Pronóstico de datos numéricos: Métodos de regresión I

Las relaciones matemáticas nos ayudan a entender muchos aspectos de la vida cotidiana. Por ejemplo, el peso corporal es una función de la ingesta calórica; los ingresos suelen estar relacionados con la educación y la experiencia laboral; y las cifras de las encuestas ayudan a estimar las probabilidades de que un candidato presidencial sea reelegido.

Cuando estos patrones se formulan con números, obtenemos mayor claridad. Por ejemplo, 250 kilocalorías adicionales consumidas diariamente pueden resultar en un aumento de peso de casi un kilogramo por mes; cada año de experiencia laboral puede valer $1,000 adicionales en salario anual; y un presidente tiene más probabilidades de ser reelegido cuando la economía es fuerte. Obviamente, estas ecuaciones no se ajustan perfectamente a todas las situaciones, pero esperamos que sean razonablemente correctas la mayor parte del tiempo.

Este tema amplía nuestro conjunto de herramientas de aprendizaje automático yendo más allá de los métodos de clasificación tratados anteriormente e introduciendo técnicas para estimar relaciones dentro de datos numéricos.

Al examinar varias tareas de predicción numérica del mundo real, aprenderás:

- Los principios estadísticos básicos utilizados en la regresión, una técnica que modela el tamaño y la fuerza de las relaciones numéricas
- Cómo preparar datos para el análisis de regresión, estimar e interpretar modelos de regresión y aplicar variantes de regresión como los modelos lineales generalizados
- Un par de técnicas híbridas conocidas como árboles de regresión y árboles de modelos, que adaptan los clasificadores de árboles de decisión para tareas de predicción numérica

Basados en un gran cuerpo de trabajo en el campo de las estadísticas, los métodos utilizados en este tema son un poco más matemáticos que los tratados anteriormente, ¡pero no te preocupes! Incluso si tus conocimientos de álgebra están un poco oxidados, R se encarga del trabajo pesado.

Comprensión de la regresión

La regresión implica especificar la relación entre una sola variable numérica dependiente (el valor que se va a predecir) y una o más variables numéricas independientes (los predictores). Como lo indica el nombre, la variable dependiente depende del valor de la variable o

variables independientes. Las formas más simples de regresión suponen que la relación entre las variables independientes y dependientes sigue una línea recta.

El origen del término "regresión" para describir el proceso de ajustar líneas a los datos se remonta a un estudio de genética realizado por Sir Francis Galton a finales del siglo XIX. Descubrió que los padres que eran extremadamente bajos o altos tendían a tener hijos cuyas estaturas se acercaban más a la estatura media. Llamó a este fenómeno "regresión a la media".

Es posible que recuerdes del álgebra básica que las líneas se pueden definir en forma de **pendiente-intersección** similar a $y = a + bx$. En esta forma, la letra y indica la variable dependiente y x indica la variable independiente. El término de **pendiente** b especifica cuánto sube la línea por cada aumento de x. Los valores positivos definen líneas que se inclinan hacia arriba, mientras que los valores negativos definen líneas que se inclinan hacia abajo. El término a se conoce como **intersección** porque especifica el punto donde la línea cruza, o intercepta, el eje vertical y. Indica el valor de y cuando $x = 0$.

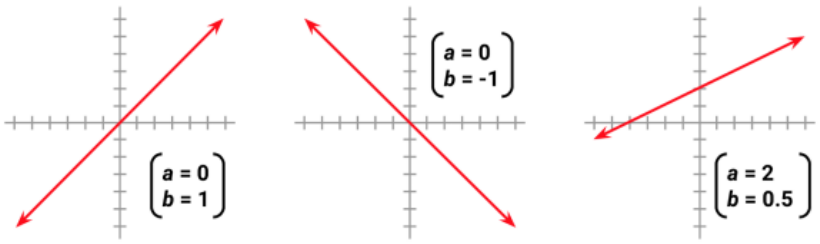

Figura 1: Ejemplos de líneas con distintas pendientes e intersecciones.

Las ecuaciones de regresión modelan los datos utilizando un formato de pendiente-intersección similar. El trabajo de la máquina es identificar los valores de a y b de modo que la línea especificada sea capaz de relacionar mejor los valores x suministrados con los valores de y.

Puede que no siempre haya un único conjunto de parámetros a y b que relacione perfectamente los valores, por lo que la máquina también debe tener alguna forma de cuantificar el margen de error y elegir el mejor ajuste. Hablaremos de esto en profundidad en breve.

El análisis de regresión se utiliza para una gran variedad de tareas; es casi seguro que es el método de aprendizaje automático más utilizado. Puede utilizarse tanto para explicar el pasado como para extrapolar al futuro y puede aplicarse a casi cualquier tarea. Algunos casos de uso específicos incluyen:

- Examinar cómo varían las poblaciones y los individuos según sus características medidas, en estudios científicos en los campos de la economía, la sociología, la psicología, la física y la ecología.
- Cuantificar la relación causal entre un evento y su respuesta, en casos como ensayos clínicos de medicamentos, pruebas de seguridad de ingeniería o investigación de mercado.
- Identificar patrones que se pueden usar para pronosticar el comportamiento futuro dados criterios conocidos, como para predecir reclamos de seguros, daños por desastres naturales, resultados electorales y tasas de criminalidad.

Los métodos de regresión también se utilizan para **pruebas de hipótesis estadísticas**, que determinan si es probable que una premisa sea verdadera o falsa a la luz de los datos observados. Las estimaciones del modelo de regresión de la fuerza y la consistencia de una relación brindan información que se puede usar para evaluar si las observaciones se deben solo al azar.

Las pruebas de hipótesis son extremadamente matizadas y quedan fuera del alcance del aprendizaje automático. Si te interesa este tema, un libro de texto introductorio de estadística es un buen lugar para comenzar, por ejemplo, Intuitive Introductory Statistics, de Wolfe, D. A. y Schneider, G., Springer, 2017.

El análisis de regresión no es sinónimo de un único algoritmo. Más bien, es un término general para muchos métodos, que se pueden adaptar a casi cualquier tarea de aprendizaje automático. Si estuvieras limitado a elegir solo un único método de aprendizaje automático para estudiar, la regresión sería una buena opción.
Uno podría dedicar toda una carrera a nada más y tal vez aún tendría mucho que aprender.

En este tema, comenzaremos con los modelos de **regresión lineal** más básicos: aquellos que utilizan líneas rectas. El caso en el que solo hay una variable independiente se conoce como **regresión lineal simple**. En el caso de dos o más variables independientes, se conoce como **regresión lineal múltiple** o simplemente regresión múltiple. Ambas técnicas suponen una única variable dependiente, que se mide en una escala continua.

La regresión también se puede utilizar para otros tipos de variables dependientes e incluso para algunas tareas de clasificación. Por ejemplo, la **regresión logística** se utiliza para modelar un resultado categórico binario, mientras que la regresión de Poisson (que recibe su nombre del matemático francés Siméon Poisson) modela datos de recuento de números enteros. El método conocido como **regresión logística multinomial** modela un resultado categórico y, por lo tanto, se puede utilizar para la clasificación.

Estos métodos de regresión especializados pertenecen a la clase de modelos lineales generalizados (GLM, **Generalized linear models**), que adaptan las líneas rectas de los modelos de regresión tradicionales para permitir el modelado de otras formas de datos. Estos se describirán más adelante en este tema.

Debido a que se aplican principios estadísticos similares en todos los métodos de regresión, una vez que comprendas el caso lineal, aprender sobre las otras variantes es sencillo. Comenzaremos con el caso básico de la regresión lineal simple. A pesar del nombre, este método no es demasiado simple para abordar problemas complejos. En la siguiente sección, veremos cómo el uso de un modelo de regresión lineal simple podría haber evitado un trágico desastre de ingeniería.

Regresión lineal simple

El 28 de enero de 1986, siete miembros de la tripulación del transbordador espacial estadounidense Challenger murieron cuando falló un propulsor del cohete, lo que provocó una desintegración catastrófica. Después del accidente, los expertos se centraron rápidamente en la temperatura de lanzamiento como posible culpable. Las juntas tóricas (O-rings) de goma responsables de sellar las juntas del cohete nunca se habían probado por debajo de los 40 °F (4 °C), y el clima el día del lanzamiento era inusualmente frío y bajo cero.

Con el beneficio de la retrospectiva, el accidente ha sido un caso de estudio sobre la importancia del análisis y la visualización de datos. Aunque no está claro qué información estaba disponible para los ingenieros de cohetes y los tomadores de decisiones antes del lanzamiento, es innegable que mejores datos, utilizados con cuidado, muy bien podrían haber evitado este desastre.

Los ingenieros de cohetes casi con certeza sabían que las temperaturas frías podrían hacer que los componentes fueran más frágiles y menos capaces de sellar correctamente, lo que daría como resultado una mayor probabilidad de una fuga de combustible peligrosa. Sin embargo, dada la presión política para continuar con el lanzamiento, necesitaban datos que respaldaran esta hipótesis. Un modelo de regresión que demostrara un vínculo entre la temperatura y los fallos de las juntas tóricas, y que pudiera predecir la probabilidad de fallo dada la temperatura esperada en el lanzamiento, podría haber sido muy útil.

Para construir el modelo de regresión, los científicos podrían haber utilizado los datos sobre la temperatura de lanzamiento y el deterioro de los componentes registrados durante los 23 lanzamientos exitosos anteriores del transbordador. El deterioro de los componentes indica uno de dos tipos de problemas. El primer problema, llamado erosión, ocurre cuando el calor

excesivo quema la junta tórica. El segundo problema, llamado fuga de gases, ocurre cuando los gases calientes se filtran a través de una junta tórica mal sellada o "pasan por encima".

Dado que el transbordador tenía un total de seis juntas tóricas primarias, podían ocurrir hasta seis averías por vuelo. Aunque el cohete podía sobrevivir a uno o más eventos de avería o destruirse con tan solo uno, cada avería adicional aumentaba la probabilidad de una falla catastrófica. El siguiente diagrama de dispersión muestra un gráfico de los problemas primarios de las juntas tóricas detectados en los 23 lanzamientos anteriores, en comparación con la temperatura en el momento del lanzamiento:

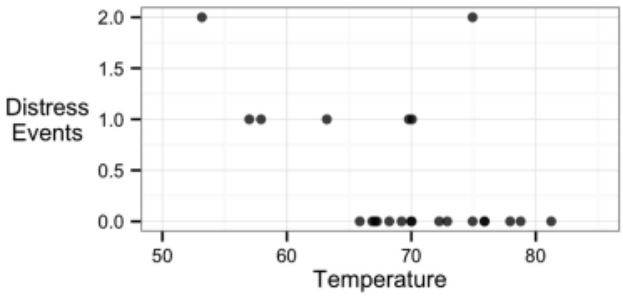

Figura 2: Visualización de los problemas de las juntas tóricas del transbordador espacial en función de la temperatura de lanzamiento.

Al examinar el gráfico, hay una tendencia aparente: los lanzamientos que se producen a temperaturas más altas tienden a tener menos eventos de problemas de las juntas tóricas. Además, el lanzamiento más frío (53 °F) tuvo dos eventos de problemas, un nivel que solo se había alcanzado en otro lanzamiento. Con esta información en mente, el hecho de que el Challenger estuviera programado para lanzarse en condiciones más de 20 grados más frías parece preocupante. Pero, ¿exactamente hasta qué punto deberían haber estado preocupados? Para responder a esta pregunta, podemos recurrir a una regresión lineal simple.

Un modelo de regresión lineal simple define la relación entre una variable dependiente y una única variable predictora independiente utilizando una línea definida por una ecuación en la siguiente forma:

$$y = \alpha + \beta x$$

Aparte de los caracteres griegos, esta ecuación es prácticamente idéntica a la forma pendiente-intersección descrita anteriormente. La intersección, α (alfa), describe el punto en el que la línea cruza el eje *y*, mientras que la pendiente, β (beta), describe el cambio en *y* dado un aumento de *x*. Para los datos del lanzamiento del transbordador, la pendiente nos indicaría el cambio esperado en las fallas de las juntas tóricas por cada grado que aumenta la temperatura de lanzamiento.

Los caracteres griegos se utilizan a menudo en el campo de la estadística para indicar variables que son parámetros de una función estadística. Por lo tanto, realizar un análisis de regresión implica encontrar **estimaciones de parámetros** para α y β.

Las estimaciones de parámetros para alfa y beta se denotan normalmente utilizando a y b, aunque es posible que descubras que parte de esta terminología y notación se utilizan indistintamente.

Supongamos que sabemos que los parámetros de regresión estimados en la ecuación para los datos de lanzamiento del transbordador son a = 3.70 y b = -0,048. En consecuencia, la ecuación lineal completa es y = 3.70 – 0.048 x. Ignorando por un momento cómo se obtuvieron estos números, podemos trazar la línea en el diagrama de dispersión de la siguiente manera:

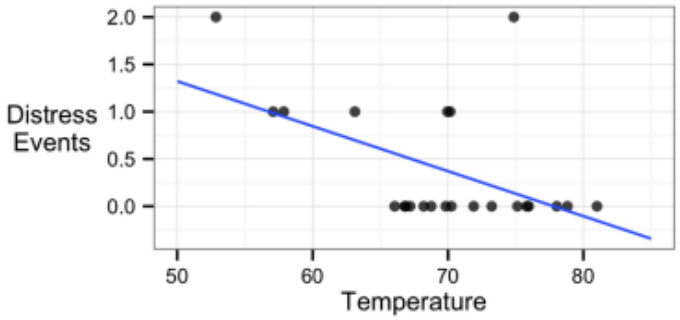

Figura 3: Una línea de regresión que modela la relación entre los eventos de avería y la temperatura de lanzamiento.

Como muestra la línea, a 60 grados Fahrenheit, predecimos menos de un evento de avería de la junta tórica. A 50 grados Fahrenheit, esperamos alrededor de 1.3 fallas. Si usamos el modelo para extrapolar hasta 31 grados (la temperatura pronosticada para el lanzamiento del Challenger), esperaríamos alrededor de 3-70 – 0.048 * 31 = 2.21 eventos de falla de junta tórica.

Suponiendo que cada falla de junta tórica tiene la misma probabilidad de causar una fuga catastrófica de combustible, esto significa que el lanzamiento del Challenger a 31 grados fue casi tres veces más riesgoso que el lanzamiento típico a 60 grados, y más de ocho veces más riesgoso que un lanzamiento a 70 grados.

Observa que la línea no pasa por cada punto de datos exactamente. En cambio, corta a través de los datos de manera más o menos uniforme, con algunas predicciones más bajas o más altas que la línea. En la siguiente sección, aprenderemos por qué se eligió esta línea en particular.

Estimación de mínimos cuadrados ordinarios

Para determinar las estimaciones óptimas de α y β se utiliza un método de estimación conocido como mínimos cuadrados ordinarios (OLS, **Ordinary least squares**). En la regresión OLS, la pendiente y la intersección se eligen de modo que minimicen la suma de los errores al cuadrado (SSE, **Sum of the squared errors**). Los errores, también conocidos como residuos (*residuales*), son la distancia vertical entre el valor de *y* predicho y el valor de *y* real. Debido a que los errores pueden ser sobreestimaciones o subestimaciones, pueden ser valores positivos o negativos; elevarlos al cuadrado hace que los errores sean positivos independientemente de la dirección. Los residuos se ilustran para varios puntos en el siguiente diagrama:

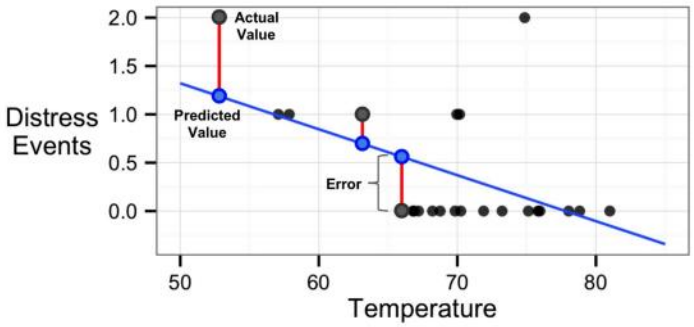

Figura 4: Las predicciones de la línea de regresión difieren de los valores reales en una cantidad residual.

En términos matemáticos, el objetivo de la regresión OLS se puede expresar como la tarea de minimizar la siguiente ecuación:

$$\sum (y_i - \hat{y}_i)^2 = \sum e_i^2$$

En lenguaje sencillo, esta ecuación define *e* (el error) como la diferencia entre el valor de *y* real y el valor de *y* predicho. Los valores de error se elevan al cuadrado para eliminar los valores negativos y se suman en todos los puntos de los datos.

El signo de intercalación (^) sobre el término *y* es una característica de notación estadística que se utiliza con frecuencia. Indica que el término es una estimación del valor *y* verdadero. Esto se conoce como *y* sombrero.

La solución para *a* depende del valor de *b*. Se puede obtener utilizando la siguiente fórmula:

$$a = \bar{y} - b\bar{x}$$

Para entender estas ecuaciones, necesitarás conocer otro dato de notación estadística. La barra horizontal que aparece sobre los términos *x* e *y* indica el valor medio de *x* o *y*. Esto se conoce como *x* barra o *y* barra.

Aunque la prueba está fuera del alcance de este curso, se puede demostrar mediante cálculo que el valor de *b* que da como resultado el error cuadrático mínimo es:

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

Si descomponemos esta ecuación en sus partes componentes, podemos simplificarla un poco. El denominador de *b* debería resultar familiar; es muy similar a la varianza de *x*, que se denota como Var(*x*). Como aprendimos en el tema, Gestión y comprensión de datos, la varianza implica encontrar la desviación cuadrática promedio de la media de *x*. Esto se puede expresar como:

$$\text{Var}(x) = \frac{\sum(x_i - \bar{x})^2}{n}$$

El numerador implica tomar la suma de la desviación de cada punto de datos con respecto al valor medio *x* multiplicada por la desviación de ese punto con respecto al valor medio *y*. Esto es similar a la función de covarianza para *x* e *y*, denotada como Cov(*x*, *y*). La fórmula de covarianza es:

$$\text{Cov}(x, y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n}$$

Si dividimos la función de covarianza por la función de varianza, los términos *n* en el numerador y el denominador se cancelan entre sí y podemos reescribir la fórmula para *b* como:

$$b = \frac{\text{Cov}(x, y)}{\text{Var}(x)}$$

Teniendo en cuenta esta reformulación, es fácil calcular el valor de *b* utilizando funciones integradas de R. Apliquémoslas a los datos del lanzamiento del transbordador para estimar la línea de regresión.

Si deseas seguir estos ejemplos, descarga el archivo challenger.csv adjunto y cárgalo en un frame de datos utilizando el comando launch<- read.csv("challenger.csv").

Si los datos del lanzamiento del transbordador se almacenan en un frame de datos llamado launch, la variable independiente *x* se llama temperature y la variable dependiente *y* se llama distress_ct, podemos utilizar las funciones R cov() y var() para estimar *b*:

```
> b <- cov(launch$temperature, launch$distress_ct)/var(launch$temperature)
> b
[1] -0.04753968
```

Podemos estimar *a* utilizando el valor *b* calculado y aplicando la función mean():

```
> a <- mean(launch$distress_ct) - b * mean(launch$temperature)
> a
[1] 3.698413
```

Estimar la ecuación de regresión a mano obviamente no es lo ideal, por lo que R proporciona, como era de esperar, una función para ajustar los modelos de regresión automáticamente. Usaremos esta función en breve. Antes de eso, es importante ampliar tu comprensión del ajuste del modelo de regresión aprendiendo primero un método para medir la fuerza de una relación lineal. Además, pronto aprenderás cómo aplicar la regresión lineal múltiple a problemas con más de una variable independiente.

Correlaciones

La correlación entre dos variables es un número que indica qué tan cerca está su relación de seguir una línea recta. Sin más aclaraciones, la correlación generalmente se refiere al **coeficiente de correlación de Pearson**, que fue desarrollado por el matemático del siglo XX Karl Pearson. Una correlación se encuentra en el rango entre -1 y +1. Los valores máximo y mínimo indican una relación perfectamente lineal, mientras que una correlación cercana a cero indica la ausencia de una relación lineal.

La siguiente fórmula define la correlación de Pearson:

$$\rho_{x,y} = \mathrm{Corr}(x, y) = \frac{\mathrm{Cov}(x, y)}{\sigma_x \sigma_y}$$

Aquí se ha introducido más notación griega: el primer símbolo (que parece una p minúscula) es rho, y se utiliza para denotar la estadística de correlación de Pearson. Los símbolos que parecen caracteres q girados en sentido antihorario son la letra griega minúscula sigma, e indican la desviación estándar de *x* o *y*.

Usando esta fórmula, podemos calcular la correlación entre la temperatura de lanzamiento y el número de eventos de avería de las juntas tóricas. Recordemos que la función de covarianza es cov() y la función de desviación estándar es sd(). Almacenaremos el resultado en r, una letra que se usa comúnmente para indicar la correlación estimada:

```
> r <- cov(launch$temperature, launch$distress_ct)/
(sd(launch$temperature) * sd(launch$distress_ct))
> r
[1] -0.5111264
```

Alternativamente, podemos obtener el mismo resultado con la función de correlación cor():

```
> cor(launch$temperature, launch$distress_ct)
[1] -0.5111264
```

La correlación entre la temperatura y el número de juntas tóricas dañadas es -0.51. La correlación negativa implica que los aumentos de temperatura están relacionados con disminuciones en el número de juntas tóricas dañadas. Para los ingenieros de la NASA que estudian los datos de las juntas tóricas, esto habría sido un indicador muy claro de que un lanzamiento a baja temperatura podría ser problemático. La correlación también nos dice sobre la fuerza relativa de la relación entre la temperatura y el deterioro de las juntas tóricas. Dado que -0.51 está a mitad de camino de la correlación negativa máxima de -1, esto implica que existe una asociación lineal negativa moderadamente fuerte.

Se utilizan varias reglas generales para interpretar la fuerza de la correlación. Un método asigna un estado de "débil" a los valores entre 0.1 y 0.3; "moderado" al rango de 0.3 a 0.5; y "fuerte" a los valores superiores a 0.5 (estos también se aplican a rangos similares de correlaciones negativas). Sin embargo, estos umbrales pueden ser demasiado estrictos o demasiado laxos para ciertos fines. A menudo, la correlación debe interpretarse en contexto.

En el caso de los datos que involucran a seres humanos, una correlación de 0.5 puede considerarse muy alta; en el caso de los datos generados por procesos mecánicos, una correlación de 0.5 puede ser muy débil.

Probablemente hayas oído la expresión "la correlación no implica causalidad".
Esto se basa en el hecho de que una correlación solo describe la asociación entre un par de variables, pero podría haber otras explicaciones que no se tienen en cuenta y que sean responsables de la relación observada. Por ejemplo, puede haber una fuerte correlación entre la esperanza de vida y el tiempo que se pasa al día viendo películas, pero antes de que los médicos recomienden que todos veamos más películas, debemos descartar otra explicación: las personas más jóvenes ven más películas y tienen (en general) menos probabilidades de morir.

Medir la correlación entre dos variables nos brinda una forma de verificar rápidamente las relaciones lineales entre las variables independientes y la variable dependiente. Esto será cada vez más importante a medida que comencemos a definir modelos de regresión con un mayor número de predictores.

Regresión lineal múltiple

La mayoría de los análisis del mundo real tienen más de una variable independiente. Por lo tanto, es probable que utilices la regresión lineal múltiple para la mayoría de las tareas de predicción numérica. Las fortalezas y debilidades de la regresión lineal múltiple se muestran en la siguiente tabla:

Fortalezas	**Debilidades**
• Es, con diferencia, el enfoque más común para modelar datos numéricos • Se puede adaptar para modelar casi cualquier tarea de modelado • Proporciona estimaciones tanto del tamaño como de la fuerza de las relaciones entre las características y el resultado	• Hace suposiciones sólidas sobre los datos • El usuario debe especificar de antemano la forma del modelo • No maneja datos faltantes • Solo funciona con características numéricas, por lo que los datos categóricos requieren preparación adicional • Requiere algunos conocimientos de estadística para comprender el modelo

Podemos entender la regresión múltiple como una extensión de la regresión lineal simple. El objetivo en ambos casos es similar: encontrar valores de coeficientes de pendiente que minimicen el error de predicción de una ecuación lineal. La diferencia clave es que hay términos adicionales para las variables independientes adicionales.

Los modelos de regresión múltiple adoptan la forma de la siguiente ecuación. La variable dependiente *y* se especifica como la suma de un término de intersección α más el producto del valor β estimado y la variable *x* para cada una de las *i* características. Aquí se ha añadido un término de error ε (denotado por la letra griega épsilon) como recordatorio de que las predicciones no son perfectas. Esto representa el término residual indicado anteriormente:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + \varepsilon$$

Consideremos por un momento la interpretación de los parámetros de regresión estimados. Notarás que en la ecuación anterior se proporciona un coeficiente para cada característica. Esto permite que cada característica tenga un efecto estimado independiente sobre el valor de *y*. En otras palabras, *y* cambia en la cantidad β_i por cada unidad de aumento en la característica x_i. La intersección α es entonces el valor esperado de *y* cuando las variables independientes son todas cero.

Dado que el término de intersección α en realidad no es diferente de cualquier otro parámetro de regresión, también se lo denota a veces como β_0 (se pronuncia beta cero, *beta naught*) como se muestra en la siguiente ecuación:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + \varepsilon$$

Al igual que antes, la intersección no está relacionada con ninguna de las variables independientes *x*. Sin embargo, por razones que se aclararán en breve, resulta útil imaginar β_0 como si se estuviera multiplicando por un término x_0. Asignamos x_0 como una constante con el valor 1:

$$y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + \varepsilon$$

Para estimar los parámetros de regresión, cada valor observado de la variable dependiente *y* debe estar relacionado con los valores observados de las variables independientes *x* utilizando la ecuación de regresión en la forma anterior. La siguiente figura es una representación gráfica de la configuración de una tarea de regresión múltiple:

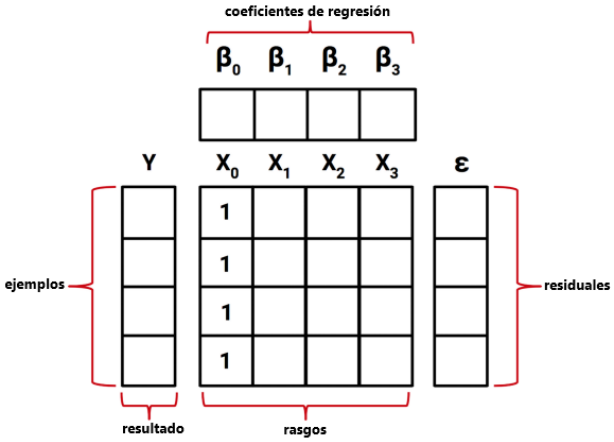

Figura 5: La regresión múltiple busca encontrar los valores β que relacionan los valores *X* con *Y* mientras minimiza ε.

Las numerosas filas y columnas de datos ilustradas en la figura anterior se pueden describir en una formulación condensada utilizando la notación matricial en negrita para indicar que cada uno de los términos representa múltiples valores. Simplificada de esta manera, la fórmula es la siguiente:

$$\boldsymbol{Y} = \boldsymbol{\beta X} + \varepsilon$$

En la notación matricial, la variable dependiente es un vector, **Y**, con una fila para cada ejemplo. Las variables independientes se han combinado en una matriz, **X**, con una columna para cada característica más una columna adicional de 1 valores para la intersección. Cada columna tiene una fila para cada ejemplo. Los coeficientes de regresión $\boldsymbol{\beta}$ y los errores residuales $\boldsymbol{\varepsilon}$ ahora también son vectores.

El objetivo ahora es resolver $\boldsymbol{\beta}$, el vector de coeficientes de regresión que minimiza la suma de los errores al cuadrado entre los valores **Y** predichos y reales. Encontrar la solución óptima

requiere el uso de álgebra matricial; por lo tanto, la derivación merece una atención más cuidadosa de la que se puede proporcionar en este documento. Sin embargo, si estás dispuesto a confiar en el trabajo de otros, la mejor estimación del vector $\boldsymbol{\beta}$ se puede calcular como:

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

Esta solución utiliza un par de operaciones matriciales: la **T** indica la transpuesta de la matriz **X**, mientras que el exponente negativo indica la **matriz inversa**. Usando las operaciones matriciales incorporadas de R, podemos implementar un aprendizaje de regresión múltiple simple. Apliquemos esta fórmula a los datos del lanzamiento del Challenger.

Usando el siguiente código, podemos crear una función de regresión básica llamada reg(), que toma un parámetro *y* y un parámetro *x*, y devuelve un vector de coeficientes beta estimados:

```
> reg <- function(y, x) {
+   x <- as.matrix(x)
+   x <- cbind(Intercept = 1, x)
+   b <- solve(t(x) %*% x) %*% t(x) %*% y
+   colnames(b) <- "estimate"
+   print(b)
+ }
```

La función reg() creada aquí usa varios comandos R que no hemos usado anteriormente. Primero, dado que usaremos la función con conjuntos de columnas de un frame de datos, la función as.matrix() convierte el frame de datos en forma de matriz.

A continuación, la función cbind() vincula una columna adicional a la matriz x; el comando Intercept = 1 indica a R que nombre la nueva columna Intercept y que la rellene con valores 1 repetidos. A continuación, se realizan una serie de operaciones matriciales en los objetos x e y:

- solve() toma la inversa de una matriz
- t() se utiliza para transponer una matriz
- %*% multiplica dos matrices

Al combinar estas funciones R como se muestra, nuestra función devolverá un vector b, que contiene parámetros estimados para el modelo lineal que relaciona x con y. Las dos líneas finales de la función le dan un nombre al vector b e imprimen el resultado en la pantalla.

Apliquemos esta función a los datos del lanzamiento del transbordador. Como se muestra en el código siguiente, el conjunto de datos incluye tres características y el recuento de averías (distress_ct), que es el resultado de interés:

```
> str(launch)

'data.frame':   23 obs. of  4 variables:
 $ distress_ct         : int  0 1 0 0 0 0 0 0 1 1 ...
 $ temperature         : int  66 70 69 68 67 72 73 70 57 63 ...
 $ field_check_pressure: int  50 50 50 50 50 50 100 100 200 200 ...
 $ flight_num          : int  1 2 3 4 5 6 7 8 9 10 ...
```

Podemos confirmar que nuestra función está funcionando correctamente comparando su resultado para el modelo de regresión lineal simple de fallas de juntas tóricas versus temperatura, que encontramos anteriormente que tiene parámetros a = 3.70 y b = -0.048. Como la temperatura está en la segunda columna de los datos de lanzamiento, podemos ejecutar la función reg() de la siguiente manera:

```
> reg(y = launch$distress_ct, x = launch[2])
                estimate
Intercept      3.69841270
temperature -0.04753968
```

Estos valores coinciden exactamente con nuestro resultado anterior, así que usemos la función para crear un modelo de regresión múltiple. Lo aplicaremos igual que antes, pero esta vez especificaremos las columnas dos a cuatro para el parámetro x para agregar dos predictores adicionales:

```
> reg(y = launch$distress_ct, x = launch[2:4])
                          estimate
Intercept                3.527093383
temperature             -0.051385940
field_check_pressure     0.001757009
flight_num               0.014292843
```

Este modelo predice el número de eventos de desgaste de la junta tórica utilizando la temperatura, la presión de verificación de campo y el número de identificación del lanzamiento. Cabe destacar que la inclusión de los dos nuevos predictores no cambió nuestro hallazgo del modelo de regresión lineal simple.

Al igual que antes, el coeficiente para la variable de temperatura es negativo, lo que sugiere que a medida que aumenta la temperatura, el número esperado de eventos de desgaste de la junta tórica disminuye. La magnitud del efecto también es aproximadamente la misma: se

esperan aproximadamente 0.05 eventos de desgaste menos por cada grado de aumento en la temperatura de lanzamiento.

Los dos nuevos predictores también contribuyen a los eventos de desgaste previstos. La presión de verificación de campo se refiere a la cantidad de presión aplicada a la junta tórica durante las pruebas previas al lanzamiento. Aunque la presión de verificación era originalmente de 50 psi, se aumentó a 100 y 200 psi para algunos lanzamientos, lo que llevó a algunos a creer que puede ser responsable de la erosión de la junta tórica. El coeficiente es positivo, pero pequeño, lo que proporciona al menos un poco de evidencia para esta hipótesis.

El número de vuelo explica la edad del transbordador. Con cada vuelo, el transbordador se hace más viejo y sus piezas pueden volverse más frágiles o propensas a fallar. La pequeña asociación positiva entre el número de vuelos y el número de averías puede reflejar este hecho.

En general, nuestro análisis retrospectivo de los datos del transbordador espacial sugiere que había razones para creer que el lanzamiento del Challenger era muy riesgoso dadas las condiciones meteorológicas. Tal vez si los ingenieros hubieran aplicado la regresión lineal de antemano, se podría haber evitado un desastre. Por supuesto, la realidad de la situación y las implicaciones políticas involucradas seguramente no eran tan simples entonces como parecen ahora en retrospectiva.

Nota. Continuamos la parte teórica en el documento 2.