tikhonov. El LASSO fue presentado por Tibshirani (1996). La monografía reciente de Hastie, Tibshirani y Wainwright (2015) cubre el desarrollo relacionado con el uso de modelos dispersos y el LASSO.

2.A Derivación de las ecuaciones normales

Las ecuaciones normales (2.17)

$$\mathbf{X}^\mathsf{T}\mathbf{X}\,\hat{\boldsymbol{\beta}} \;=\; \mathbf{X}^\mathsf{T}\mathbf{y},$$

pueden derivarse a partir de (2.16)

$$\hat{\boldsymbol{\beta}} \;=\; \arg\min_{\boldsymbol{\beta}} \left\|\mathbf{X}\boldsymbol{\beta}-\mathbf{y}\right\|_2^2,$$

de distintas maneras. Presentaremos una basada en cálculo (matricial) y otra en geometría y álgebra lineal.

Sea cual sea la derivación de (2.17), si $\mathbf{X}^\mathsf{T}\mathbf{X}$ es invertible, entonces (únicamente) se obtiene

$$\hat{\boldsymbol{\beta}} \;=\; \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}.$$

Si $\mathbf{X}^\mathsf{T}\mathbf{X}$ no es invertible, entonces (2.17) tiene infinitas soluciones $\hat{\boldsymbol{\beta}}$, todas ellas igualmente buenas para el problema (2.16).

2.A.1 Un enfoque mediante cálculo

Sea

$$V(\beta) = \|\mathbf{X}\beta-\mathbf{y}\|_2^2 = (\mathbf{X}\beta-\mathbf{y})^\mathsf{T}(\mathbf{X}\beta-\mathbf{y}) = \mathbf{y}^\mathsf{T}\mathbf{y} - 2\,\mathbf{y}^\mathsf{T}\mathbf{X}\beta + \beta^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\beta, \tag{2.33}$$

y derivemos $V(\beta)$ con respecto al vector β,

$$\frac{\partial}{\partial\beta}V(\beta) = -2\,\mathbf{X}^\mathsf{T}\mathbf{y} + 2\,\mathbf{X}^\mathsf{T}\mathbf{X}\beta. \tag{2.34}$$

Dado que $V(\beta)$ es una forma cuadrática positiva, su mínimo debe alcanzarse cuando $\frac{\partial}{\partial\beta}V(\beta)=0$, lo que caracteriza la solución $\hat{\beta}$ como:

$$\frac{\partial}{\partial\beta}V(\hat{\beta})=0 \iff -2\,\mathbf{X}^\mathsf{T}\mathbf{y} + 2\,\mathbf{X}^\mathsf{T}\mathbf{X}\beta = 0 \iff \mathbf{X}^\mathsf{T}\mathbf{X}\,\hat{\beta} = \mathbf{X}^\mathsf{T}\mathbf{y}, \tag{2.35}$$

es decir, las ecuaciones normales.

2.A.2 Un enfoque mediante álgebra lineal

Sea $\mathbf{X}$ una matriz cuyas $p+1$ columnas son c_j, $j=1,\ldots,p+1$. Primero mostraremos que $\|\mathbf{X}\beta-\mathbf{y}\|_2^2$ se minimiza si β se elige tal que $\mathbf{X}\beta$ sea la proyección ortogonal de $\mathbf{y}$ sobre el (sub)espacio generado por las columnas c_j de $\mathbf{X}$, y luego que esta proyección ortogonal se obtiene mediante las ecuaciones normales.

Descomponemos $\mathbf{y}$ como $\mathbf{y}_\perp + \mathbf{y}_\parallel$, donde $\mathbf{y}_\perp$ es ortogonal al (sub)espacio generado por todas las columnas c_i, y $\mathbf{y}_\parallel$ está en el (sub)espacio generado por todas las columnas c_i. Como $\mathbf{y}_\perp$ es ortogonal tanto a $\mathbf{y}_\parallel$ como a $\mathbf{X}\beta$, se sigue que

$$\|\mathbf{X}\beta-\mathbf{y}\|_2^2 = \|\mathbf{X}\beta-(\mathbf{y}_\perp+\mathbf{y}_\parallel)\|_2^2 = \|(\mathbf{X}\beta-\mathbf{y}_\parallel)-\mathbf{y}_\perp\|_2^2 \geq \|\mathbf{y}_\perp\|_2^2, \tag{2.36}$$

y la desigualdad triangular también nos da

$$\|\mathbf{X}\beta-\mathbf{y}\|_2^2 = \|\mathbf{X}\beta-\mathbf{y}_\perp-\mathbf{y}_\parallel\|_2^2 \leq \|\mathbf{y}_\perp\|_2^2 + \|\mathbf{X}\beta-\mathbf{y}_\parallel\|_2^2. \tag{2.37}$$

Esto implica que si elegimos β tal que $\mathbf{X}\beta=\mathbf{y}_\parallel$, el criterio $\|\mathbf{X}\beta-\mathbf{y}\|_2^2$ alcanza su mínimo. Por lo tanto, nuestra solución $\hat{\beta}$ debe ser tal que $\mathbf{X}\beta-\mathbf{y}$ sea ortogonal al (sub)espacio generado por todas las columnas c_i, es decir:

$$(\mathbf{y}-\mathbf{X}\hat{\beta})^\mathsf{T}c_j = 0, \quad j=1,\ldots,p+1. \tag{2.38}$$

Recordemos que dos vectores $\mathbf{u}, \mathbf{v}$ son ortogonales si y solo si su producto escalar $\mathbf{u}^\mathsf{T}\mathbf{v}$ es cero. Dado que las columnas c_j forman la matriz $\mathbf{X}$, podemos escribir esto de forma compacta como

$$(\mathbf{y} - \mathbf{X}\hat{\beta})^\mathsf{T}\mathbf{X} = 0, \tag{2.39}$$

donde el lado derecho es el vector cero de dimensión $p+1$. Esto es equivalente a

$$\mathbf{X}^\mathsf{T}\mathbf{X}\,\hat{\beta} = \mathbf{X}^\mathsf{T}\mathbf{y},$$

es decir, las ecuaciones normales.