1

Regresión lineal

1.1 Regresión lineal simple

Matemáticamente, podemos escribir una regresión lineal simple como:

$$Y \approx \beta_0 + \beta_1 X$$

β_0 y β_1 son conocidos cómo coeficientes o parámetros del modelo. Estos valores son desconocidos y deben producir estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$, así

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

donde $\hat{y}$ es la predicción de Y basada en el valor de $X = x$. Utilizamos el símbolo sobrero $'\hat{}'$ para indicar el valor estimado de un parámetro o coeficiente desconocido, o para indicar el valor predictivo de la respuesta.

1.1.1 Estimación de los coeficientes

Sea $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ el predictor para Y pasado en el i-ésimo valor de X. Entonces, $e_i = y_i - \hat{y}_i$ representada como el residuo i-ésimo. Definamos el cuadrado de la suma residual como

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2,$$

que es equivalente a decir que

$$RSS = \left(y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1\right)^2 + \left(y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2\right)^2 + \cdots + \left(y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n\right)^2.$$

El método de mínimos cuadrados elige $\hat{\beta}_0$ y $\hat{\beta}_1$ para minimizar RSS. Utilizando algún cálculo, se puede demostrar que los minimizadores son

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

donde $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ y $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ son las medias muestrales. Estos son conocidos como las estimaciones de los coeficientes de mínimos cuadrados para la regresión lineal simple.

1.1.2 Evaluación de la exactitud de las estimaciones de los coeficientes

Si f debe aproximarse mediante una función lineal, entonces podemos escribir esta relación como:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Normalmente suponemos que el término de error es independiente de X, este término de error abarca todo lo que se nos escapa en el modelo simple, y se genera a partir de una distribución normal con media cero.

Fundamentalmente, se utiliza un enfoque estadístico estándar donde utilizamos la información de una muestra para estimar las características de una población. Cómo la media muestral proporciona una buena estimación de la media poblacional, así los coeficientes β_0 y β_1 es la regresión lineal definen la recta de regresión poblacional. Intentamos estimar estos coeficientes desconocidos utilizando $\hat{\beta}_0$ y $\hat{\beta}_1$, que definen la recta de mínimos cuadrados.
La analogía entre la regresión lineal y la estimación de la media de una variable aleatoria es adecuada y se basa en el concepto de sesgo. Si utilizamos la media muestral $\hat{\mu}$ para estimar μ, esta estimación es insesgada, en el sentido de que, en promedio, esperamos que $\hat{\mu}$ sea igual a μ. Esto significa que sobre la bases de un conjunto particular de observaciones $y_1, \ldots, y_n$, $\hat{\mu}$ podría sobrestimar μ, y sobre la base de otro conjunto de observaciones, $\hat{\mu}$ podría subestimar μ. Pero si pudiéramos promediar un gran número de estimaciones de μ obtenidas a partir de un gran número de conjuntos de observaciones, entonces este promedio sería exactamente igual a μ. Por lo tanto, un estimador insesgado no sobrestima ni subestima sistemáticamente el parámetro verdadero.
La propiedad de insesgadez también es válida para los estimadores de coeficientes por mínimos cuadrados. Si estimamos β_0 y β_1 a partir de un conjunto de datos concreto, nuestras estimaciones no serán exactamente iguales al de los parámetros poblacionales, Pero si pudiéramos promediar las estimaciones obtenidas a partir de un gran número de conjuntos de datos, podremos estimar la línea de regresión.

Ahora, ¿hasta que punto es exacta la estimación de $\hat{\mu}$ para hallar μ?. Sabemos que una única estimación podría subestimar o sobrestimar considerablemente la media poblacional. Por lo que debemos preguntarnos es: ¿Cuánto, se alejará esa única estimación de $\hat{\mu}$?; en general, responderemos a esta pregunta calculando el error estándar de $\hat{\mu}$ escrito como $SE(\hat{\mu})$, acá escribimos la conocida fórmula

$$\text{Var}(\hat{\mu}) = SE(\hat{\mu})^2 = \frac{\sigma^2}{n},$$

donde σ es la desviación estándar de cada de las realizaciones y_i de Y. En términos generales, el error estándar nos dice en que medida esta estimación $\hat{\mu}$ difiere del valor real de μ. Esta ecuación también nos dice cómo esta desviación se reduce con n: cuantas más observaciones tengamos, menor será el error estándar de $\hat{\mu}$.
De manera similar podemos preguntarnos que tan cerca están $\hat{\beta}_0$ y $\hat{\beta}_1$ de los valores verdaderos de β_0 y β_1. Para calcular los errores estándar asociados con $\hat{\beta}_0$ y $\hat{\beta}_1$, utilizamos la siguiente fórmula:

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right], \qquad SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

donde $\sigma^2 = \text{Var}(\epsilon)$. Para que se cumplan estas fórmulas, necesitamos suponer que los errores ϵ_i para cada observación tiene varianza común σ^2 y no están correlacionadas. Obsérvese en la fórmula que $SE(\hat{\beta}_1)$ es menor cuando las x_i están más dispersas; intuitivamente tenemos más ventaja para estimar una pendiente cuando éste es el caso. También vemos que $SE(\hat{\beta}_0)$ sería igual a $SE(\hat{\mu})$ si $\overline{x}$ fuera cero (en cuyo caso $\hat{\beta}_0$ sería igual a $\overline{y}$). En general, σ^2 no se conoce, pero puede estimarse a partir de los datos. Esta estimación de σ se conoce como error estándar residual, y viene dada por la fórmula

$$RSE = \sqrt{\frac{RSS}{(n-2)}}.$$

Estrictamente hablando, cuando σ^2 se estima a partir de los datos deberíamos escribir $\hat{SE}(\hat{\beta}_1)$ para indicar que se ha hecho una estimación, pero para simplificar la notación nos ahorraremos este "sombrero" extra.
Los errores estándar pueden utilizarse para calcular intervalos de confianza. Un intervalo de confianza del 95% se define como un rango de valores tal que, con una probabilidad del 95%, el rango contendrá el verdadero valor desconocido del parámetro. El intervalo se define en términos de límites inferior y

superior calculados a partir de la muestra de datos. Un intervalo de confianza del 95% tiene la siguiente propiedad: si tomamos muestras repetidas y construimos el intervalo de confianza para cada muestra, el 95% de los intervalos contendrán el verdadero valor desconocido del parámetro. Para la regresión lineal, el intervalo de confianza del 95% para β_1 adopta aproximadamente la forma

$$\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1).$$

Esto, es que existe aproximadamente un 95% de que el intervalo

$$\left[\hat{\beta}_1 - 2 \cdot SE(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot SE(\hat{\beta}_1)\right]$$

contenga el verdadero valor de β_1. De manera similar un intervalo de confianza para β_0 tomará aproximadamente la forma

$$\hat{\beta}_0 \pm 2 \cdot SE(\hat{\beta}_0).$$

Los errores estándar pueden también ser usados para realizar las pruebas de hipótesis sobre los coeficientes. La prueba de hipótesis más común consiste en probar la hipótesis nula de

$$H_0 : \text{No existe una relación entre } X \text{ e } Y$$

frente a la alternativa de que

$$H_a : \text{Existe una relación entre } X \text{ e } Y.$$

Matemáticamente, esto corresponde a probar

$$H_0 : \beta_1 = 0$$

frente a

$$H_a : \beta_1 \neq 0,$$

Si, $\beta_1 = 0$, entonces el modelo $Y = \beta_0 + \beta_1 X + \epsilon$ se reduce a $Y = \beta_0 + \epsilon$, y X no está asociado con Y. Para probar la hipótesis nula, necesitamos determinar si $\hat{\beta}_1$, está lo suficientemente lejos de cero como para que podamos confiar en que $\hat{\beta}_1$ es distinto de cero. ¿Qué distancia es suficiente?, esto depende de $SE(\hat{\beta}_1)$. Si $SE(\hat{\beta}_1)$ es pequeño, entonces incluso valores relativamente pequeños de $\hat{\beta}_1$ pueden proporcionar una fuerte evidencia de que $\beta_1 = 0$, y por lo tanto de que existe una relación entre X e Y. Por el contrario, si $SE(\hat{\beta}_1)$ es grande, entonces $\hat{\beta}_1$ debe ser grande en valor absoluto para que rechacemos la hipótesis nula. En la práctica calculamos el estadístico t, dado por

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)},$$

que mide el número de desviaciones típica que $\hat{\beta}_1$ se aleja de 0. Si realmente no hay relación entre X e Y, entonces esperamos que $t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$ tenga una distribución t con $n-2$ grados de libertad. Es muy sencillo calcular la probabilidad de observar cualquier número igual a $|t|$ o mayor en valor absoluto, suponiendo que $\beta_1 = 0$. Llamamos a esta probabilidad el valor-p. Si vemos un valor-p pequeño, podemos inferir que existe una asociación entre el el predictor y la respuesta. Rechazamos la hipótesis nula, es decir declaramos que existe una relación entre X e Y. Cuando $n = 30$ los valores 5% y 1% corresponden a estadísticos t de alrededor de 2 y 2.75, respectivamente.

1.1.3 Evaluación de la precisión del modelo

Luego de rechazar la hipótesis nula, debemos cuantificar en qué medida el modelo se ajusta a los datos. La calidad del ajuste de una regresión lineal suele evaluarse con el error estándar residual (RSE) y el estadístico R^2.

Error estándar residual

El RSE es una estimación de la desviación típica de ϵ. En términos generales, es la desviación media de la respuesta con respecto a la linea de regresión real. Se calcula mediante la fórmula

$$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

Notemos que RSS ya fue definido dada la ecuación:

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

En otras palabras significa que $\hat{y}$ se desvía de la regresión real en aproximadamente RSS unidades de media. En el conjunto de datos de publicidad, el valor medio de las ventas en todos los mercados es de aproximadamente 14.000 unidades, por lo que el porcentaje de error es de $3.260/14.000 = 23\%$.
EL RSE se considera una medida de la falta de ajuste del modelo $Y = \beta_0 + \beta_1 X + \epsilon$ a los datos. , si $\hat{y}_i$ está muy lejos de y_i para una o más observaciones, entonces el RSE puede ser bastante grande, lo que indica que el modelo no se ajusta bien a los datos.

Estadístico R^2

El RSE proporciona una medida absoluta de la falta de ajuste del modelo $Y = \beta_0 + XY_1 + \epsilon$ a los datos. Pero como se mide en las unidades de Y, no siempre está claro qué constituye un buen RSE. Cómo alternativa proporcionamos una medida de ajuste llamado estadístico R^2, adopta la forma de una proporción de varianza explicada, por lo que toma valores de 0 a 1 y es independiente de la escala de Y. R^2 se calcula de la siguiente forma

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

donde $TSS = \sum (y_i + \overline{y})^2$ es la suma total de los cuadrados y RSS se define como $RSS = \sum (y_i - \hat{y}_1)^2$. La TSS, mide la varianza total de la respuesta Y, y se considera la cantidad de variabilidad inherente a la respuesta antes de realizar la regresión. Por el contrario el RSS mide la cantidad de variabilidad de la respuesta que queda sin explicar después de realizar la regresión. Por lo tanto, $TSS - RSS$ mide la cantidad de variabilidad de la respuesta que se explica o elimina al realizar la regresión. Un número cercano a 0 indica que la regresión no explica gran parte de la variabilidad en la respuesta; esto podría ocurrir porque el modelo lineal es erróneo, o la varianza de error σ^2 es alta, o ambas cosas.

A pesar de que podemos demostrar que la correlación al cuadrado (r^2)

$$\text{Cor}(X, Y) = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}}$$

es igual a R^2. El problema a la hora de aplicar regresiones multiples, es que r^2 cuantifica la asociación entre un único par de variables.

1.2 Regresión lineal múltiple

Cuando hablamos de regresión lineal múltiple, podemos asociarlo a varios modelos simple. Pero el enfoque de ajustar un modelo de regresión lineal simple independiente para cada predictor no es del todo satisfactorio. En primer lugar, no está claro cómo hacer una predicción única a partir de los varios presupuestos, ya que cada uno de los presupuestos está asociado a una ecuación de regresión independiente. En segundo lugar, cada una de las ecuaciones de regresión ignora los otros medios a la hora de calcular los coeficientes de regresión.

Supongamos que tenemos p predictores distintos. Entonces, el modelo de rerpesión lineal múltiple toma la forma

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon.$$

donde X_j representa el j-ésimo predictor y β_j cuantifica la asociación entre esa variable y la respuesta. Interpretamos β_j como el efecto medio sobre Y de un aumento de una unidad en X_j, manteniendo fijo todos los demás predictores.

1.2.1 Estimación de los coeficientes de regresión

Cómo el caso de la regresión lineal simple, los coeficientes $\beta_0, \beta_1, \ldots, \beta_p$ son desconocidos y se deben estimar. Usamos la siguiente fórmula:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p.$$

Los parámetros se estiman utilizando el mismo método de mínimos cuadrados múltiples, que se representan de forma matricial. Sea $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}$, entonces los coeficientes de regresión se eligen para minimizar la suma de los cuadrados de los residuos

$$\begin{aligned} RSS &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip}\right)^2 \end{aligned}$$

1.2.2 Algunas cuestiones importantes

Cuando realizamos una regresión lineal múltiple, normalmente estamos interesados en responder a las siguientes preguntas:

1. ¿Al menos uno de los predictores $X_1, X2, \ldots, X_p$ útil para predecir la respuesta?
2. ¿Ayudan todos los predictores a explicar Y , o sólo es útil un subconjunto de los predictores?
3. ¿En qué medida se ajusta el modelo a los datos?
4. Dado un conjunto de valores predictores, ¿qué valor de respuesta deberíamos predecir? y ¿qué precisión tiene nuestra predicción?

Uno: ¿Existe una relación entre la respuesta y los predictores?

Para determinar si existe una relación, debemos preguntarnos si todos los coeficientes de regresión son cero; es decir, si $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$. Para lo cual utilizaremos una prueba de hipótesis nula:

$$H : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

contra la hipótesis alternativa

$$H_a : \text{ al menos un } \beta_j \text{ no es cero.}$$

Esta prueba de hipótesis se realiza calculando la estadística F:

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

donde, al igual que con la regresión lineal simple, $TSS = \sum (y_i - \overline{y})^2$ y $RSS = \sum (y_i - \hat{y}_i)^2$. Si el supuesto del modelo lineal son correctos, podemos demostrar que

$$E\left[\frac{RSS}{n-p-1}\right] = \sigma^2$$

y, siempre que H_0 sea verdad

$$E\left[\frac{TSS - RSS}{p}\right] = \sigma^2$$

Por lo tanto, cuando no hay relación entre la respuesta y los predictores, se esperaría que el estadístico F tomara un valor cercano a 1. Por otro lado, si H_a es cierto, entonces $E\left[(TSS - RSS)/p\right] > \sigma^2$, así esperamos que F sea mayor que 1. Cuando interpretamos el estadístico F sugiere que al menos uno de los predictores debe estar relacionado con la respuesta. Ahora, cuando n es grande, un estadístico F que sea sólo un poco mayor que 1 podría proporcionar pruebas en contra de H_0. Por el contrario, se necesita un estadístico F mayor para rechazar H_0 si n es pequeño. Cuando H_0 es cierta y los errores e_i tiene una distribución normal, el estadístico F sigue una distribución F, donde para cualquier valor dado n y p sigue una distribución normal.

A veces queremos probar que un subconjunto particular tiene los coeficientes igual a cero. Esto corresponde a una hipótesis nula:

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0.$$

Supongamos que la suma residual de cuadrados para ese modelo es RSS_0. Entonces, podemos calcular el estadístico F:

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n - p - 1)}$$

El enfoque de utilizar un estadístico F para probar cualquier asociación entre los predictores y la respuesta funciona cuando p es relativamente pequeño, y ciertamente pequeño comparado con n.

Dos: Variables relevantes

El primer paso para analizar una regresión múltiple es calcular el estadístico F y examinar el valor-p asociado. Si a partir de ese valor p llegamos a la conclusión de que al menos uno de los predictores está relacionado con la respuesta, es natural preguntarse cuáles son los predictores causantes. Para ello realizamos una selección de variables. Por ejemplo, si $p = 2$, podemos considerar cuatro modelos: (1) un modelo que no contenga variables, (2) un modelo que contenga sólo X_1, (3) un modelo que contenga sólo X_2 y (4) un modelo que contenga tanto X_1 como X_2 . A continuación, podemos seleccionar el mejor modelo de todos los que hemos considerado. ¿Cómo se determina cuál es el mejor modelo? Para juzgar la calidad de un modelo se pueden utilizar varios estadísticos. Entre ellos se incluyen el C_p de Mallow, el criterio de información de Akaike (AIC), el criterio de información bayesiano (BIC) y el R^2 ajustado. Ahora, tomemos en cuenta que existe un total de 2^p modelos que contiene subconjuntos de p valores, probar todos es inviable. Por lo que existen tres enfoques clásicos para esta tarea:

1. Selección directa. Comenzamos con el modelo nulo, un modelo que contiene un intercepto pero no predictores. A continuación, ajustamos p regresiones lineales simples r añadimos al modelo nulo la variable que da como resultado el RSS más bajo. A continuación, añadimos a ese modelo la variable que da como resultado el RSS más bajo para el nuevo modelo de dos variables. Este proceso continúa hasta que se cumple alguna regla de parada.

2. Selección hacia atrás. Empezamos con todas las variables del modelo y eliminamos la variable con el valor p más alto, es decir, la variable menos significativa estadísticamente. Se ajusta el nuevo modelo de $(p - 1)$ variables y se elimina la variable con el valor p más alto. Este procedimiento continúa hasta que se alcanza una regla de parada. Por ejemplo, podemos detenernos cuando todas las variables restantes tengan un valor p inferior a algún umbral.

3. Selección mixta. Se trata de una combinación de selección hacia delante y hacia atrás. Empezamos sin variables en el modelo y, al igual que con la selección hacia delante, añadimos la variable que proporcione el mejor ajuste. Seguimos añadiendo variables una a una. Por supuesto, como observamos en el ejemplo de la publicidad, los valores p de las variables pueden aumentar a medida que se añaden nuevos predictores al modelo. Por lo tanto, si en algún momento el valor-p de una de las variables del modelo supera un determinado umbral, eliminamos esa variable del modelo. Seguimos realizando

estos pasos hacia delante y hacia atrás hasta que todas las variables del modelo tienen un valor-p suficientemente bajo y todas las variables fuera del modelo tendrían un valor-p grande si se añadieran al modelo.

Tres: Ajuste del modelo

Dos de la medidas numéricas más comunes del ajuste del modelo son el RSE y R^2. En la regresión múltiple resulta que R^2 es igual a $\text{Cor}(Y, \hat{Y})^2$, el cuadrado de la correlación entre la respuesta y el modelo ajustado; de esto una propiedad del modelo ajustado es que maximiza esta correlación entre todos los modelos lineales posibles. Resulta que R^2 siempre aumentará cuando se añadan más variables al modelo, incluso si esas variables sólo están débilmente asociadas con la respuesta. Esto se debe al hecho de que añadir otra variable siempre provoca una disminución de la suma residual de cuadrados en los datos de entrenamiento (aunque no necesariamente en los datos de prueba). Por lo tanto, el estadístico R^2, que también se calcula con los datos de entrenamiento, debe aumentar.

Cuatro: Predicciones

Luego de ajustar el modelo, existen tres tipos de incertidumbre asociados a esta predicción:

1. La inexactitud en las estimaciones de los coeficientes está relacionada con el error reducible. Podemos calcular un intervalo de confianza para determinar lo cerca que estará $\hat{Y}$ de $f(X)$.

2. En la práctica suponer un modelo lineal para $f(X)$ es casi siempre una aproximación a la realidad, por lo que existe una fuente adicional de error potencialmente reducible que denominamos sesgo del modelo. Así pues, cuando utilizamos un modelo lineal, en realidad estamos estimando la mejor aproximación lineal a la superficie real.

3. Nunca podemos predecir perfectamente debido al error aleatorio ϵ (error irreducible). Los intervalos de predicción son siempre más amplios que los intervalos de confianza, porque incorporan tanto el error en la estimación de $f(X)$ (el error reducible) como la incertidumbre sobre cuánto diferirá un punto individual del plano de regresión de la población (el error irreducible).

1.3 Otras consideraciones en el modelo de regresión

1.3.1 Predictores cualitativos

predictores con solo dos niveles

Supongamos que tenemos otra variable para la fórmula de regresión lineal dada por

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{si la i-ésima persona tiene una casa.} \\ \beta_0 + \beta_1 \cdot 0 + \epsilon_i & \text{si la i-ésima persona no tiene una casa.} \end{cases}$$

donde

$$x_i = \begin{cases} 1 & \text{si la i-ésima persona tiene una casa.} \\ 0 & \text{si la i-ésima persona no tiene una casa.} \end{cases}$$

Ahora, $\beta_0 + \beta_1 \cdot 0$ puede ser interpretado como el saldo promedio de las tarjetas de crédito entre quienes no poseen una casa, $\beta_0 + \beta_1$ como el saldo promedio de las tarjetas de crédito entre quienes si poseen su casa, y β_1 como la diferencia promedio en el saldo de las tarjetas de crédito entre propietarios y no propietarios.

Predictores cualitativos con más de dos niveles

Sea

$$x_{i1} = \begin{cases} 1 & \text{si la i-ésima persona es del sur.} \\ 0 & \text{si la i-ésima persona no es del sur.} \end{cases}$$

y

$$x_{i2} = \begin{cases} 1 & \text{si la i-ésima persona es del oeste.} \\ 0 & \text{si la i-ésima persona no es del oeste.} \end{cases}$$

Entonces,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{si la i-ésima persona es del sur.} \\ \beta_0 + \beta_2 + \epsilon_i & \text{si la i-ésima persona es del oeste.} \\ \beta_0 + \epsilon_i & \text{si la i-ésima persona es del este.} \end{cases}$$

Ahora β_0 puede interpretarse como el saldo medio de las tarjetas de crédito de los individuos del Este, β_1 puede interpretarse como la diferencia en el saldo medio entre las personas del Sur frente a las del Este, y β_2 puede interpretarse como la diferencia en el saldo medio entre los del Oeste frente a los del Este. Por ejemplo, el saldo estimado para la línea de base, Este, es de 531,00 dólares. Se estima que los del Sur tendrán 18,69 $ menos de deuda que los del Este, y que los del Oeste tendrán 12,50 $ menos de deuda que los del Este.
Ahora en lugar de analizar los coeficientes individuales podemos basarnos en los coeficientes individuales, utilizando una prueba F para probar $H_0 : \beta_1 = \beta_2 = 0$; esto no depende de la codificación.

1.3.2 Extenciones del modelo lineal

Dos de **los supuestos más importantes establecen que la relación entre los predictores y la respuesta son aditivos y lineales**. El supuesto de aditividad significa que la asociación entre un predictor X_j y la respuesta Y no depende de los valores de los demás predictores. El supuesto de linealidad establece que el cambio en la respuesta Y asociado a un cambio de una unidad en X_j es constante, independientemente del valor de X_j.

Eliminar la hipótesis aditiva

Consideremos el modelo de regresión lineal con dos variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Según este modelo, un aumento de una unidad en X_1 se asocia con un aumento promedio en Y de β_1 unidades. Donde β_2 no altera esta información; Una manera de ampliar este modelo es incluir un tercer predictor, llamado término de interacción, que se construye calculando el producto de X_1 y X_2. Da como resultado:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

¿Cómo la inclusión de este término de interacción relaja el supuesto aditivo? Observe que (3.31) se puede reescribir como

$$Y = \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon = \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon.$$

Dado que $\tilde{\beta}_1$ es una función de X_2, la relación entre X_1 e Y ya no es constante, ya que un cambio en el valor de X_2 cambiará la asociación entre X_1 e Y. Un argumento similar demuestra que un cambio en el valor de X_1 modifica la asociación entre X_2 e Y.
Por ejemplo, supongamos que nos interesa estudiar la productividad de una fábrica. Queremos predecir el número de unidades producidas a partir del número de líneas de producción y del número total de trabajadores. Parece probable que el efecto de aumentar el número de líneas de producción dependa del

número de trabajadores, ya que si no hay trabajadores disponibles para manejar las líneas, el aumento del número de líneas no aumentará la producción. Esto sugiere que sería apropiado incluir un término de interacción entre líneas y trabajadores en un modelo lineal para predecir las unidades. Supongamos que al ajustar el modelo obtenemos

$$\begin{aligned} \text{unidades} &= 1.2 + 3.4 \cdot \text{ lineas} + 0.22 \cdot \text{ trabajadores} + 1.4 \cdot (\text{lineas} \cdot \text{trabajadores}) \\ &= 1.2 + (3.4 + 1.4 \cdot \text{trabajadores}) \cdot \text{lineas} + 0.22 \cdot \text{trabajadores} \end{aligned}$$

En otras palabras, añadir una línea adicional aumentará el número de unidades producidas en $3.4 + 1.4 \cdot$ trabajadores.

Si la interacción entre X_1 y X_2 parece importante, debemos incluir tanto X_1 como X_2 en el modelo, aunque las estimaciones de sus coeficientes tengan valores p elevados. El fundamento de este principio es que si $X_1 \cdot X_2$ está relacionado con la respuesta, el hecho de que los coeficientes de X_1 o X_2 sean exactamente cero o no tiene poco interés. Además, $X_1 \cdot X_2$ suele estar correlacionado con X_1 y X_2, por lo que omitirlos tiende a alterar el significado de la interacción.

El concepto de interacción se aplica igualmente a las variables cualitativas o a una combinación de variables cuantitativas y cualitativas. De hecho, una interacción entre una variable cualitativa y una variable cuantitativa tiene una interpretación especialmente agradable. Consideremos el conjunto de datos de crédito y supongamos que deseamos predecir el saldo utilizando las variables de ingresos (cuantitativa) y estudiante (cualitativa). En ausencia de un término de interacción, el modelo adopta la forma

$$\begin{aligned} \text{balance} &= \beta_0 + \beta_1 \cdot \text{ingreso}_i + \begin{cases} \beta_2 & \text{si la persona i-ésima es estudiante} \\ 0 & \text{si la persona i-ésima no es estudiante.} \end{cases} \\ &= \beta_1 \cdot \text{ingreso}_i + \begin{cases} \beta_0 + \beta_2 & \text{si la persona i-ésima es estudiante} \\ \beta_0 & \text{si la persona i-ésima no es estudiante.} \end{cases} \end{aligned}$$

Observe que esto equivale a ajustar dos rectas paralelas a los datos, una para los estudiantes y otra para los no estudiantes. Esto representa una limitación potencialmente grave del modelo, ya que, de hecho, un cambio en los ingresos puede tener un efecto muy diferente en el saldo de la tarjeta de crédito de un estudiante frente a un no estudiante.
Esta limitación puede abordarse añadiendo una variable de interacción, creada multiplicando los ingresos por la variable ficticia de estudiante. Nuestro modelo modelo pasa a ser

$$\begin{aligned} \text{balance} &= \beta_0 + \beta_1 \cdot \text{ingreso}_i + \begin{cases} \beta_2 + \beta_3 \cdot \text{ingreso}_i & \text{si es estudiante} \\ 0 & \text{si no es estudiante.} \end{cases} \\ &= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \cdot \text{ingreso}_i & \text{si es estudiantes} \\ \beta_0 + \beta_1 \cdot \text{ingreso}_i & \text{si no es estudiante.} \end{cases} \end{aligned}$$

Una vez más, tenemos dos líneas de regresión diferentes para los estudiantes y los no estudiantes. Pero ahora esas líneas de regresión tienen diferentes interceptos, $\beta_0 + \beta_2$ frente a β_0, así como diferentes pendientes, $\beta_1 + \beta_3$ frente a β_1. Esto permite la posibilidad de que los cambios en los ingresos afecten de forma diferente a los saldos de las tarjetas de crédito de los estudiantes y de los no estudiantes.

Relaciones no lineales

En algunos casos, la verdadera relación entre la respuesta y los predictores puede no ser lineal. Podemos utilizar una forma cuadrática como se verá a continuación:

$$\text{mpg} = \beta_0 + \beta_1 \cdot \text{horsepower} + \beta_2 \cdot \text{horsepower}^2 + \epsilon.$$

Notemos que sigue siendo un modelo lineal; ya que $X_1 = horspower$ y $X_2 = horsepower^2$. Así que podemos estimar los coeficientes usando mínimos cuadrados para producir un ajuste no lineal. El enfoque que acabamos de describir para ampliar el modelo lineal y dar cabida a relaciones no lineales se conoce como regresión polinómica, ya que hemos incluido funciones polinómicas de los predictores en el modelo de regresión.

1.3.3 Problemas potenciales

Cuando ajustamos un modelo de regresión lineal a un conjunto de datos concreto, pueden surgir muchos problemas. Los más comunes son los siguientes

1. No linealidad de las relaciones respuesta-predictor.
2. Correlación de los términos de error.
3. Varianza no constante de los términos de error.
4. Valores atípicos.
5. Puntos de gran influencia.
6. Colinealidad.

En la práctica, identificar y superar estos problemas es tanto un arte como una ciencia.

1. No linealidad de las relaciones respuesta-predictor

Los gráficos de residuos son una herramienta gráfica útil para identificar la no linealidad. Dado un modelo de regresión lineal simple, podemos trazar los residuos, $\epsilon_i = y_i - \hat{y}_i$, frente al predictor x_i . En el caso de un modelo de regresión múltiple, puesto que hay múltiples predictores, trazamos los residuos frente a los valores predichos (o ajustados) $\hat{y}_i$. Idealmente, el gráfico de residuos no mostrará ningún patrón discernible. La presencia de un patrón puede indicar un problema con algún aspecto del modelo lineal.

Si el gráfico de residuos indica que existen asociaciones no lineales en los datos, un enfoque sencillo es utilizar transformaciones no lineales de los predictores, como $\log X$, $\sqrt{X}$ y X^2 en el modelo de regresión.

2. Correlación de los términos de error

Un supuesto importante del modelo de regresión lineal es que los términos de error, $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$, no están correlacionados. Esto significa que si los errores no están correlacionados, el hecho de que ϵ_i sea positivo proporciona poca o ninguna información sobre el signo de ϵ_{i+1}. Los errores estándar que se calculan para los coeficientes de regresión estimados o los valores ajustados se basan en el supuesto de términos de error no correlacionados. Si de hecho existe correlación entre los términos de error, los errores estándar estimados tenderán a subestimar los verdaderos errores estándar. Como resultado, los intervalos de confianza y predicción serán más estrechos de lo que deberían ser. En resumen, si los términos de error están correlacionados, podemos tener una sensación de confianza injustificada en nuestro modelo.

Por qué podrían producirse correlaciones entre los términos de error? Tales correlaciones se producen con frecuencia en el contexto de los datos de series temporales, que consisten en observaciones cuyas mediciones se obtienen en puntos discretos en el tiempo. Para determinar si éste es el caso de un determinado conjunto de datos, podemos representar gráficamente los residuos de nuestro modelo en función del tiempo. Si los errores no están correlacionados, no debería haber ningún patrón apreciable. Por otra parte, si los términos de error están correlacionados positivamente, es posible que veamos un seguimiento en los residuos, es decir, que los residuos próximos tengan valores similares. En general, el supuesto de errores no correlacionados es extremadamente importante para la regresión lineal, así como para otros métodos estadísticos, y un buen diseño experimental es crucial para mitigar el riesgo de tales correlaciones.

3. Varianza no constante de los términos de error

Otro supuesto importante del modelo de regresión lineal es que los términos de error tienen una varianza constante, $\text{Var}(\epsilon_i) = \sigma^2$. Los errores estándar, los intervalos de confianza y las pruebas de hipótesis asociadas al modelo lineal se basan en este supuesto. Se pueden identificar varianzas no constantes en los errores, o heteroscedasticidad, por la presencia de una forma de embudo en el gráfico de residuos.

La magnitud de los residuos tiende a aumentar con los valores ajustados. Ante este problema, una posible solución es transformar la respuesta Y utilizando una función cóncava como $\log Y$ o $\sqrt{Y}$. Una transformación de este tipo produce una mayor contracción de las respuestas más grandes, lo que conduce a una reducción de la heteroscedasticidad.

A veces tenemos una buena idea de la varianza de cada respuesta. Por ejemplo, la i-ésima respuesta podría ser una media de ni observaciones brutas. Si cada una de estas observaciones brutas no está correlacionada con una varianza σ^2, entonces su media tiene una varianza $\sigma_i^2 = \sigma^2/n_i$. En este caso, un remedio sencillo es ajustar nuestro modelo por mínimos cuadrados ponderados, con ponderaciones proporcionales a las varianzas inversas, es decir, $w_i = n_i$ en este caso. La mayoría de los programas de regresión lineal permiten ponderar las observaciones.

Se realizará el gráfico de valores ajustados contra residuos.

4. Valores atípicos

Los valores atípicos pueden surgir por diversas razones, como el registro incorrecto de una observación durante la recogida de datos.

Es típico que un valor atípico que no tiene un valor predictor inusual tenga poco efecto en el ajuste por mínimos cuadrados. Sin embargo, aunque un valor atípico no afecte mucho al ajuste por mínimos cuadrados, puede causar otros problemas. Los gráficos de residuos pueden utilizarse para identificar valores atípicos.

Para resolver este problema, en lugar de trazar los residuos, podemos trazar los residuos estudiados, calculados dividiendo cada residuo ei por su error estándar estimado. Las observaciones cuyo valor absoluto de los residuos estudiados es superior a 3 son posibles valores atípicos. Si creemos que se ha producido un valor atípico debido a un error en la recogida o el registro de datos, una solución es simplemente eliminar la observación. Sin embargo, hay que tener cuidado, ya que un valor atípico puede indicar una deficiencia en el modelo, como la falta de un predictor.

5. Puntos de gran influencia

Las observaciones de alto apalancamiento tienden a tener un impacto considerable en la línea de regresión estimada. Es motivo de preocupación si la línea de mínimos cuadrados se ve muy afectada por sólo un par de observaciones, porque cualquier problema con estos puntos puede invalidar todo el ajuste. Por esta razón, es importante identificar las observaciones de alto apalancamiento.

Este problema es más pronunciado en entornos de regresión múltiple con más de dos predictores, porque entonces no hay una forma sencilla de trazar todas las dimensiones de los datos simultáneamente. Para cuantificar el apalancamiento de una observación, calculamos el estadístico de apalancamiento. Un valor elevado de este estadístico indica una observación con un apalancamiento elevado. Para una regresión lineal simple,

$$h_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{i'=1}^{n} (x_{i'} - \overline{x})^2}.$$

De esta ecuación se deduce claramente que h_i aumenta con la distancia de x_i a $\overline{x}$.

6. Colinealidad

La colinealidad se refiere a la situación en la que dos o más variables predictoras están estrechamente relacionadas entre sí. La presencia de colinealidad puede plantear problemas en el contexto de la regresión, ya que puede ser difícil separar los efectos individuales de las variables colineales en la respuesta.

Dado que la colinealidad reduce la precisión de las estimaciones de los coeficientes de regresión, hace que el error estándar de β_j aumente. Recordemos que el estadístico t para cada predictor se calcula dividiendo β_j por su error estándar. En consecuencia, la colinealidad provoca una disminución del estadístico t. Como resultado, en presencia de colinealidad, es posible que no rechacemos $H_0 : \beta_j = 0$. Esto significa que la potencia de la prueba de hipótesis la probabilidad de detectar correctamente un coeficiente distinto de cero se ve reducida por la colinealidad.

Una forma sencilla de detectar la colinealidad es observar la matriz de correlaciones de los predictores. Un elemento de esta matriz que sea grande en valor absoluto indica un par de variables altamente correlacionadas y, por tanto, un problema de colinealidad en los datos. Por desgracia, no todos los problemas de colinealidad pueden detectarse mediante la inspección de la matriz de correlaciones: es posible que exista colinealidad entre tres o más variables aunque ningún par de variables tenga una correlación especialmente alta. Esta situación se denomina multicolinealidad. En lugar de inspeccionar la matriz de correlaciones, una forma mejor de evaluar la multicolinealidad es calcular el factor de inflación de la varianza (VIF). El VIF es el cociente de la varianza de β_j cuando se ajusta el modelo completo dividido por la varianza de β_j si se ajusta por sí solo. El menor valor posible para VIF es 1, que indica la ausencia total de colinealidad. Normalmente, en la práctica existe una pequeña cantidad de colinealidad entre los predictores. Como regla general, un valor VIF superior a 5 o 10 indica una cantidad problemática de colinealidad. El VIF de cada variable puede calcularse mediante la fórmula

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}.$$

donde $R^2_{X_j|X_{-j}}$ es el R^2 de una regresión de X_j sobre todos los otros predictores. Si $R^2_{X_j|X_{-j}}$ es cercano a uno, entonces existe colinealidad, y por tanto el VIF será grande.

Cuando nos enfrentamos al problema de la colinealidad, hay dos soluciones sencillas. La primera es eliminar una de las variables problemáticas de la regresión. Por lo general, esto puede hacerse sin comprometer mucho el ajuste de la regresión, ya que la presencia de colinealidad implica que la información que esta variable proporciona sobre la respuesta es redundante en presencia de las otras variables. La segunda solución consiste en combinar las variables colineales en un único predictor. Por ejemplo, podríamos tomar la media de las versiones estandarizadas de límite y calificación para crear una nueva variable que mida la solvencia.

1.4 Comparación de la regresión lineal con K-vecinos más cercanos

Los métodos paramétricos tienen varias ventajas. Suelen ser fáciles de ajustar, porque sólo es necesario estimar un pequeño número de coeficientes. En el caso de la regresión lineal, los coeficientes tienen interpretaciones sencillas y las pruebas de significación estadística pueden realizarse fácilmente. Pero los métodos paramétricos tienen una desventaja: por construcción, hacen fuertes suposiciones sobre la forma de f (X). Si la forma funcional especificada dista mucho de la realidad, y nuestro objetivo es la precisión de la predicción, el método paramétrico no funcionará bien.

Por el contrario, los métodos no paramétricos no asumen explícitamente una forma paramétrica para $f(X)$ y, por lo tanto, proporcionan un enfoque alternativo y más flexible para realizar la regresión. Aquí consideramos uno de los métodos no paramétricos más sencillos y conocidos, la regresión K-nearest

neighbors (regresión KNN). El método de regresión KNN está estrechamente relacionado con el clasificador KNN. Dado un valor para K y un punto de predicción x_0 , la regresión KNN identifica primero las K observaciones de entrenamiento más cercanas a x_0 , representadas por $\mathcal{N}_0$. A continuación, estima $f(x_0)$ utilizando la media de todas las respuestas de entrenamiento en $\mathcal{N}_0$. Dicho de otro modo,

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} y_i.$$

Vemos que cuando $K = 1$, el ajuste KNN interpola perfectamente las observaciones de entrenamiento y, en consecuencia, adopta la forma de una función escalonada. Cuando $K = 9$, el ajuste KNN sigue siendo una función escalonada, pero al promediar nueve observaciones se obtienen regiones mucho más pequeñas de predicción constante y, en consecuencia, un ajuste más suave. En general, el valor óptimo de K dependerá de la relación entre sesgo y varianza. Un valor pequeño de K proporciona la mayor precisión posible. Un valor pequeño de K proporciona el ajuste más flexible, que tendrá un sesgo bajo pero una varianza alta. Esta varianza se debe al hecho de que la predicción en una región dada depende totalmente de una sola observación. Por el contrario, valores mayores de K proporcionan un ajuste más suave y menos variable; la predicción en una región es una media de varios puntos, por lo que cambiar una observación tiene un efecto menor. Sin embargo, el suavizado puede causar sesgos al ocultar parte de la estructura de $f(X)$.

¿En qué situación un enfoque paramétrico, como la regresión lineal por mínimos cuadrados, superará a un enfoque no paramétrico, como la regresión KNN? La respuesta es sencilla: el enfoque paramétrico superará al no paramétrico si la forma paramétrica que se ha seleccionado se aproxima a la verdadera forma de f.

Cuando el valor de K es grande, el rendimiento de KNN es sólo un poco peor que el de la regresión por mínimos cuadrados en términos de MSE. Su rendimiento es mucho peor cuando K es pequeño.

Observe que a medida que aumenta el grado de no linealidad, hay pocos cambios en el MSE del conjunto de prueba para el método KNN no paramétrico, pero hay un gran aumento en el MSE del conjunto de prueba de la regresión lineal.

La disminución del rendimiento a medida que aumenta la dimensión es un problema común para KNN, y resulta del hecho de que en dimensiones más altas hay efectivamente una reducción del tamaño de la muestra.

Como regla general, los métodos paramétricos tenderán a superar a los enfoques no paramétricos cuando haya un pequeño número de observaciones por predictor.