Regresion lineal simple y multiple

Evelyn Faican - Alex Guaman

Capitulo 2

¿Qué es el aprendizaje estadístico?

Para motivar nuestro estudio del aprendizaje estadístico, comenzamos con un simple ejemplo. Supongamos que somos consultores estadísticos contratados por un cliente para investigar la asociación entre la publicidad y las ventas de un determinado producto. El conjunto de datos de Publicidad consiste en las ventas de ese producto en 200 mercados diferentes, junto con los presupuestos de publicidad del producto en cada uno de esos mercados para tres medios diferentes: televisión, radio y periódicos. Los datos se muestran en la Figura 2.1. No es posible para nuestro cliente aumentar directamente las ventas del producto. Por otro lado, pueden controlar el gasto publicitario en cada uno de los tres medios. Por lo tanto, si nosotros determinar que existe una asociación entre la publicidad y las ventas, entonces podemos instruir a nuestro cliente para que ajuste los presupuestos de publicidad, indirectamente aumentando las ventas. En otras palabras, nuestro objetivo es desarrollar un modelo preciso que se puede utilizar para predecir las ventas sobre la base de los tres presupuestos de medios.

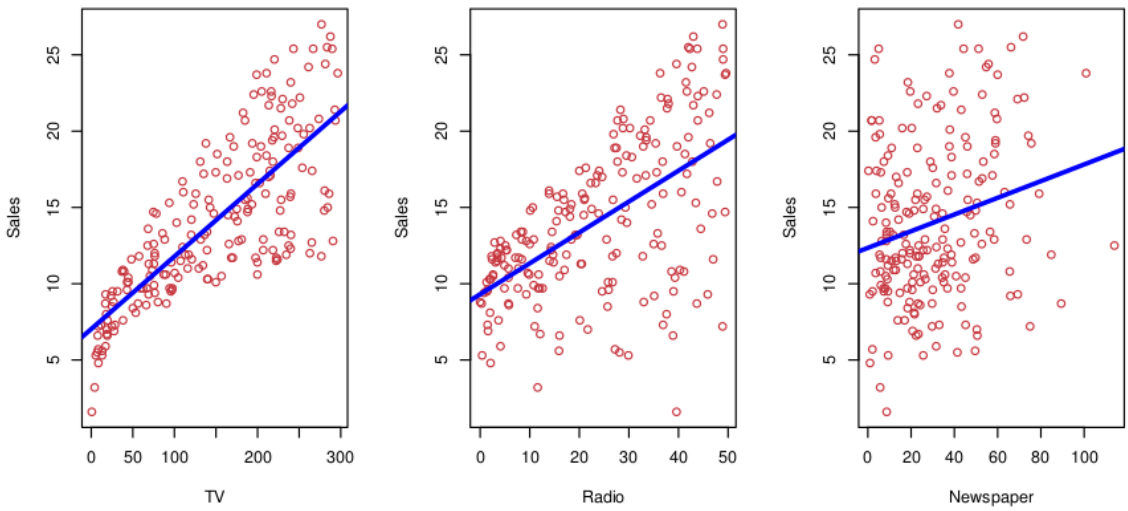


FIGURA 2.1. El conjunto de datos de publicidad. El gráfico muestra las ventas, en miles de unidades, en función de los presupuestos de TV, radio y periódicos, en miles de dólares, para 200 mercados diferentes. En cada gráfico mostramos los mínimos cuadrados simples ajuste de las ventas a esa variable, como se describe en el Capítulo 3. En otras palabras, cada azul línea representa un modelo simple que se puede usar para predecir las ventas usando TV, radio, y periódico, respectivamente.

De manera más general, supongamos que observamos una respuesta cuantitativa Y y p diferentes predictores, X1, X2, . . . , XP . Suponemos que hay alguna relación entre Y y X = (X1 , X2 , . . . , Xp), que se puede escribir en la forma muy general.

$$Y = f(X) + \epsilon. \tag{2.1}$$

Aquí f es una función fija pero desconocida de X1, . . . , Xp , y es aleatorio término de error, que es independiente de X y tiene media cero. En esta fórmulación, f representa la información sistemática que X proporciona sobre Y . Como otro ejemplo, considere el panel de la izquierda de la figura 2.2, un gráfico de ingresos versus años de educación para 30 personas en el conjunto de datos de Ingresos. La trama sugiere que uno podría predecir el ingreso usando años de educación. Sin embargo, la función f que conecta la variable de entrada a la variable de salida es en general desconocida. En esta situación se debe estimar f en base a los puntos observados. Como Ingreso es un conjunto de datos simulados, f es conocido y se muestra mediante la curva azul en el panel de la derecha de la Figura 2.2.

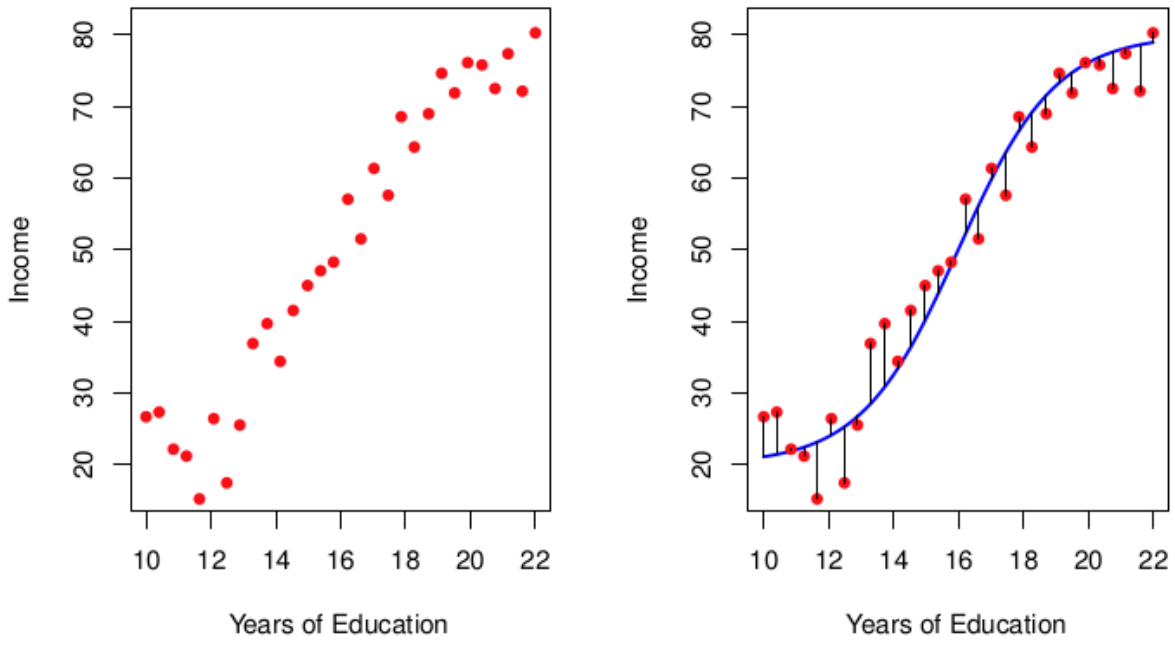


FIGURA 2.2. El conjunto de datos de ingresos. Izquierda: Los puntos rojos son los valores observados de ingresos (en decenas de miles de dólares) y años de educación para 30 personas Derecha: La curva azul representa la verdadera relación subyacente entre ingresos y años de educación, que generalmente se desconoce (pero se conoce en este caso porque los datos fueron simulados). Las líneas negras representan el error, asociado a cada observación. Tenga en cuenta que algunos errores son positivos (si una observacion se encuentra por encima de la curva azul) y algunos son negativos (si una observacion se encuentra debajo de la curva). En general, estos errores tienen aproximadamente una media de cero.

¿Por qué estimar f?

Hay dos razones principales por las que podemos desear estimar f: predicción e inferencia. Discutimos cada uno en su turno.

Predicción

En muchas situaciones, un conjunto de entradas X están fácilmente disponibles, pero la salida y no se puede obtener fácilmente. En este escenario, dado que el término de error promedia cero, podemos predecir y usando:

$$\hat{Y} = \hat{f}(X), \tag{2.2}$$

En general, la función f puede involucrar más de una variable de entrada. En la Figura 2.3 representamos el ingreso en función de los años de educación y antigüedad. Aquí f es una superficie bidimensional que debe ser estimada en base a los datos observados.

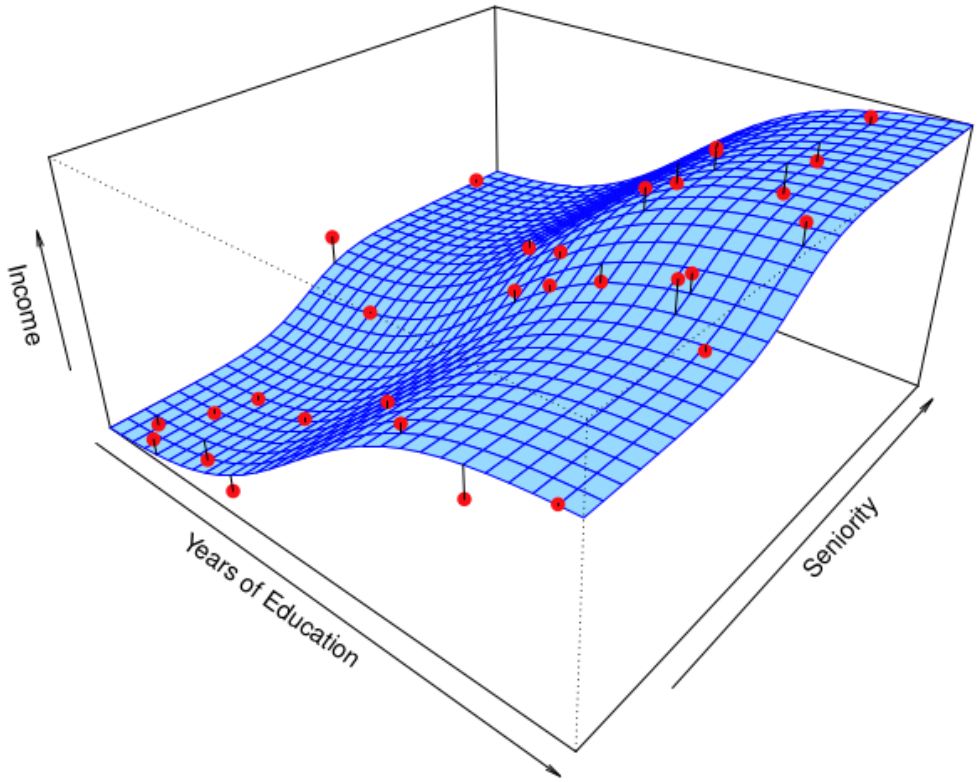


FIGURA 2.3. El gráfico muestra los ingresos como una función de los años de educación y antigüedad en el conjunto de datos Ingresos. La superficie azul representa la verdadera relación subyacente entre ingresos y años de educación y antigüedad, que se conoce ya que los datos son simulados. Los puntos rojos indican los valores observados de estas cantidades para 30 individuos.

¿Cómo estimamos f?

Se explora muchos enfoques lineales y no lineales para estimar f. Sin embargo, estos métodos generalmente comparten ciertas características. Proporcionamos una descripción general de estas características compartidas en esta sección. Siempre supondremos que hemos observado

un conjunto de n puntos de datos diferentes. Por ejemplo, en la Figura 2.2 observamos n = 30 puntos de datos.

Estas observaciones se denominan datos de entrenamiento porque usaremos estas observaciones para entrenar, o enseñar, nuestro método para estimar f . Sea xij el valor del j-ésimo predictor, o entrada, para la observación i, donde i = 1, 2, . . . , nyj = 1, 2, . . . , pag. En consecuencia, sea yi la variable de respuesta para la i-ésima observación. Entonces nuestros datos de entrenamiento consisten en

{(x1, y1), (x2, y2), . . . , (xn , yn)} donde xi = (xi1 , xi2 , . . . , xip)T .

Métodos paramétricos

Los métodos paramétricos implican un enfoque basado en modelos de dos pasos. **1.** Primero, hacemos una suposición sobre la forma funcional, o forma, apagado . Por ejemplo, una suposición muy simple es que f es lineal en

X: f (X) = 0 + 1 X1 + 2 X2 + · · · + p Xp .

Este es un modelo lineal, que se analizará extensamente en el capítulo ter 3. Una vez que hemos supuesto que f es lineal, el problema de estimación ing f se simplifica enormemente. En lugar de tener que estimar un total función p-dimensional arbitraria f (X), solo se necesita estimar los coeficientes **p + 1 0 , 1 , . . . , p . 2.** Después de seleccionar un modelo, necesitamos un procedimiento que use el datos de entrenamiento para ajustar o entrenar el modelo. En el caso del modelo lineal, necesitamos estimar los parámetros 0, 1, . . . , p . Es decir, queremos encontrar valores de estos parámetros tales que

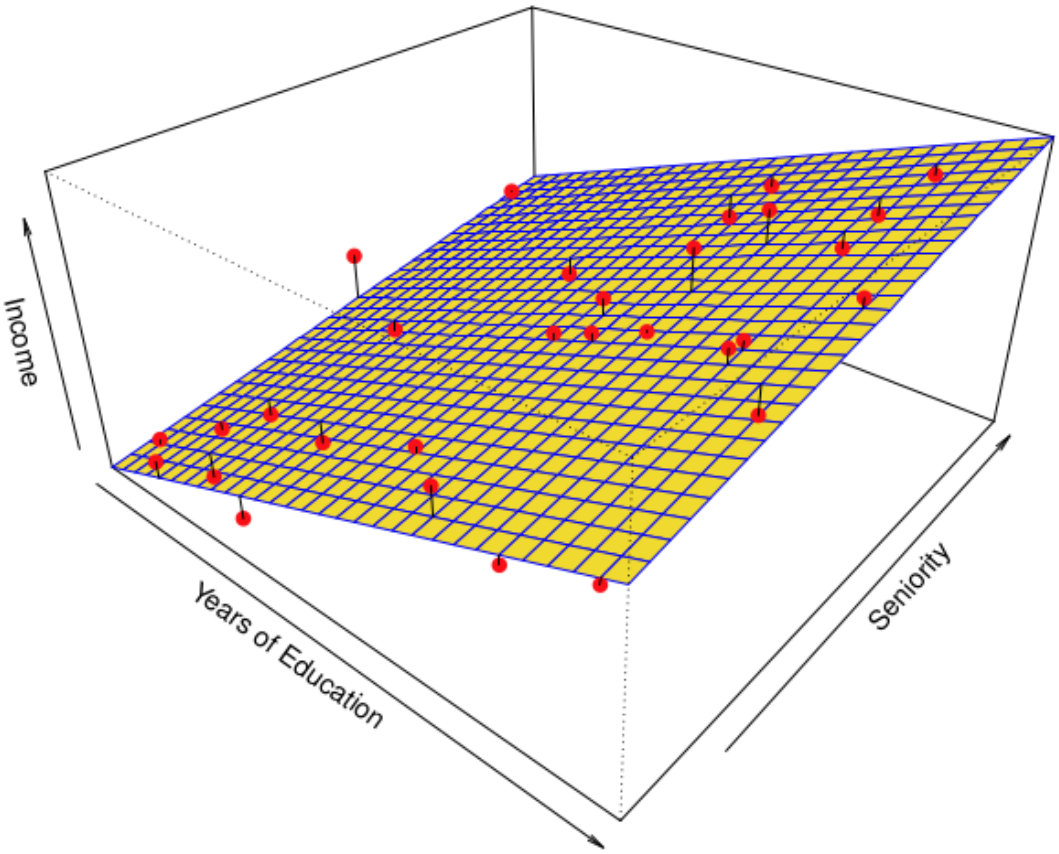
Y 0 + 1 X1 + 2 X2 + · · · + p Xp .

FIGURA 2.4. Un modelo lineal ajustado por mínimos cuadrados a los datos de ingresos de la figura 2.3. Las observaciones se muestran en rojo y el plano amarillo indica el ajuste de mínimos cuadrados a los datos.

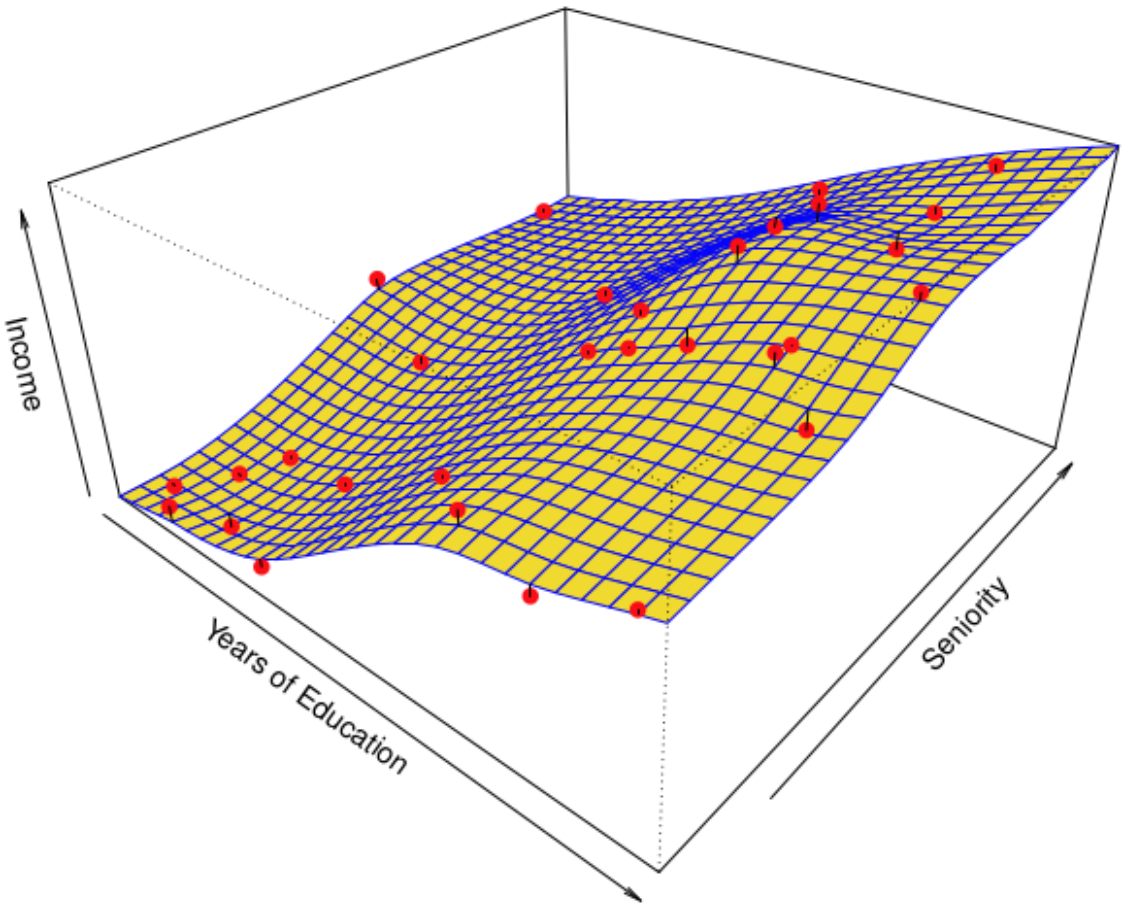
El enfoque basado en modelos que se acaba de describir se conoce como paramétrico; reduce el problema de estimar f a uno de estimar un conjunto de parámetros. Asumir una forma paramétrica para f simplifica el problema de estimar f porque generalmente es mucho más fácil estimar un conjunto de parámetros, como 0, 1, . . . , p en el modelo lineal, que ajustar una función f completamente arbitraria. La desventaja potencial de un enfoque paramétrico es que el modelo que elegimos generalmente no coincidirá con la verdadera forma desconocida de f .

La figura 2.4 muestra un ejemplo del enfoque paramétrico aplicado a los datos de ingresos de la figura 2.3. Hemos ajustado un modelo lineal de la forma

FIGURA 2.5. En amarillo, se muestra un ajuste spline de placa delgada uniforme a los datos de ingresos de la figura 2.3; las observaciones se muestran en rojo.



$$\texttt{income} \approx \beta_0 + \beta_1 \times \texttt{education} + \beta_2 \times \texttt{seniority}.$$


Income
Years of Education
Seniority

Métodos no paramétricos

Los métodos no paramétricos no hacen suposiciones explícitas sobre la forma funcional de f. En su lugar, buscan una estimación de f que se acerque lo más posible a los puntos de datos sin ser demasiado tosco o ondulado. Dichos enfoques pueden tener una gran ventaja sobre los enfoques paramétricos: al evitar la suposición de una forma funcional particular para f , tienen el potencial de adaptarse con precisión a una gama más amplia de formas posibles para f . Cualquier enfoque paramétrico trae consigo la posibilidad de que la forma funcional utilizada para estimar f sea muy diferente de la f verdadera, en cuyo caso el modelo resultante no se ajustará bien a los datos.

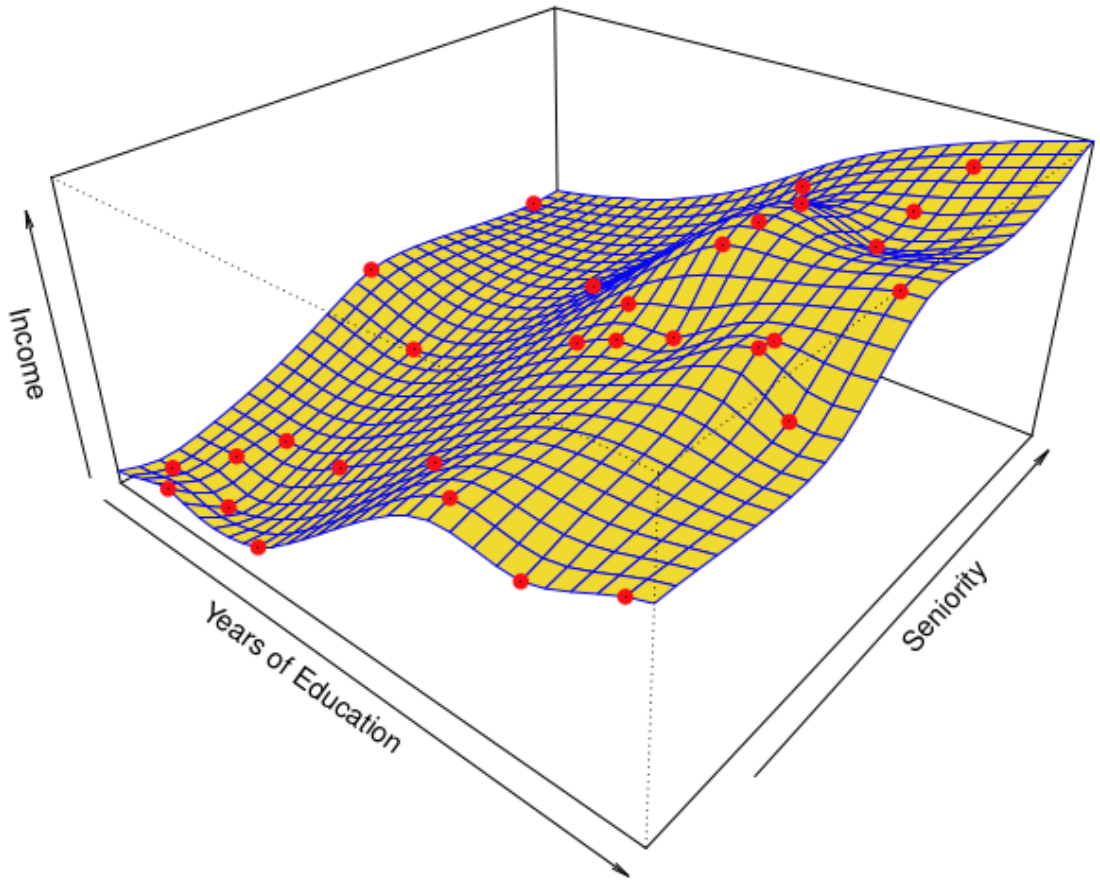


FIGURA 2.6. Un spline de placa delgada áspera se ajusta a los datos de ingresos de la figura 2.3. Este ajuste no comete errores en los datos de entrenamiento.

La compensación entre la precisión de la predicción y el modelo Interpretabilidad

De los muchos métodos que examinamos en este libro, algunos son menos flexibles o más restrictivos, en el sentido de que pueden producir solo un rango relativamente pequeño de formas para estimar f. Por ejemplo, la regresión lineal es un enfoque relativamente inflexible, porque solo puede generar funciones lineales como las líneas que se muestran en la Figura 2.1 o el plano que se muestra en la Figura 2.4.

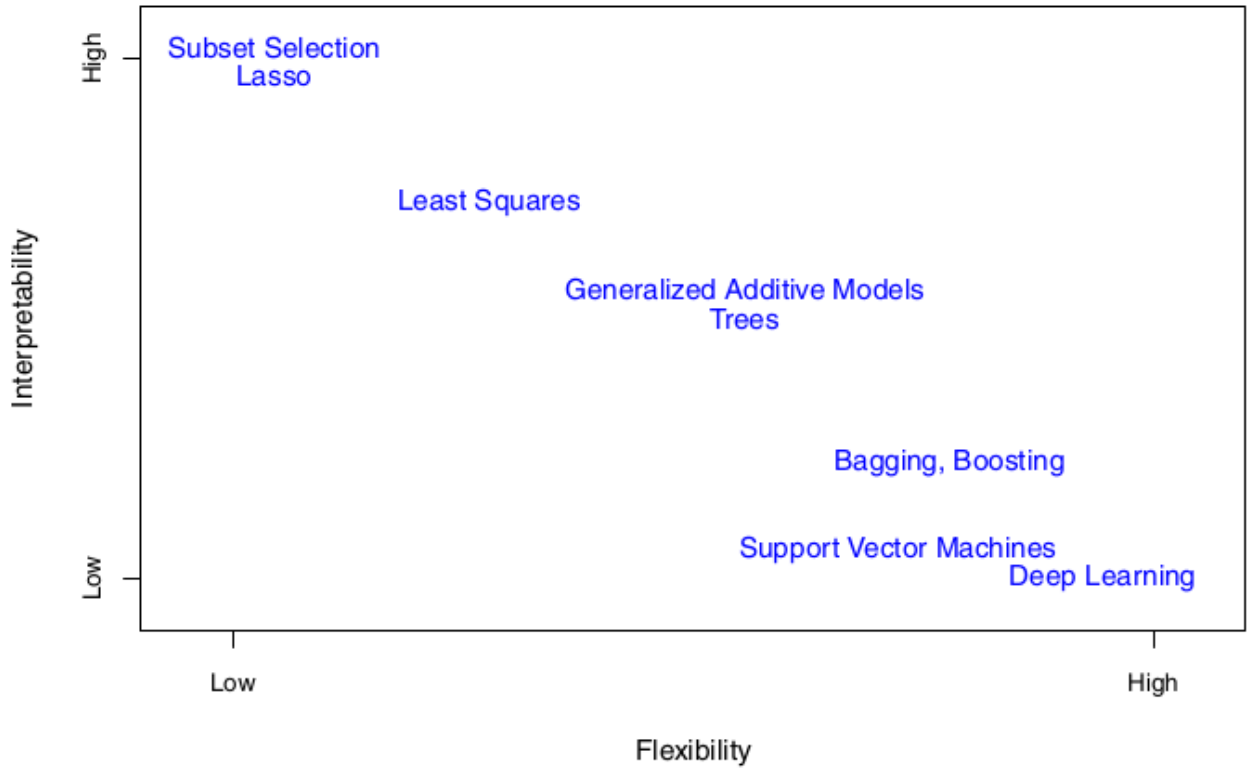


FIGURA 2.7. Una representación de la compensación entre flexibilidad e interpretabilidad, utilizando diferentes métodos de aprendizaje estadístico. En general, a medida que aumenta la flexibilidad de un método, disminuye su interpretabilidad.

Hemos establecido que cuando la inferencia es el objetivo, existen claras ventajas en el uso de métodos de aprendizaje estadístico simples y relativamente inflexibles. En algunos entornos, sin embargo, solo estamos interesados en la predicción, y la interpretabilidad del modelo predictivo simplemente no es de interés. Por ejemplo, si buscamos desarrollar un algoritmo para predecir el precio de una acción, nuestro único requisito para el algoritmo es que prediga con precisión; la interpretabilidad no es una preocupación. En este escenario, podemos esperar que sea mejor usar el modelo más flexible disponible.

Aprendizaje supervisado versus no supervisado

La mayoría de los problemas de aprendizaje estadístico se clasifican en una de dos categorías:

supervisado o no supervisado. Todos los ejemplos que hemos discutido hasta ahora en este capítulo caen en el dominio del aprendizaje supervisado.

Por el contrario, el aprendizaje no supervisado describe la situación algo más desafiante en la que para cada observación i = 1, . . . , n, observamos un vector de medidas xi pero sin respuesta asociada yi .

Deseamos ajustar un modelo que relacione la respuesta con los predictores, con el objetivo de predecir con precisión la respuesta para futuras observaciones (predicción) o comprender mejor la relación entre

la respuesta y los predictores (inferencia). No es posible ajustar un modelo de regresión lineal, ya que no hay una variable de respuesta que predecir. En este escenario, en cierto sentido estamos trabajando a ciegas; la situación se denomina no supervisada porque carecemos de una variable de respuesta que pueda supervisar nuestro análisis. ¿Qué tipo de análisis estadístico es posible? Podemos buscar comprender las relaciones entre las variables o entre las observaciones. Una herramienta de aprendizaje estadístico que podemos usar.

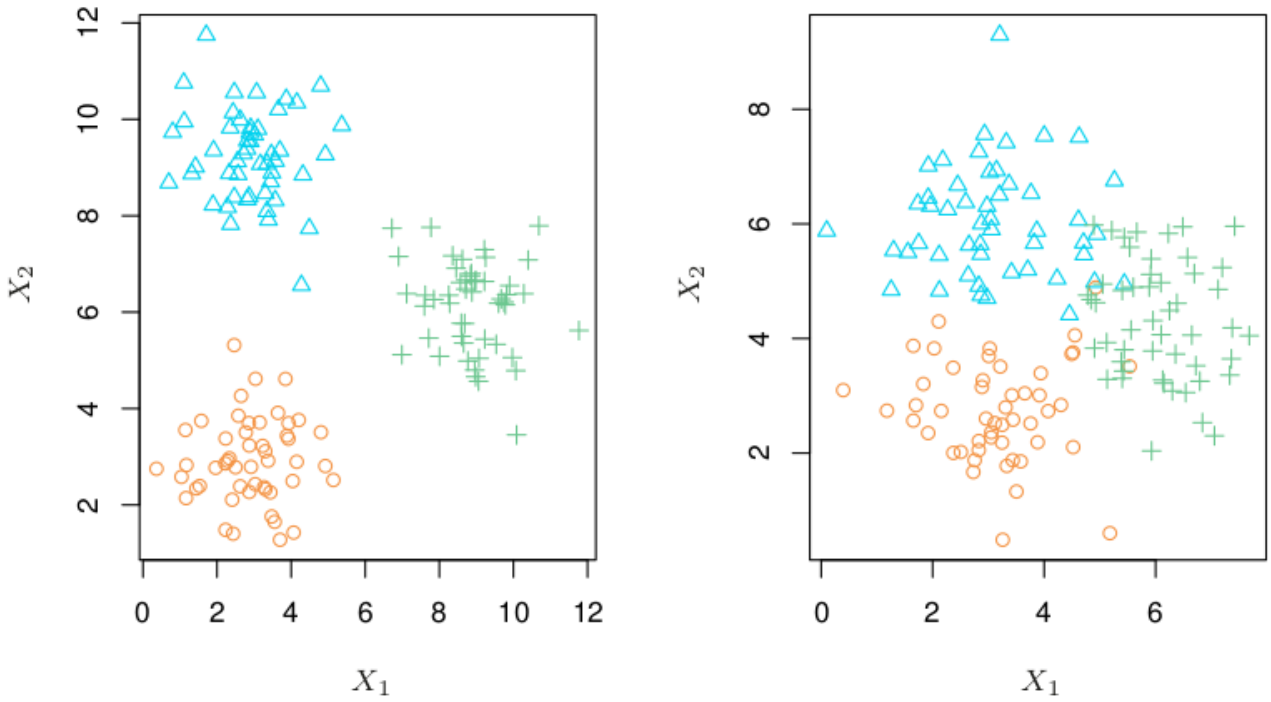


FIGURA 2.8. Un conjunto de datos de agrupamiento que involucra tres grupos. Cada grupo se muestra con un símbolo de color diferente. Izquierda: Los tres grupos están bien separados.

En este contexto, un enfoque de agrupamiento debería identificar con éxito los tres grupos. Derecha: Hay cierta superposición entre los grupos. Ahora la tarea de agrupamiento es más desafiante.

Muchos problemas caen naturalmente en los paradigmas de aprendizaje supervisado o no supervisado. Sin embargo, a veces la cuestión de si un análisis debe considerarse supervisado o no supervisado es menos clara. Por ejemplo, supongamos que tenemos un conjunto de n observaciones. Para m de las observaciones, donde $m < n$, tenemos medidas predictoras y una medida de respuesta. Para las $n - m$ observaciones restantes, tenemos mediciones predictivas pero no mediciones de respuesta. Tal escenario puede surgir si los predictores se pueden medir de manera relativamente económica, pero las respuestas correspondientes son mucho más costosas de recopilar.

Problemas de regresión versus clasificación

Las variables se pueden caracterizar como cuantitativas o cualitativas (también conocidas como categóricas). Las variables cuantitativas toman valores numéricos. Los ejemplos incluyen la edad, la altura o los ingresos de una persona, el valor de una casa y el precio de una acción. Por el contrario, las variables cualitativas toman valores en una de K clases o categorías diferentes. Los ejemplos de variables cualitativas incluyen el estado civil de una persona (casada o no), la marca del producto comprado (marca A, B o C), si una persona no paga una deuda (sí o no) o un diagnóstico de cáncer (mielogenosis aguda). leucemia, leucemia linfoblástica aguda o sin leucemia).

Tendemos a seleccionar métodos de aprendizaje estadístico sobre la base de si la respuesta es cuantitativa o cualitativa; es decir, podríamos usar la regresión lineal cuando sea cuantitativa y la regresión logística cuando sea cualitativa. Sin embargo, si los predictores son cualitativos o cuantitativos generalmente se considera menos importante.

Evaluación de la precisión del modelo

Uno de los objetivos clave de este libro es presentar al lector una amplia gama de métodos de aprendizaje estadístico que se extienden mucho más allá del enfoque de regresión lineal estándar. ¿Por qué es necesario introducir tantos enfoques diferentes de aprendizaje estadístico, en lugar de un único método óptimo? No hay comida gratis en estadística: ningún método domina a todos los demás sobre todos los conjuntos de datos posibles. En un conjunto de datos en particular, un método específico puede funcionar mejor, pero algún otro método puede funcionar mejor en un conjunto de datos similar pero diferente. Por lo tanto, es una tarea importante decidir para cualquier conjunto dado de datos

qué método produce los mejores resultados. Seleccionar el mejor enfoque puede ser una de las partes más desafiantes de realizar el aprendizaje estadístico en la práctica.

Medición de la calidad del ajuste

Para evaluar el rendimiento de un método de aprendizaje estadístico en un conjunto de datos dado, necesitamos alguna forma de medir qué tan bien sus predicciones coinciden realmente con los datos observados. Es decir, necesitamos cuantificar hasta qué punto el valor de respuesta pronosticado para una observación determinada se acerca al valor de respuesta real para esa observación. En el entorno de regresión, la medida más utilizada es el error cuadrático medio (MSE), dado por:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2,$$

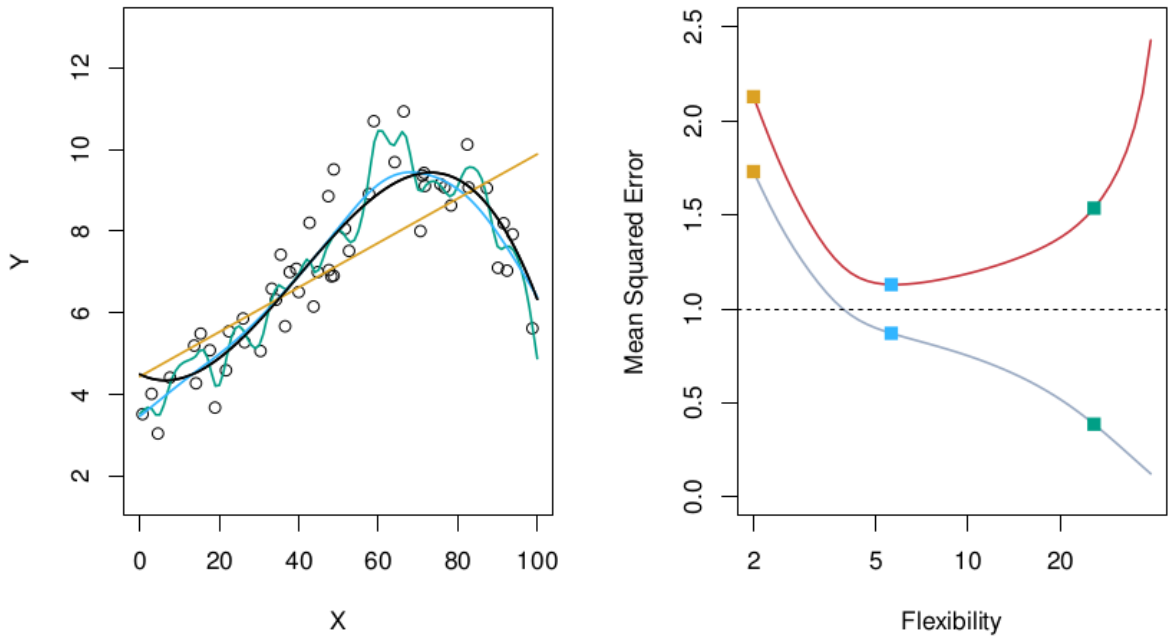
Formula (2.5)

Expresándolo más matemáticamente, supongamos que ajustamos nuestro método de aprendizaje estadístico a nuestras observaciones de entrenamiento {(x1, y1), (x2, y2), . . . , (xn , yn)}, y obtenemos la estimación fˆ. Entonces podemos calcular fˆ(x1), fˆ(x2), . . . , fˆ(xn). Si estos son aproximadamente iguales a y1 , y2 , . . . , yn , entonces el MSE de entrenamiento dado por (2.5) es pequeño. Sin embargo, en realidad no nos interesa saber si fˆ(xi) yi ; en su lugar, queremos saber si fˆ(x0) es aproximadamente igual a y0 , donde (x0 , y0) es una observación de prueba no vista anteriormente que no se usa para entrenar el método de aprendizaje estadístico. Queremos elegir el método que proporcione el MSE de prueba más bajo, a diferencia del MSE de entrenamiento más bajo.

$$\text{Ave}(y_0 - \hat{f}(x_0))^2,$$

Formula (2.6)

FIGURA 2.9. Izquierda: Datos simulados de f , mostrados en negro. Se muestran tres estimaciones de f: la línea de regresión lineal (curva naranja) y dos ajustes spline de suavizado (curvas azul y verde). Derecha: MSE de entrenamiento (curva gris), MSE de prueba (curva roja) y MSE de prueba mínimo posible sobre todos los métodos (línea discontinua). Los cuadrados representan los MSE de entrenamiento y prueba para los tres ajustes que se muestran en el panel de la izquierda.

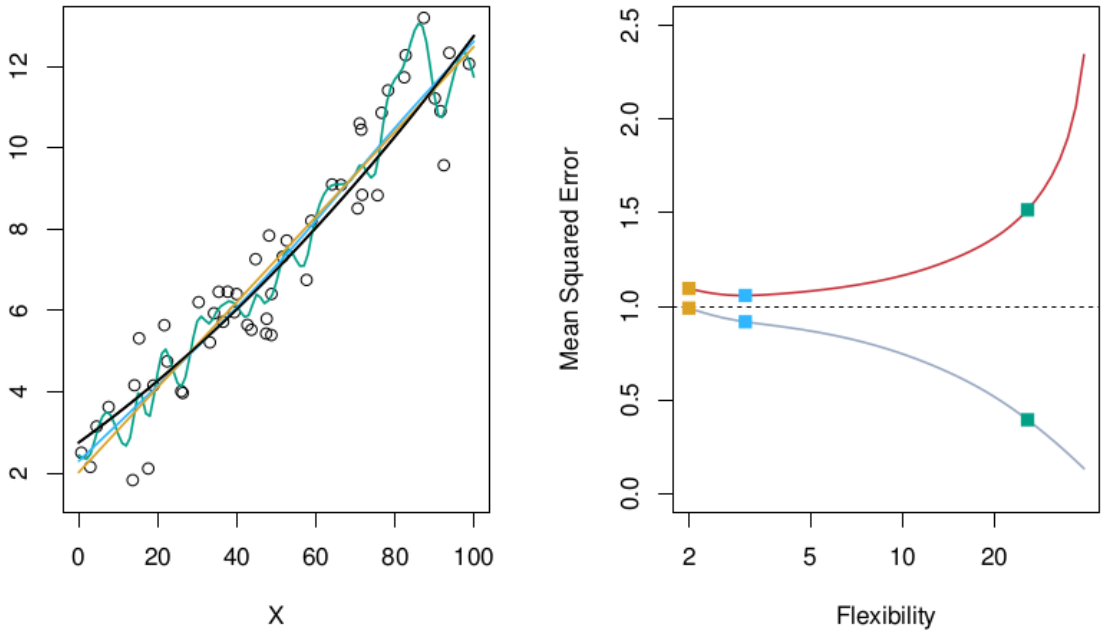


Desafortunadamente, hay un problema fundamental con esta estrategia: hay no hay garantía de que el método con el MSE de entrenamiento más bajo también tienen el MSE de prueba más bajo. A grandes rasgos, el problema es que muchos los métodos estadísticos estiman específicamente los coeficientes para minimizar el MSE del conjunto de entrenamiento. Para estos métodos, el MSE del conjunto de entrenamiento puede ser bastante pequeño, pero el MSE de prueba suele ser mucho mayor. La Figura 2.9 ilustra este fenómeno con un ejemplo simple. En el panel de la izquierda de la figura 2.9, hemos generado observaciones a partir de (2.1) con la verdadera f dada por la curva negra. Las curvas naranja, azul y verde ilustran tres posibles estimaciones de f obtenidas utilizando métodos con niveles crecientes de flexibilidad. La línea naranja es el ajuste de regresión lineal, que es relativamente inflexible.

En el panel de la derecha de la Figura 2.9, como la flexibilidad de la estadística aumenta el método de aprendizaje, observamos una disminución monótona en el MSE de entrenamiento y una forma de U en el MSE de prueba. Esta es una propiedad fundamental del aprendizaje estadístico que se mantiene independientemente del conjunto de datos en cuestión y del método estadístico que se utilice. A medida que aumenta la flexibilidad del modelo, el MSE de entrenamiento disminuirá, pero es posible que no lo haga el MSE de prueba. Cuando un método dado produce un MSE de entrenamiento pequeño pero un MSE de prueba grande, se dice que estamos sobreajustando los datos.

FIGURA 2.10. Los detalles son como en la figura 2.9, usando una f verdadera diferente que es mucho más cercana a la lineal. En esta configuración, la regresión lineal proporciona un muy buen ajuste a los datos.

La prueba MSE. Sin embargo, debido a que la verdad es casi lineal, el MSE de prueba solo



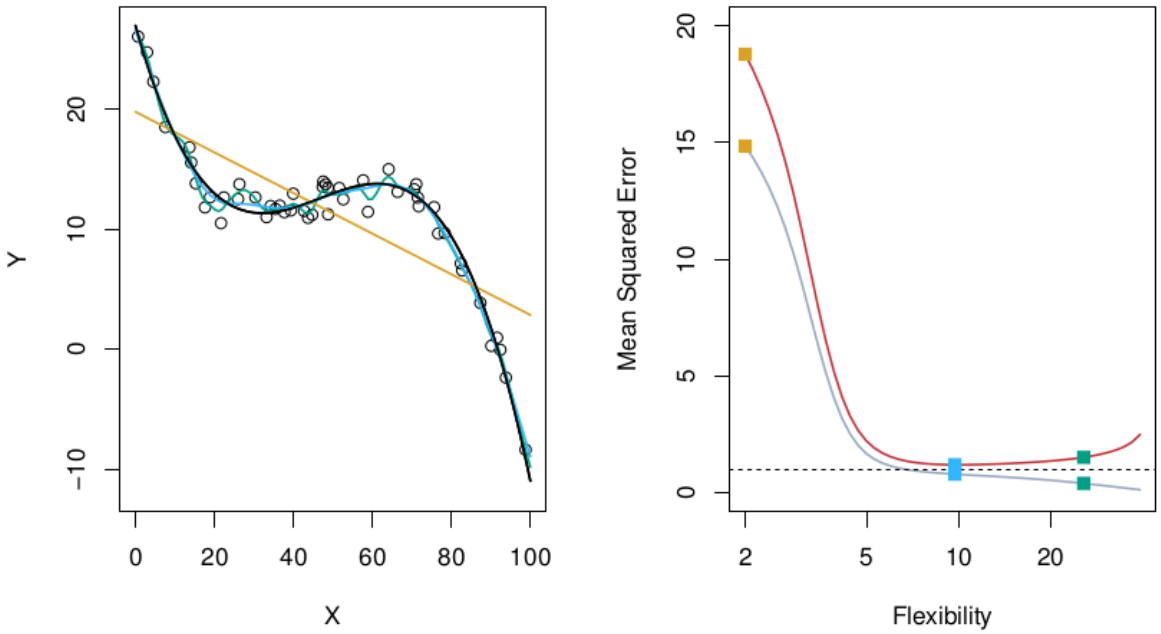
disminuye ligeramente antes de volver a aumentar, de modo que el ajuste de mínimos cuadrados naranja es sustancialmente mejor que la curva verde altamente flexible. Finalmente, la figura 2.11 muestra un ejemplo en el que f es altamente no lineal. Las curvas MSE de entrenamiento y de prueba siguen mostrando los mismos patrones generales, pero ahora hay una disminución rápida en ambas curvas antes de que la MSE de prueba comience a aumentar lentamente. En la práctica, normalmente se puede calcular el MSE de entrenamiento con relativa facilidad, pero estimar el MSE de prueba es considerablemente más difícil porque normalmente no hay datos de prueba disponibles. Como ilustran los tres ejemplos anteriores.

El nivel de flexibilidad correspondiente al modelo con el MSE de prueba mínimo puede variar considerablemente entre los conjuntos de datos.

El equilibrio entre sesgo y varianza

La forma de U observada en las curvas de prueba MSE (Figuras 2.9–2.11) resulta ser el resultado de dos propiedades en competencia de los métodos de aprendizaje estadístico. Aunque la prueba matemática está más allá del alcance de este libro, es posible demostrar que el MSE de prueba esperado, para un valor x0 dado, siempre se puede descomponer en la suma de tres cantidades fundamentales: la varianza de f^(x0), la sesgo al cuadrado de f^(x0) y la varianza del error.

FIGURA 2.11. Los detalles son como en la Figura 2.9, usando una f diferente que está lejos de ser lineal. En esta configuración, la regresión lineal proporciona un ajuste muy pobre a los datos.



términos . Eso es,

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\epsilon). \qquad (2.7)$$

Por otro lado, el sesgo se refiere al error que se introduce al aproximar un problema de la vida real, que puede ser extremadamente complicado, por un modelo mucho más simple. Por ejemplo, la regresión lineal supone que existe una relación lineal entre Y y X1, X2, . . ., XP . Es poco probable que cualquier problema de la vida real realmente tenga una relación lineal tan simple, por lo que realizar una regresión lineal sin duda dará como resultado algún sesgo en la estimación de f. En la figura 2.11, la verdadera f es sustancialmente no lineal, por lo que no importa cuántas observaciones de entrenamiento recibamos, no será posible producir una estimación precisa mediante la regresión lineal. En otras palabras, la regresión lineal da como resultado un alto sesgo en este ejemplo. Sin embargo, en la figura 2.10, la verdadera f es muy cercana a la lineal, por lo que, dados suficientes datos, debería ser posible que la regresión lineal produzca una estimación precisa. Generalmente, los métodos más flexibles resultan en menos sesgo.

Capitulo 3

Regresión lineal

La regresión lineal es una herramienta útil para predecir una respuesta cuantitativa es un método de aprendizaje estadístico útil y ampliamente utilizado

Regresión lineal simple

Se denomina como un enfoque sencillo que tiene un enfoque para predecir una respuesta cuantitativav y sobre la base de una única variable predictora X. Es muy importate que existe una relación aproximadamente lineal entre X y Y.

Matemáticamente, podemos escribir esta relación lineal como:

Y 0+ 1X.

Son dos constantes desconocidas que representan:

0 = El interceptor

1= pendiente

Representan un modelo de coeficientes de parámetros el mismo que nos permite predecir estimaciones ˆ0 y ˆ1para los coeficientes del modelo:

ŷ=ˆ0+ˆ1X

- ŷ indica una predicción de Y sobre la base deX=X.

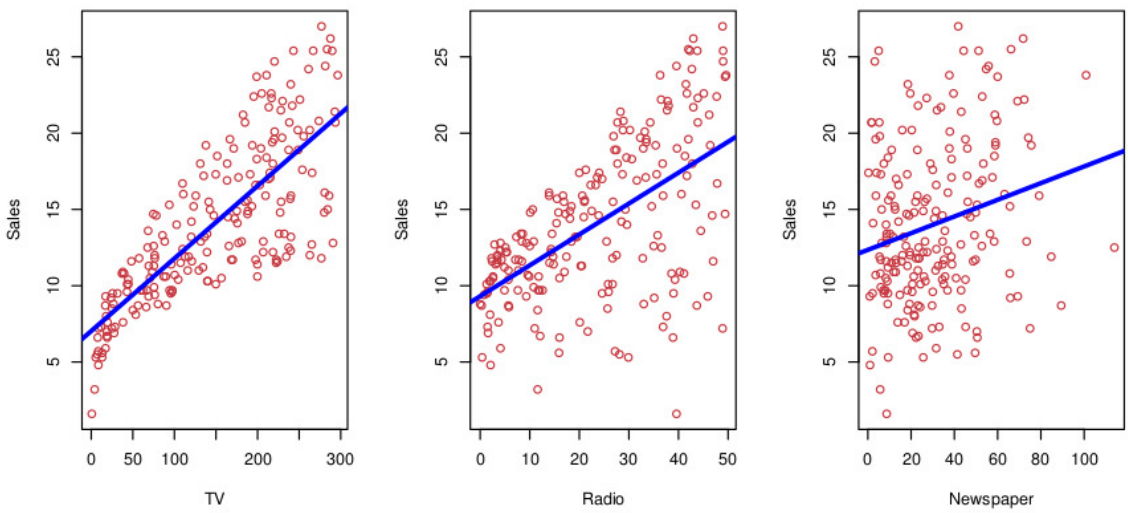
Aquí usamos un sombrero símbolo, ˆ , para denotar el valor estimado de un parámetro o coeficiente desconocido, o para denotar el valor predicho de la respuesta.

Estimación de los coeficientes

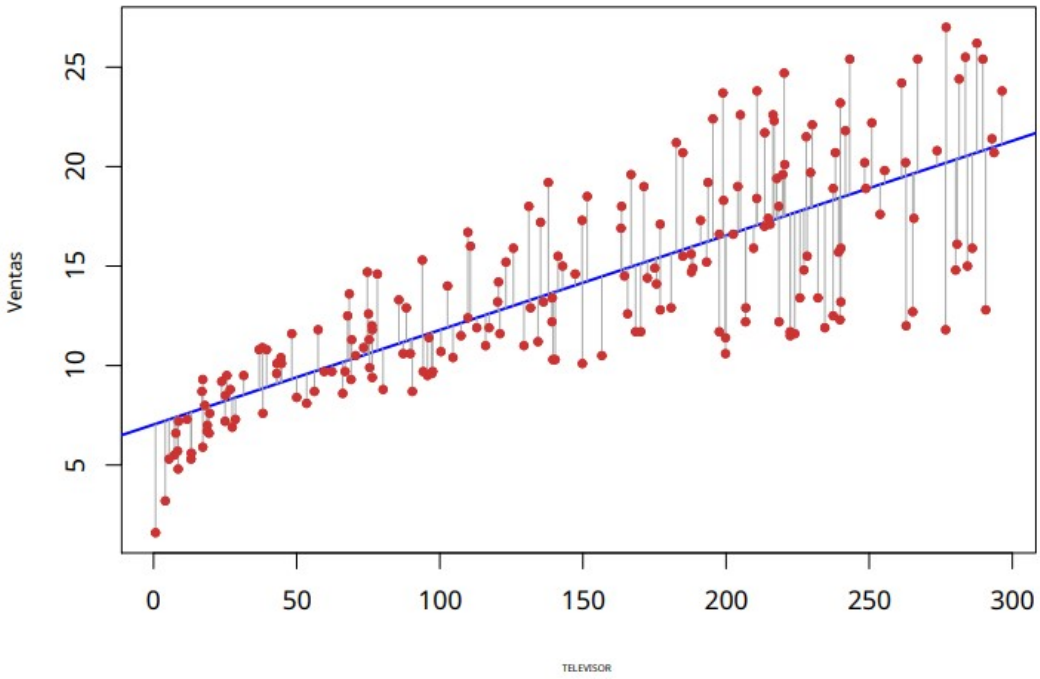
Antes de utilizar diferentes formulas comerzaremos por estimar coeficientes:

(X1, y1),(X2, y2), . . . ,(Xnorte, ynorte)

Ejemplo: Presupuesto de publicidad televisiva y las ventas de productos en norte=200 mercados diferentes



El objetivo: Es obtener estimaciones de coeficientes ˆ0 y ˆi tal que el modelo lineal se ajuste bien a los datos disponibles, hay varias formas de medircercanía. Sin embargo el enfoque más común consiste en minimizar el mínimos cuadrados

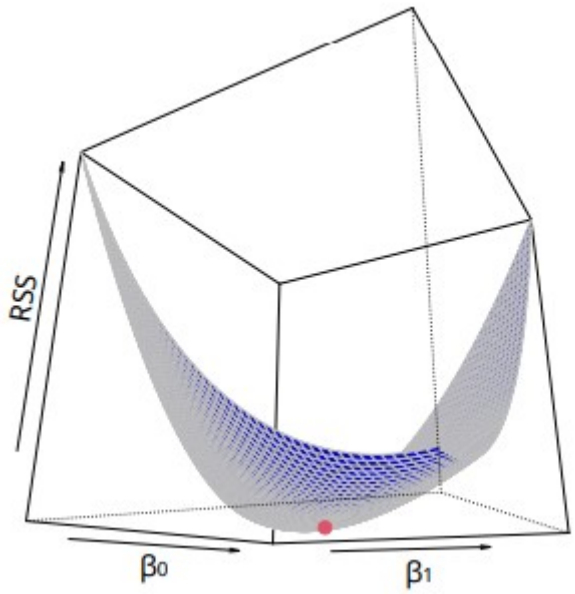


Es asi que llegamos a la conclusion que:

- Publicidad de datos y los mínimos cuadrados se ajustan a la regresión de ventas sobreTELEVISOR.
- El ajuste se obtiene minimizando la suma residual de cuadrados.
- Cada segmento de línea gris representa un residuo.
- En este caso, un ajuste lineal captura la esencia de la relación.

Existen gráficos de contorno y tridimensionales del RSS en el Publicidad de datos, utilizando ventas como respuesta y TELEVISOR como predictor.

Los puntos rojos corresponden a las estimaciones de mínimos cuadrados ˆ0y ˆ1



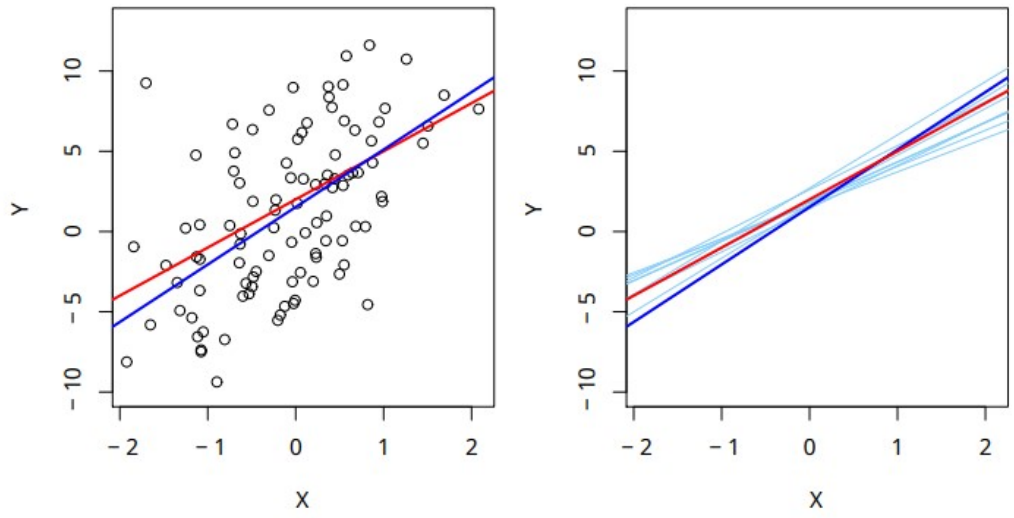
Evaluación de la precisión de las estimaciones del coeficiente La verdadera relación entre X y Y toma la forma de:

Y=F(X) +

- Para alguna función desconocida F, dónde es un término de error aleatorio de media cero.
- Si Fdebe aproximarse mediante una función lineal, entonces podemos escribir esta relación como **Y= 0+ 1X+ .**
- 0es el término de intersección en pocas palabras podemos decir que es el valor esperado de Y cuando X=0.
- 1es la pendiente asociado con un aumento de una unidad en X.

- es el error de un cajón de sastre

La línea de regresión es la mejor aproximación lineal entre X y Y



- La línea roja representa la verdadera relación,F(X) = 2 + 3X,que se conoce como la línea de regresión de la población.
- La línea azul es la línea de mínimos cuadrados; es la estimación de mínimos cuadrados para F(X) basado en los datos observados, mostrados en negro
- La línea de regresión de población se muestra nuevamente en rojo y la línea de mínimos cuadrados en azul oscuro.
- En azul claro, se muestran diez líneas de mínimos cuadrados, cada una calculada sobre la base de un conjunto aleatorio separado de observaciones.
- Cada línea de mínimos cuadrados es diferente, pero en promedio, las líneas de mínimos cuadrados están bastante cerca de la línea de regresión de población.

Es importante recalcar que la media muestral y la población son diferentes pero en general la media de la muestra proporcionará una buena estimación de la media de la población.

- La analogía entre la regresión lineal y la estimación de la media de una variable aleatoria es adecuada basada en el concepto deinclinación.
- Si usamos la media muestra lμ̂ para estimar μ, esta estimación es imparcial

La analogía con la estimación de la media poblacional μ cde una variable aleatoria Y, calculando el Error estándar de μ

Var(μ̂) = SE(μ̂)2= 2

- dónde es la desviación estándar.
- El error estándar nos dice la cantidad promedio que esta estimación μ̂ difiere del valor real de μ

Para la regresión lineal, el intervalo de confianza del 95% para 1 aproximadamente toma la forma confianza**^1±2 · SE(^1).**

Hay aproximadamente un 95% de probabilidad de que el intervalo **^1−2 · SE(^1), ^1+2 · SE(^1)**

Contendrá el verdadero valor de 1, de manera que el intervalo de confianza para 0 aproximadamente toma la forma: **^0±2 · SE(^0).**

Para probar la hipótesis nula, necesitamos determinar si^1, por lo cuál es recomendado calcular unt-estadística que mide el número de desviaciones estándar que^1está lejos de 0.

Evaluación de la precisión del modelo Al cuantificar la medida en que el modelo se ajusta a los datos. La calidad de un ajuste de regresión lineal generalmente se evalúa utilizando dos cantidades relacionadas:

- El error estándar residual (RSE) y
- El R2 estadística

Error estándar residual

Hay un término de error . Debido a la presencia de estos términos de error, incluso si conociéramos la verdadera línea de regresión.

- El RSE es una estimación de la desviación estándar de .
- En términos generales, es la cantidad promedio que la respuesta se desviará de la verdadera línea de regresión. Se calcula usando la fórmula:

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

- El RSE se considera una medida de la falta de ajuste del modelo a los datos.

- Si las predicciones obtenidas con el modelo están muy cerca de los valores reales de los resultados, es decir, siŷi yiparai=1, . . . , norte—entonces será pequeña y podemos concluir que el modelo se ajusta muy bien a los datos.
- Por otro lado, si ŷ iestá muy lejos deyipara una o más observaciones, entonces el RSE puede ser bastante grande, lo que indica que el modelo no se ajusta bien a los datos

R2 Estadística

- El RSE proporciona una medida absoluta de la falta de ajuste del modelo a los datos. Pero como se mide en las unidades de Y,no siempre está claro qué constituye una buena RSE.
- El R2 estadístico proporciona una medida alternativa de ajuste por lo que siempre toma un valor entre 0 y 1, y es independiente de la escala de Y.

$$R2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

- El R2 estadística es una medida de la relación lineal entre X y Y.
- Para probar la hipótesis nula, necesitamos determinar sî 1, nuestra estimación de 1, está lo} suficientemente lejos de cero para que podamos estar seguros de que 1es distinto de cero.
- Pero nosotros podemos rechazar la hipótesis nula—es decir podemos declararlo como una relación entre X y Y.

Evaluación de la precisión del modelo

Una vez que hemos rechazado la hipótesis nula a favor de la hipótesis alternativa , es natural querer cuantificarla medida en que el modelo se ajusta a los datos. La calidad de un ajuste de regresión lineal generalmente se evalúa utilizando dos cantidades relacionadas:

- El error estándar
- Error residual (RSE) y el
- R2 estadística.

```
### **Error estándar residual**
```

- El RSE es una estimación de la desviación estándar de . En términos generales, es la cantidad promedio que la respuesta se desviará de la verdadera línea de regresión.
- El RSE se considera una medida de lafalta de ajuste del modelo a los datos. Si las predicciones obtenidas con el modelo están muy cerca de los valores reales de los resultados.

- Podenis concluir que el modelo se ajusta muy bien a los datos.

R2 Estadística

Existen caracteristicas muy esenciales como:

- Proporciona una medida absoluta de la falta de ajuste del modelo a los datos
- Proporciona una medida alternativa de ajuste.
- Toma la forma de un proporciónque se denomina como la varianza vpor lo que siempre toma un valor entre 0 y 1.

Regresión lineal múltiple

- La regresión lineal simple es un enfoque útil para predecir una respuesta sobre la base de una única variable predictora. Sin embargo, en la práctica a menudo tenemos más de un predictor.
- El enfoque de ajustar un modelo de regresión lineal simple separado para cada predictor no es del todo satisfactorio.

Es recomendado:

Debe tener un enfoque de ajustar un modelo de regresión lineal simple separado para cada predictor no es del todo satisfactorio

- En primer lugar, no está claro cómo hacer una única predicción de ventas dados los tres presupuestos de medios publicitarios, ya que cada uno de los presupuestos está asociado con una ecuación de regresión separada.
- Segundo, cada una de las tres ecuaciones de regresión ignora los otros dos medios al formar estimaciones para los coeficientes de regresión.

En lugar de ajustar un modelo de regresión lineal simple separado para cada predictor, un mejor enfoque es extender el modelo de regresión lineal simple para que pueda acomodar directamente múltiples predictores.

Estimación de los coeficientes de regresión

- Los parámetros se estiman utilizando el mismo enfoque de mínimos cuadrados que vimos en el contexto de la regresión lineal simple.

- El enfoque de usar estadística para probar cualquier asociación entre los predictores y la respuesta funciona cuandopages relativamente pequeño, y ciertamente pequeño en comparación connorte

Selección de reenvío.

- Empezamos con elmodelo nulo—un modelo que conadelante contiene un intercepto pero no predictores. Luego encajamospagre-lineal simpleselección regresiones y agregar al modelo nulo la variable que resulta en elmodelo nulo RSS más bajo

Extensiones del Modelo Lineal

- El modelo de regresión lineal estándar (3.19) proporciona resultados interpretables y funciona bastante bien en muchos problemas del mundo real.

Correlación de términos de error

- Una suposición importante del modelo de regresión lineal es que los términos de error, 1, 2, . . . , norte, no están correlacionados.

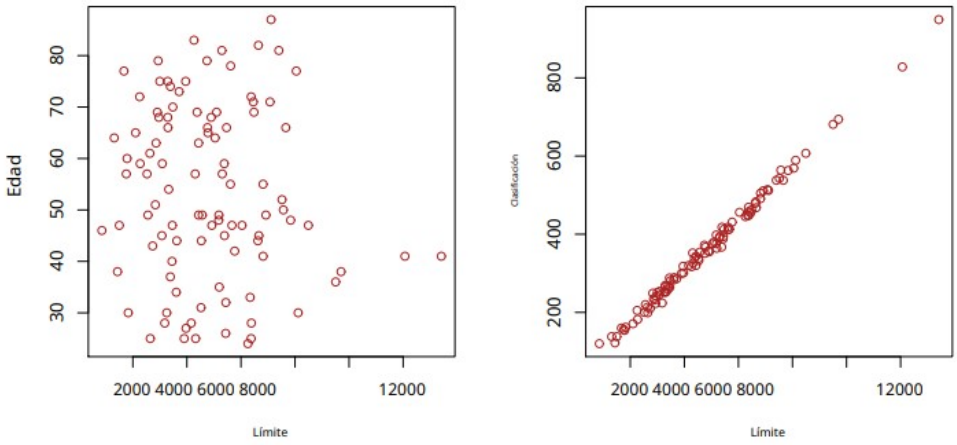
Variación no constante de los términos de error

- Los errores estándar, los intervalos de confianza y las pruebas de hipótesis asociadas con el modelo lineal se basan en esta suposición.

Altos puntos de apalancamiento

En una regresión lineal simple, las observaciones de alto apalancamiento son bastante fáciles de identificar, ya que simplemente podemos buscar observaciones para las cuales el valor del predictor está fuera del rango normal de las observaciones. Pero en una regresión lineal múltiple con muchos predictores, es posible tener una observación que esté dentro del rango de los valores de cada predictor individual, pero que sea inusual en términos del conjunto completo de predictores

Colinealidad Se refiere a la situación en la que dos o más variables predictoras están estrechamente relacionadas entre si, como se evidencia a continuación:



Comparación de regresión lineal con k-Vecinos más cercanos

- Los métodos paramétricos tienen varias ventajas.
- Suelen ser fáciles de ajustar, porque solo se necesita estimar un pequeño número de coeficientes.
- En el caso de la regresión lineal, los coeficientes tienen interpretaciones simples y las pruebas de significancia estadística se pueden realizar fácilmente.
- Pero los métodos paramétricos tienen una desventaja: por construcción, hacen fuertes suposiciones sobre la forma deF(X).
- Si la forma funcional especificada está lejos de la verdad, y nuestro objetivo es la precisión de la predicción, entonces el método paramétrico tendrá un desempeño deficiente.
- A diferencia de,no paramétricométodos no asumen explícitamente una forma paramétrica paraF(X) y, por lo tanto, proporcionar un enfoque alternativo y más flexible para realizar la regresión.

3.6 Laboratorio

Librerias

```
library(MASS)
library(ISLR2)
```

```
Attaching package: 'ISLR2'
```

```
The following object is masked from 'package:MASS':

    Boston
```

Regresión Lineal Simple

- `ISRL2` Contiene informacion de Boston
- `medv` Valor medio de la vivienda de 506 secciones ceensales de Boston
- `rm` Número medio de habitaciones por casa
- `lstat` Porcentaje de hogares con un status socioeconómico bajo

```
head(Boston)
```

```
     crim zn indus chas   nox    rm  age    dis rad tax ptratio lstat medv
1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3  4.98 24.0
2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8  9.14 21.6
3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8  4.03 34.7
4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7  2.94 33.4
5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7  5.33 36.2
6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7  5.21 28.7
```

- La función `lm()` sirve para ajustar a una regresión simple
- La función `medv()` Es la respuesta y
- La función `lstat()` Es el predictor x

El error se debe a que R no sabe donde encontrar esos valores, por lo que en la siguiente linea se coloca Boston.

```
lm.fit <- lm ( medv~lstat , data = Boston )
attach ( Boston )
lm.fit <- lm ( medv~lstat )
```

- La función `lm.fit` permite acceder a información básica sobre el modelo.
- La función `summary(lm.fit)` nos da los valores p y los errores estándar de los coeficientes, así como el estadístico R2 y el estadístico F del modelo.

```
lm.fit
```

```
Call:
lm(formula = medv ~ lstat)

Coefficients:
(Intercept)        lstat
      34.55        -0.95
```

```
summary(lm.fit)
```

```
Call:
lm(formula = medv ~ lstat)

Residuals:
    Min      1Q  Median      3Q     Max
-15.168  -3.990  -1.318   2.034  24.500

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.55384    0.56263   61.41   <2e-16 ***
lstat       -0.95005    0.03873  -24.53   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.216 on 504 degrees of freedom
Multiple R-squared:  0.5441,    Adjusted R-squared:  0.5432
F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

- La función `names()` permit4e averiguar qué otras piezas de información se almacenan en `lm.fit`. Aunque podemos extraer por su nombre `lm.fit$coefficients`. Además, se pueden utilizar las funciones extractoras como `coef()` para acceder a ellos.

```
names(lm.fit)
```

```
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"
```

```
coef(lm.fit)
```

```
(Intercept)       lstat
 34.5538409  -0.9500494
```

- La función `confint()` obtiene un intervalo de confianza para los coeficientes estimados.

```
confint(lm.fit)
```

```
                2.5 %     97.5 %
(Intercept) 33.448457 35.6592247
lstat       -1.026148 -0.8739505
```

- La función `predict()` puede utilizarse para producir intervalos de confianza y para la predicción de `medv()` para un valor dado de `lstat()`.

```
predict(lm.fit , data.frame(lstat = (c(5 , 10 , 15))) ,
interval = "confidence")
```

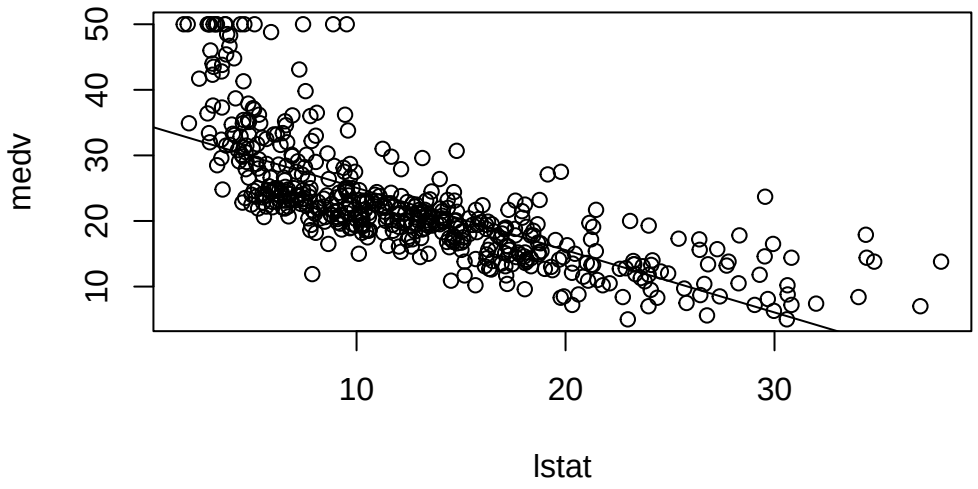
```
       fit      lwr      upr
1 29.80359 29.00741 30.59978
2 25.05335 24.47413 25.63256
3 20.30310 19.73159 20.87461
```

```
predict(lm.fit , data.frame(lstat = (c(5 , 10 , 15))) ,
interval = "prediction")
```

```
       fit       lwr      upr
1 29.80359 17.565675 42.04151
2 25.05335 12.827626 37.27907
3 20.30310  8.077742 32.52846
```

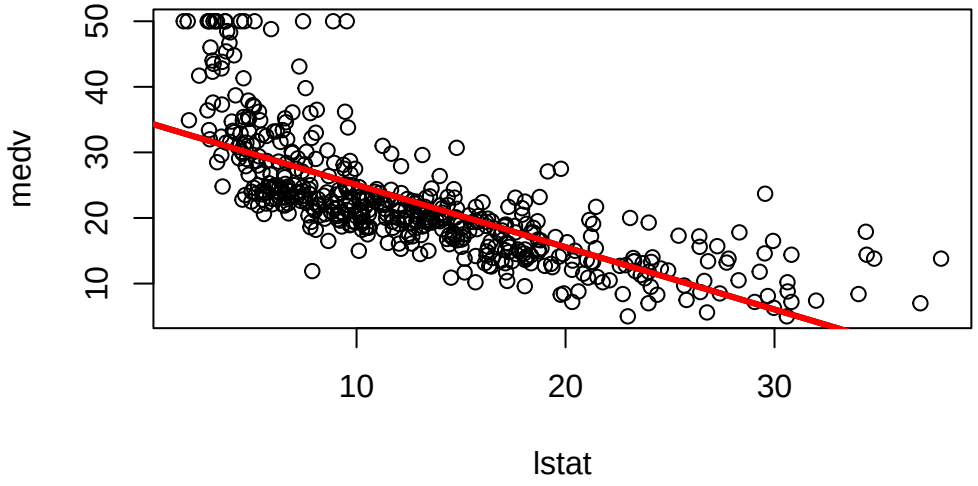
Ahora trazaremos `medv()` y `lstat()` junto con la línea de regresión por mínimos cuadrados mediante las funciones `plkot()` y `abline()`.

```
plot ( lstat , medv )
abline ( lm.fit )
```

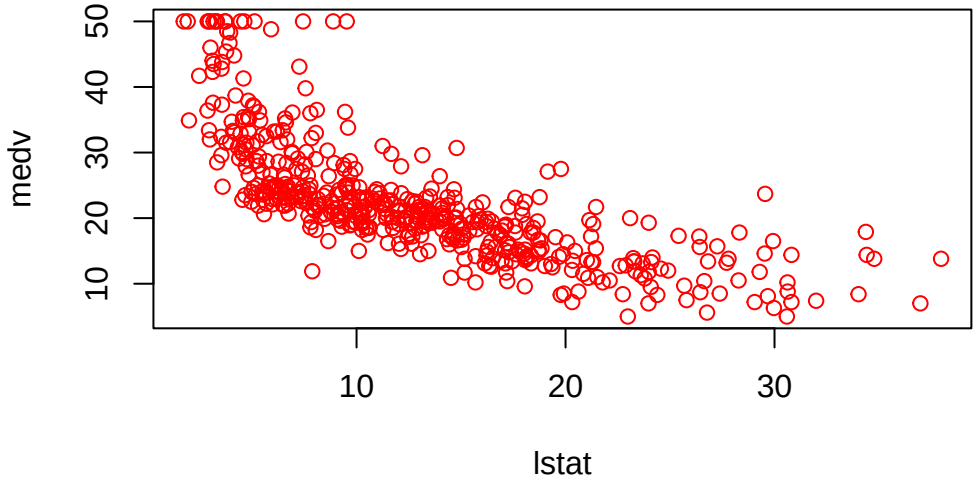


- La función `abline()` sirve para dibujar cualquier línea, no sólo la línea de regresión por mínimos cuadrados. Podemos dibujar una recta con intercepto a y pendiente b, escribimos abline(a, b).
- El comando `lwd=3` hace que la anchura de la línea de regresión se incremente en un factor de 3.
- La funcion `pch()` se utiliza para crear símbolos de trazado diferentes.

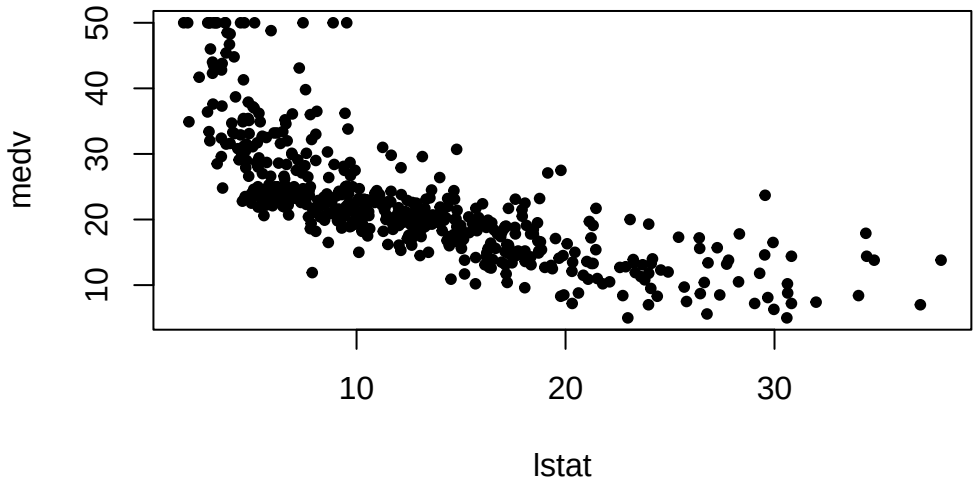
```
plot(lstat , medv )
abline(lm.fit, lwd = 3)
abline(lm.fit, lwd= 3 , col = " red ")
```



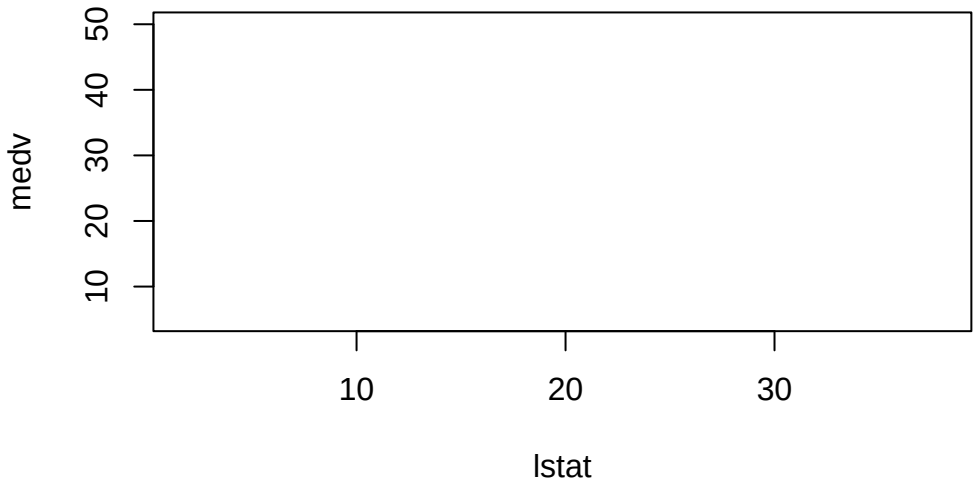

```
plot(lstat, medv , col = " red ")
```



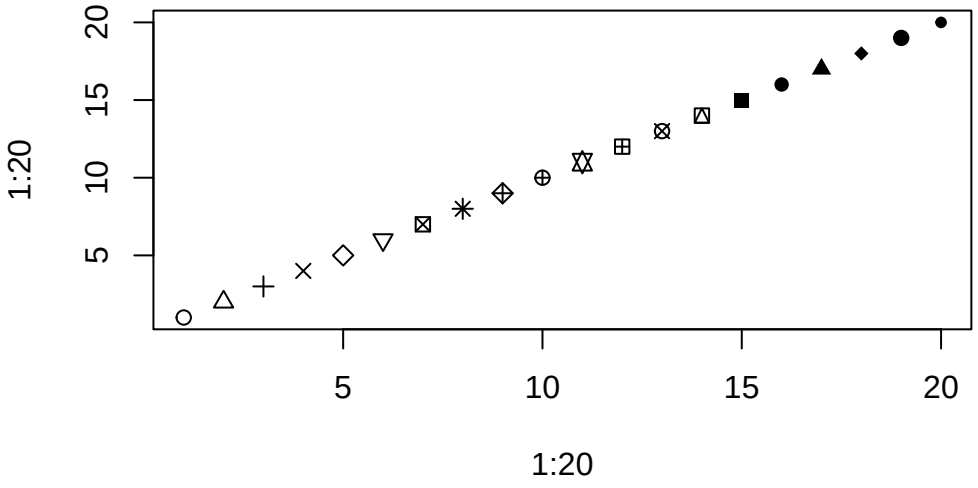
```
plot(lstat, medv , pch = 20)
```




```
plot(lstat, medv , pch = " + ")
```

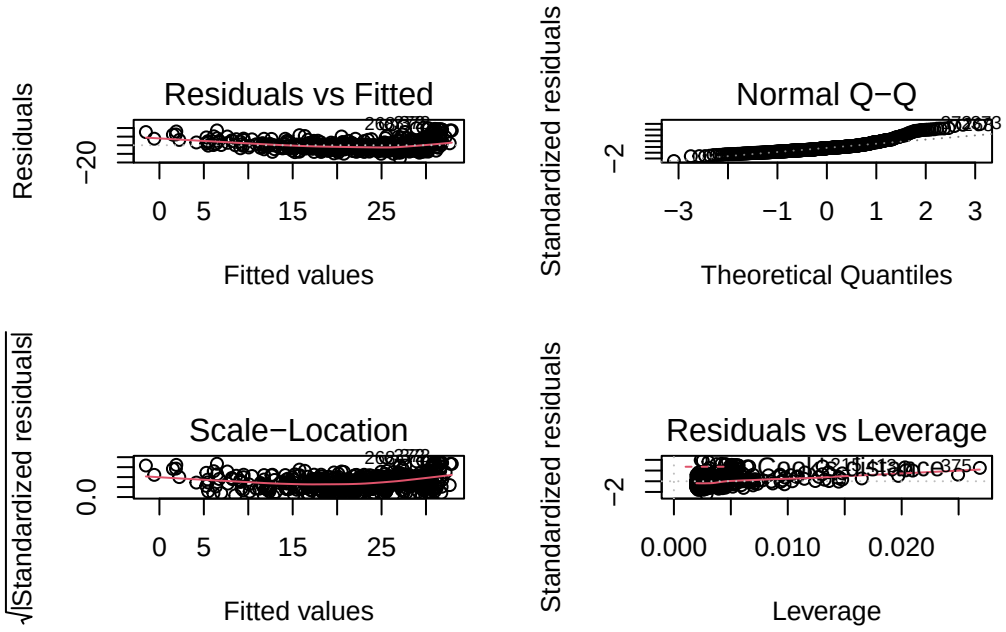



```
plot(1:20, 1:20 , pch = 1:20)
```



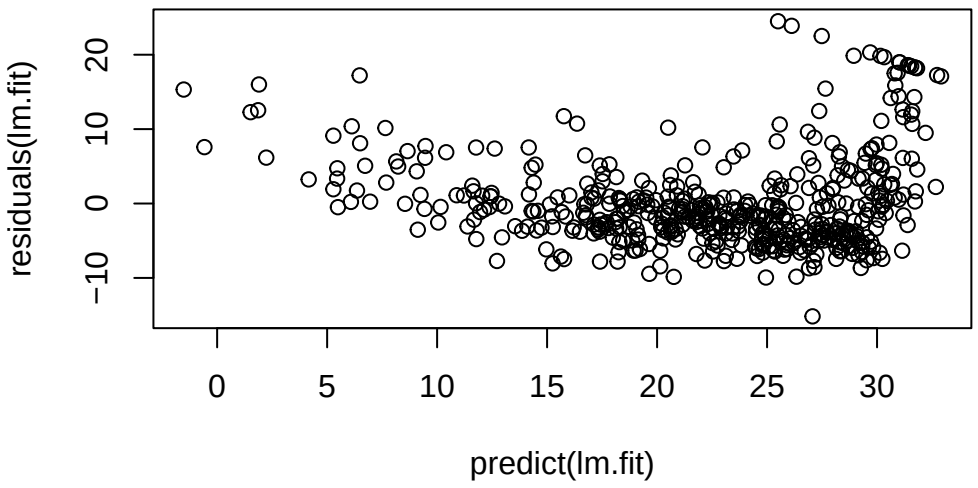
- Cuatro diagramas de diagnóstico se producen al aplicar la función `plot()` directamente a la salida de `lm()`. Este comando producirá un gráfico a la vez, y al presionar Enter se generará el siguiente gráfico. Sin embargo, es conveniente ver los cuatro gráficos juntos.
- Las funciones `par()` y `mfrow()`, dicen a R que divida la pantalla en paneles separados para ver varios gráficos simultáneamente.

```
par(mfrow = c(2,2))
plot(lm.fit)
```

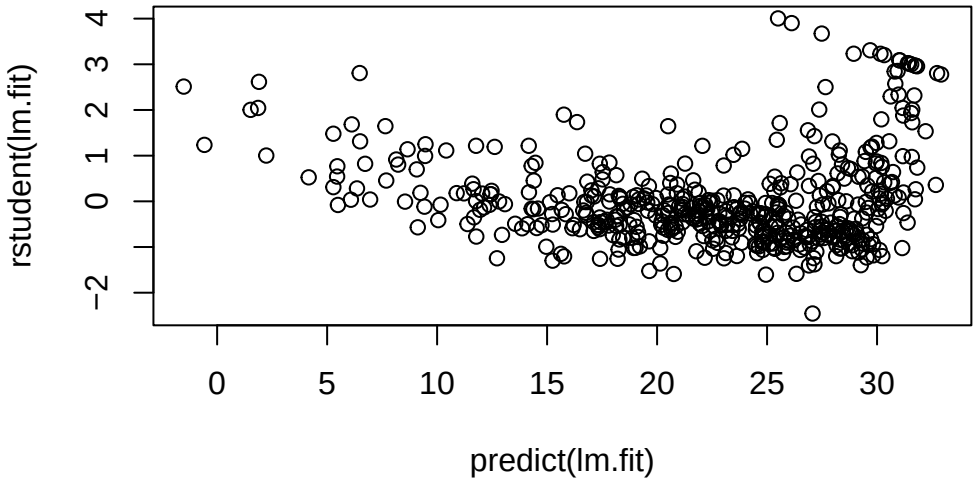


- La función `residuals()` permite calcular los residuos de un ajuste de regresión lineal.
- La función `rstudent()` devolverá los residuos estudentizados, y podemos usarla para graficar los residuos contra los valores ajustados.

```
plot(predict(lm.fit), residuals(lm.fit))
```

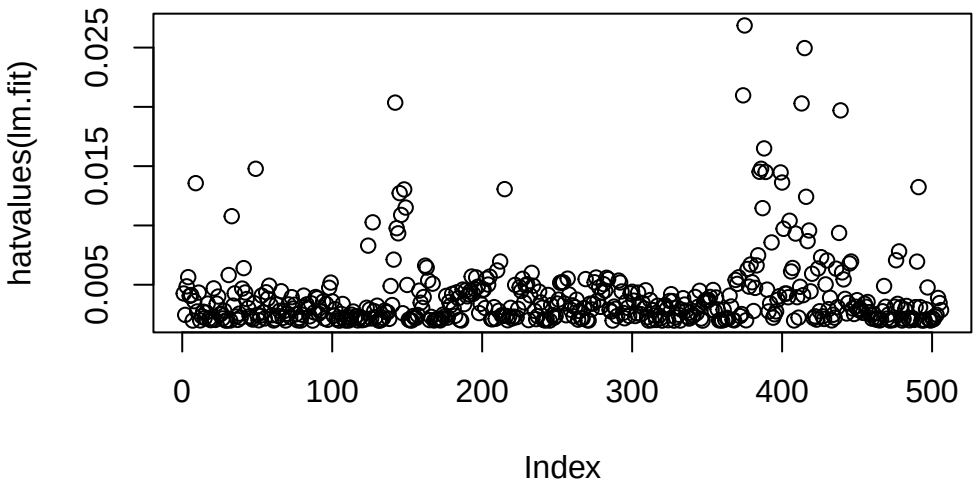



```
plot(predict(lm.fit), rstudent(lm.fit))
```



- Mediante la función `hatvalues()` se pueden calcular las estadísticas de apalancamiento para cualquier número de predictores.

```
plot(hatvalues(lm.fit))
```




```
which.max(hatvalues(lm.fit))
```

```
375
375
```

- La función `which.max()` identifica el índice del elemento más grande de un vector (observación que tiene la estadística de apalancamiento más grande).

Regresión Lineal Múltiple

- La función `lm()`se utiliza para ajustar un modelo de regresión lineal múltiple por mínimos cuadrados
- La sintaxis `lm(y  x1 + x2 + x3)` se utiliza para ajustar un modelo con tres predictores, x1, x2 y x3.

- La función `summary()` muestra ahora los coeficientes de regresión de todos los predictores.

```
lm.fit <- lm(medv ~ lstat + age, data = Boston)
summary(lm.fit)
```

```
Call:
lm(formula = medv ~ lstat + age, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-15.981  -3.978  -1.283   1.968  23.158

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.22276    0.73085  45.458  < 2e-16 ***
lstat       -1.03207    0.04819 -21.416  < 2e-16 ***
age          0.03454    0.01223   2.826  0.00491 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.173 on 503 degrees of freedom
Multiple R-squared:  0.5513,    Adjusted R-squared:  0.5495
F-statistic:   309 on 2 and 503 DF,  p-value: < 2.2e-16
```

El conjunto de datos de Boston contiene 12 variables, y no se pretende realizar una regresión utilizando todos los predictores:

```
lm.fit <- lm(medv ~ ., data = Boston)
summary(lm.fit)
```

```
Call:
lm(formula = medv ~ ., data = Boston)

Residuals:
     Min       1Q   Median       3Q      Max
-15.1304  -2.7673  -0.5814   1.9414  26.2526

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  41.617270   4.936039   8.431 3.79e-16 ***
crim         -0.121389   0.033000  -3.678 0.000261 ***
zn            0.046963   0.013879   3.384 0.000772 ***
indus         0.013468   0.062145   0.217 0.828520
chas          2.839993   0.870007   3.264 0.001173 **
nox         -18.758022   3.851355  -4.870 1.50e-06 ***
rm            3.658119   0.420246   8.705  < 2e-16 ***
age           0.003611   0.013329   0.271 0.786595
dis          -1.490754   0.201623  -7.394 6.17e-13 ***
rad           0.289405   0.066908   4.325 1.84e-05 ***
tax          -0.012682   0.003801  -3.337 0.000912 ***
ptratio      -0.937533   0.132206  -7.091 4.63e-12 ***
lstat        -0.552019   0.050659 -10.897  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.798 on 493 degrees of freedom
Multiple R-squared:  0.7343,    Adjusted R-squared:  0.7278
F-statistic: 113.5 on 12 and 493 DF,  p-value: < 2.2e-16
```

- La función `summary(lm.fit)$r.sq`nos permite obtener el R^2, y `summary(lm.fit)$sigma` nos da el RSE.
- La función `vif()` se utiliza para calcular los factores de inflación de la varianza.

```
library(car)
```

```
Loading required package: carData
```

```
vif(lm.fit)
```

```
    crim       zn    indus     chas      nox       rm      age      dis
1.767486 2.298459 3.987181 1.071168 4.369093 1.912532 3.088232 3.954037
     rad      tax  ptratio    lstat
7.445301 9.002158 1.797060 2.870777
```

Si queremos realizar una regresión excluyendo un predictor, en este caso vamos a excluir a la edad:

```r
lm.fit1 <- lm(medv ~ . - age, data = Boston)
summary(lm.fit1)
```

```
Call:
lm(formula = medv ~ . - age, data = Boston)

Residuals:
     Min       1Q   Median       3Q      Max
-15.1851  -2.7330  -0.6116   1.8555  26.3838

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  41.525128   4.919684   8.441 3.52e-16 ***
crim         -0.121426   0.032969  -3.683 0.000256 ***
zn            0.046512   0.013766   3.379 0.000785 ***
indus         0.013451   0.062086   0.217 0.828577
chas          2.852773   0.867912   3.287 0.001085 **
nox         -18.485070   3.713714  -4.978 8.91e-07 ***
rm            3.681070   0.411230   8.951  < 2e-16 ***
dis          -1.506777   0.192570  -7.825 3.12e-14 ***
rad           0.287940   0.066627   4.322 1.87e-05 ***
tax          -0.012653   0.003796  -3.333 0.000923 ***
ptratio      -0.934649   0.131653  -7.099 4.39e-12 ***
lstat        -0.547409   0.047669 -11.483  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.794 on 494 degrees of freedom
Multiple R-squared:  0.7343,    Adjusted R-squared:  0.7284
F-statistic: 124.1 on 11 and 494 DF,  p-value: < 2.2e-16
```

- Otra opción para excluir a un predictor es `update()`.

```r
lm.fit1 <- update(lm.fit, ~ . - age)
```

Términos de Interacción

- La función `lm()`permite incluir términos de interacción en un modelo lineal.

- La sintaxis `lstat:black` indica a R que incluya un término de interacción entre lstat y black.
- La sintaxis `lstat * edad` incluye lstat, la edad.
- El término de interacción `lstat * edad` se refiere que son predictores.

```
summary(lm(medv ~ lstat * age, data = Boston))
```

```
Call:
lm(formula = medv ~ lstat * age, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-15.806  -4.045  -1.333   2.085  27.552

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.0885359  1.4698355  24.553  < 2e-16 ***
lstat       -1.3921168  0.1674555  -8.313 8.78e-16 ***
age         -0.0007209  0.0198792  -0.036   0.9711
lstat:age    0.0041560  0.0018518   2.244   0.0252 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.149 on 502 degrees of freedom
Multiple R-squared:  0.5557,    Adjusted R-squared:  0.5531
F-statistic: 209.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

Transformaciones No Lineales de los Predictores

- La función `lm()` puede acomodar transformaciones no lineales de los predictores.

Tenemos un predictor X y se puede crear un predictor X^2 utilizando `I(X^2)`.

```
lm.fit2 <- lm(medv ~ lstat + I(lstat^2))
summary(lm.fit2)
```

```
Call:
lm(formula = medv ~ lstat + I(lstat^2))

Residuals:
     Min       1Q   Median       3Q      Max
-15.2834  -3.8313  -0.5295   2.3095  25.4148

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 42.862007   0.872084   49.15   <2e-16 ***
lstat       -2.332821   0.123803  -18.84   <2e-16 ***
I(lstat^2)   0.043547   0.003745   11.63   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.524 on 503 degrees of freedom
Multiple R-squared:  0.6407,    Adjusted R-squared:  0.6393
F-statistic: 448.5 on 2 and 503 DF,  p-value: < 2.2e-16
```

El valor p casi nulo asociado al término cuadrático conduce a un modelo mejorado.

- La función**anova()** sirve para profundizar el análisis de en qué medida el ajuste cuadrático es superior al lineal.

```
lm.fit <- lm(medv ~ lstat)
anova(lm.fit, lm.fit2)
```

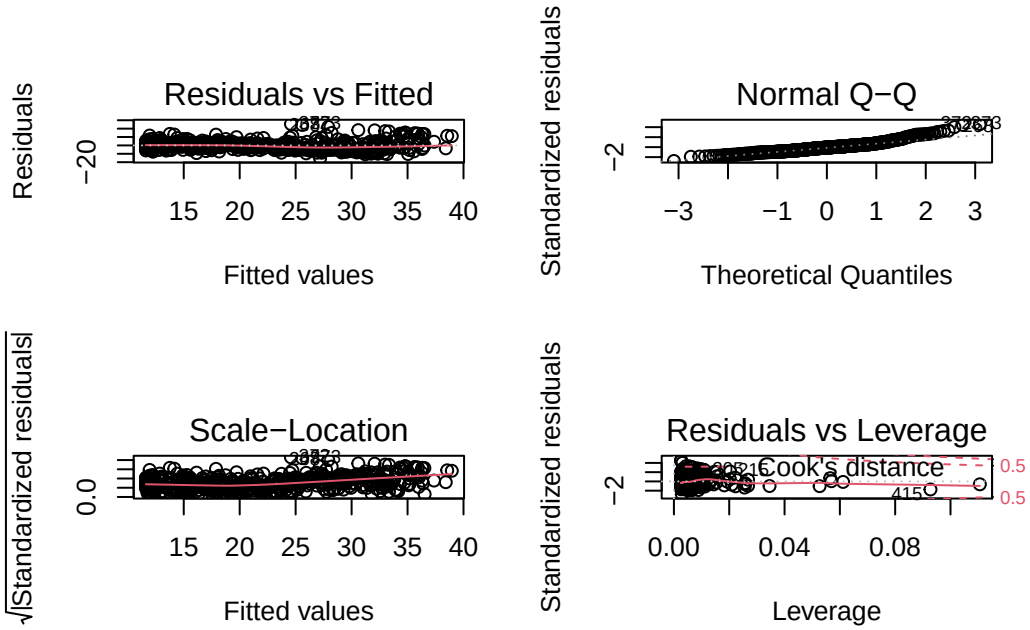
```
Analysis of Variance Table

Model 1: medv ~ lstat
Model 2: medv ~ lstat + I(lstat^2)
  Res.Df   RSS Df Sum of Sq     F    Pr(>F)
1    504 19472
2    503 15347  1    4125.1 135.2 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En estas líneas se puede observar que el Modelo 1 representa el submodelo lineal que contiene un solo predictor lstat, mientras que el Modelo 2 corresponde al modelo cuadrático más amplio que tiene dos predictores, lstat y lstat2.

- La función `anova()` realiza una prueba de hipótesis, compara los dos modelos.
- La hipótesis nula es que los dos modelos se ajustan igual de bien a los datos.
- La hipótesis alternativa es que el modelo completo es superior.

Se puede observar que el estadístico F es 135 y el valor p asociado es prácticamente cero. Esto demuestra que el modelo que contiene los predictores `lstat` y `lstat2` es muy superior al modelo que sólo contiene el predictor `lstat`.

```
par(mfrow = c(2,2))
plot(lm.fit2)
```



- La función `poly()` se utiliza para crear un polinomio dentro de `lm()`, se va a producir un ajuste polinómico de quinto orden.

```
lm.fit5 <- lm(medv ~ poly(lstat, 5))
summary(lm.fit5)
```

```
Call:
lm(formula = medv ~ poly(lstat, 5))
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-13.5433  -3.1039  -0.7052   2.0844  27.1153

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       22.5328     0.2318  97.197  < 2e-16 ***
poly(lstat, 5)1 -152.4595     5.2148 -29.236  < 2e-16 ***
poly(lstat, 5)2   64.2272     5.2148  12.316  < 2e-16 ***
poly(lstat, 5)3  -27.0511     5.2148  -5.187 3.10e-07 ***
poly(lstat, 5)4   25.4517     5.2148   4.881 1.42e-06 ***
poly(lstat, 5)5  -19.2524     5.2148  -3.692 0.000247 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.215 on 500 degrees of freedom
Multiple R-squared:  0.6817,    Adjusted R-squared:  0.6785
F-statistic: 214.2 on 5 and 500 DF,  p-value: < 2.2e-16
```

Esto sugiere que la inclusión de términos polinómicos adicionales, de hasta quinto orden, mejora el ajuste del modelo. Sin embargo los datos revela que ningún término polinómico más allá del quinto orden tiene valores p significativos en un ajuste de regresión.

- Un modelo lineal aplicado a la salida de la función `poly()` tendrá los mismos valores ajustados, que un modelo lineal aplicado a los polinomios brutos.

```r
summary(lm(medv ~ log(rm), data = Boston))
```

```
Call:
lm(formula = medv ~ log(rm), data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-19.487  -2.875  -0.104   2.837  39.816

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -76.488      5.028  -15.21   <2e-16 ***
log(rm)       54.055      2.739   19.73   <2e-16 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.915 on 504 degrees of freedom
Multiple R-squared:  0.4358,    Adjusted R-squared:  0.4347
F-statistic: 389.3 on 1 and 504 DF,  p-value: < 2.2e-16
```

Predictores Cualitativos

Se examinarán los datos de Carseats que forman parte de la biblioteca ISLR2. Intentaremos predecir las Ventas en 400 localidades basándonos en una serie de predictores.

```r
head(Carseats)
```

```
  Sales CompPrice Income Advertising Population Price ShelveLoc Age Education
1  9.50       138     73          11        276   120       Bad  42        17
2 11.22       111     48          16        260    83      Good  65        10
3 10.06       113     35          10        269    80    Medium  59        12
4  7.40       117    100           4        466    97    Medium  55        14
5  4.15       141     64           3        340   128       Bad  38        13
6 10.81       124    113          13        501    72       Bad  78        16
  Urban  US
1   Yes Yes
2   Yes Yes
3   Yes Yes
4   Yes Yes
5   Yes  No
6    No Yes
```

Los datos de Carseats incluyen predictores cualitativos como el `Shelveloc`, un indicador de la calidad de la estantería, etc.

El indicador `Shelveloc` adopta tres valores posibles: Malo, Medio y Bueno, a partir de esto R genera variables ficticias automáticamente.

```r
lm.fit <- lm(Sales ~ . + Income:Advertising + Price:Age, data = Carseats)
summary(lm.fit)
```

```
Call:
lm(formula = Sales ~ . + Income:Advertising + Price:Age, data = Carseats)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.9208 -0.7503  0.0177  0.6754  3.3413

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         6.5755654  1.0087470   6.519 2.22e-10 ***
CompPrice           0.0929371  0.0041183  22.567  < 2e-16 ***
Income              0.0108940  0.0026044   4.183 3.57e-05 ***
Advertising         0.0702462  0.0226091   3.107 0.002030 **
Population          0.0001592  0.0003679   0.433 0.665330
Price              -0.1008064  0.0074399 -13.549  < 2e-16 ***
ShelveLocGood       4.8486762  0.1528378  31.724  < 2e-16 ***
ShelveLocMedium     1.9532620  0.1257682  15.531  < 2e-16 ***
Age                -0.0579466  0.0159506  -3.633 0.000318 ***
Education          -0.0208525  0.0196131  -1.063 0.288361
UrbanYes            0.1401597  0.1124019   1.247 0.213171
USYes              -0.1575571  0.1489234  -1.058 0.290729
Income:Advertising  0.0007510  0.0002784   2.698 0.007290 **
Price:Age           0.0001068  0.0001333   0.801 0.423812
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.011 on 386 degrees of freedom
Multiple R-squared:  0.8761,    Adjusted R-squared:  0.8719
F-statistic:   210 on 13 and 386 DF,  p-value: < 2.2e-16
```

- La función `contrasts()`devuelve la codificación que R utiliza para la variable ficticia.

```
attach(Carseats)
contrasts(ShelveLoc)
```

```
       Good Medium
Bad       0      0
Good      1      0
Medium    0      1
```

- La variable ficticia `ShelveLocGood` toma un valor de 1 si la ubicación de la estantería es buena, y 0 en caso contrario.
- La variable ficticia `ShelveLocMedium` es igual a 1 si la ubicación de la estantería es media, y 0 en caso contrario.

- Si el `coe cient` para `ShelveLocGood` de la regresión es positivo indica que una buena ubicación de las estanterías, mientras que si `ShelveLocMedium` tiene un `coe cient` positivo menor, indica una ubicación media de la estantería la cual se asocia con mayores ventas.

Funciones de Escritura

R viene con muchas funciones útiles, y disponibles a través de las bibliotecas. Sin embargo, a menudo vamos a realizar operaciones para las que no hay ninguna función disponible.

Una función que lee las bibliotecas `ISRL2` y `MASS` es `LoadLibraries()`, sin embargo sale error ya que no está creada una función.

```
> LoadLibraries
Error: object 'LoadLibraries' not found
> LoadLibraries()
Error: could not find function "LoadLibraries"
```

- Los símbolos `+` son impresos por R y no deben escribirse.
- El símbolo `{` informa a R que varios comandos están por ser ingresados, finalmente el símbolo `}` informa que no se introducirán más comandos.

```
LoadLibraries <- function() {
  library(ISLR2)
  library(MASS)
  print("Las bibliotecas han sido cargadas.")
  }
```

- La función `LoadLibraries` permite a R que nos informe que hay en la función.

```
LoadLibraries
```

```
function() {
  library(ISLR2)
  library(MASS)
  print("Las bibliotecas han sido cargadas.")
  }
```

```
  function(){
    library(ISLR2)
    library(MASS)
    print("Las bibliotecas han sido cargadas.")
  }
```

```
function(){
  library(ISLR2)
  library(MASS)
  print("Las bibliotecas han sido cargadas.")
}
```

Si llamamos a la funciones, las librerias se cargan.

```
  LoadLibraries()
```

```
[1] "Las bibliotecas han sido cargadas."
```