Introducción al Aprendizaje Estadístico con Aplicaciones en R

Daniela Cuesta - Paola Peralta Flores

CAPITULO 2

Aprendizaje estadístico

2.1 ¿Qué es el Aprendizaje Estadístico?

Para entender mejor vamos a empezar con que somos consultores estadísticos para la asociación entre la publicidad y las ventas de un producto, esta publicidad consiste en las ventas en 200 mercados, junto con los presupuestos publicitarios en la TV, radio y prensa.

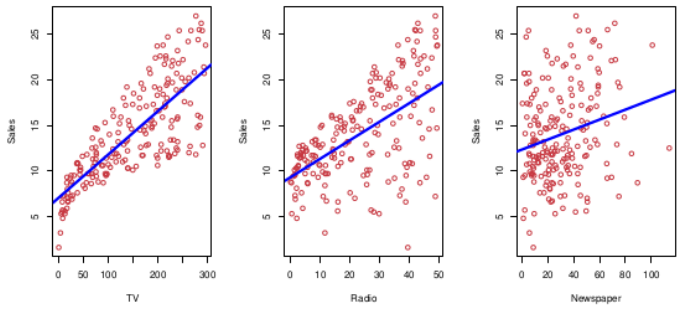


Figure 1: El gráfico muestra las ventas en función de los presupuestos de televisión, radio y prensa, para 200 mercados diferentes. En otras palabras, cada línea azul representa un modelo sencillo que puede utilizarse para predecir las ventas en cada medio

Se puede ver en la imagen que se puede controlar el gasto publicitarios en cada medio, se puede ordenar al cliente que ajuste los presupuestos publicitarios con las ventas, es decir se va a desarrollar un modelo que permita predecir las ventas a partir de los presupuestos de los tres medios.

- Los presupuestos publicitarios son variables de entrada.
- Las variables de entrada reciben distintos nombres, como predictores, variables independientes o características, o a veces simplemente variables. (X con subíndice)
- Las ventas son variables de salida
- La variable de salida es denominada variable de respuesta o dependiente (Y)

En términos más generales, si hay una respuesta cuantitativa Y y p diferentes predictores, X1 , X2 , . . . Xp, va a existir una relación entre Y y X que se puede escribir de forma muy general

Y = f (X) + error.

- f representa la información sistemática que X proporciona sobre Y.
- error es un término de error aleatorio que es independiente de X y tiene media cero.

En la figura 2 se observa los ingresos frente a los años de educación de 30 personas. El gráfico sugiere que se podría predecir la renta utilizando los años de educación.

- La función f que conecta la variable de entrada con la variable de salida es, en general, desconocida.
- f es es conocida y se muestra mediante la curva azul.

Las líneas verticales representan los términos de error .

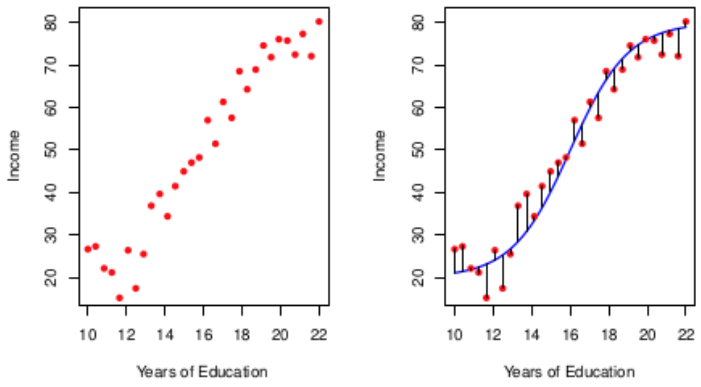


Figure 2: Izquierda: Los puntos rojos son los valores observados de ingresos (en decenas de miles de dólares) y los años de educación de 30 individuos. Derecha: La curva azul representa la verdadera relación subyacente entre ingresos y los años de educación. Las líneas negras representan el error asociado a cada observación.

2.1.1 ¿Por qué estimar f?

Para la predicción y la inferencia

Predicción

Tenemos un conjunto de entradas X, pero la salida Y no puede obtenerse fácilmente. En este caso, podemos predecir Y utilizando $\hat{Y} = \hat{f}(X)$

- $\hat{f}$ representa estimación de f .
- $\hat{Y}$ representa la predicción para Y .

A modo de ejemplo, supongamos que X1 , . . . Xp son características de la muestra de sangre de un paciente que pueden medirse fácilmente en un laboratorio, e Y es una variable que codifica el riesgo del paciente de sufrir una reacción adversa grave a un determinado medicamento.

La precisión de $\hat{Y}$ como predicción de Y depende de error reducible y error irreducible, ya que $\hat{f}$no será una estimación perfecta de f , y esta inexactitud introducirá algún error. Este error es reducible porque podemos mejorar potencialmente la precisión de $\hat{f}$ utilizando la técnica de aprendizaje estadístico más adecuada para estimar f .

Inferencia

A menudo estamos interesados en comprender la asociación entre Y y X1 . . . , Xp , para eso se tienen que responder las siguientes preguntas:

- **¿Qué predictores están asociados con la respuesta?**

Amenudo que sólo una pequeña fracción de los predictores disponibles están sustancialmente con Y .

- **¿Cuál es la relación entre la respuesta y cada predictor?**

Algunos predictores pueden tener una relación positiva con Y , en el sentido de que los valores mayores del predictor se asocian con valores mayores de Y .

- **¿Puede resumirse adecuadamente la relación entre Y y cada predictor mediante una ecuación lineal?**

La mayoría de los métodos para estimar f han adoptado una forma lineal. Pero a menudo un modelo lineal puede no ofrecer una representación exacta de la relación entre las variables de entrada y de salida.

Por ejemplo, en una empresa interesada en llevar a cabo una campaña de marketing directo. El objetivo es identificar a las personas responder positivamente a un mailing.

En este caso, las variables demográficas sirven como predictores, y la respuesta a la campaña de marketing (positiva o negativa) es el resultado, para así predecir con exactitud la respuesta utilizando los predictores.

En este modelo se puede responder a preguntas como:

- ¿Qué medios se asocian a las ventas?
- ¿Qué medios generan el mayor aumento de ventas?
- ¿Cuál es la magnitud del aumento de ventas asociado a un incremento de la publicidad televisiva?

Por último, existen algunos modelos que pueden utilizarse tanto para la predicción como para la inferencia, dependiendo de si nuestro objetivo final es la predicción, la inferencia o una combinación de ambos, pueden ser adecuados distintos métodos para estimar f.

- Los modelos lineales permiten una inferencia relativamente sencilla pero pueden no producir predicciones.
- Algunos de los enfoques no lineales pueden proporcionar predicciones bastante precisas de Y.

2.1.2 ¿Cómo estimamos f?

Siempre supondremos que hemos observado un conjunto de n puntos de datos diferentes, los cuales se denominan datos de entrenamiento porque utilizaremos estos para entrenar, o enseñar, a nuestro método a estimar f .

Nuestro objetivo es aplicar un método de aprendizaje estadístico a los datos de entrenamiento para estimar la función desconocida f . En términos generales, la mayoría de los métodos de aprendizaje estadístico para esta tarea se pueden caracterizar como paramétricos o paramétricos.

Métodos Paramétricos

Los métodos paramétricos implican un enfoque basado en un modelo en dos pasos.

1. Hacemos una suposición sobre la forma funcional, o forma de f .

X:

f(X) = 0 + 1 X1 + 2 X2 + - - - + p Xp

2. Una vez seleccionado un modelo, necesitamos un procedimiento que utilice los datos de entrenamiento para ajustar o entrenar el modelo.

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 totalmente arbitraria.

La desventaja potencial de un enfoque paramétrico es que si el modelo elegido se aleja demasiado de la verdadera f , nuestra estimación será deficiente, esto se puede resolver este problema eligiendo modelos flexibles que puedan ajustarse a muchas variables diferentes para f .

Métodos no paramétricos

Buscan una estimación de f que se acerque lo más posible a los puntos de datos sin ser demasiado aproximada o imprecisos.

En los enfoques no paramétricos no se hace ninguna suposición sobre la forma de f. Sin embargo, los enfoques no paramétricos tienen una gran desventaja, dado que no reducen el problema de estimar f a un número de parámetros, se necesita un gran número de observaciones para obtener una estimación precisa de f.

2.1.3 El Equilibrio entre la Precisión de la Predicción y la Interpretabilidad del Modelo

De los muchos métodos que existen, algunos son menos flexibles o más restrictivos. Por ejemplo, la regresión lineal es relativamente inflexible, porque sólo puede generar funciones lineales, y hay otros métodos, como los splines de placa delgada, son más flexibles porque pueden generar una gama mucho más amplia de formas posibles para estimar f .

Podemos usar un modelo restrictivo si nos interesa la inferencia, ya que son mucho mas interpretables, debido a que será bastante fácil entender la relación entre Y y X1 , X2 , . . . Xp . Por el contrario, los enfoques flexibles, como los splines pueden dar lugar a estimaciones tan complicadas de f que resulta difícil comprender cómo se asocia cada predictor individual con la respuesta.

- La regresión lineal por mínimos cuadrados es relativamente inflexible pero es bastante interpretable.

- El lazo, se basa en el método lineal, pero utiliza un procedimiento de ajuste alternativo para estimar los coeficientes 0 , 1 , . . . , p, es por eso que el lazo es un enfoque menos flexible que la regresión lineal, y más interpretable que la regresión lineal.
- Los modelos aditivos generalizados son más flexibles que la regresión lineal. También son menos interpretables que la regresión lineal.
- Los métodos totalmente no lineales como bagging, boosting, máquinas de vectores soporte con núcleos no lineales, y las redes neuronales (aprendizaje profundo), son enfoques muy flexibles que resultan más difíciles de interpretar.

Hemos establecido que cuando el objetivo es la inferencia, hay claras ventajas en utilizar métodos de aprendizaje estadístico sencillos y relativamente inflexibles.

2.1.4 Aprendizaje Supervisado frente a Aprendizaje No Supervisado

La mayoría de los problemas de aprendizaje estadístico se dividen en dos categorías: supervisado o no supervisados.

Aprendizaje supervisado : Para cada observación de medida(s) predictora(s) xi , i = 1, . . . n, existe una medida de respuesta asociada respuesta yi . Deseamos ajustar un modelo que relacione la respuesta con los predictores, con el fin de comprender mejor la relación entre la respuesta y los predictores (inferencia).

- Muchos métodos clásicos de aprendizaje estadístico como la regresión lineal y la regresión logística de aprendizaje supervisado.

Aprendizaje no supervisado: Para cada observación i = 1, . . . n, observamos un vector de medidas xi pero ninguna respuesta asociada a yi, la situación se denomina no supervisada porque carecemos de una variable de respuesta que pueda supervisar nuestro análisis.

Una herramienta de aprendizaje estadístico que podemos utilizar en este contexto es el análisis de conglomerados o clustering., cuyo objetivo es determinar, a partir de x1 , . . . xn , si las observaciones pertenecen a grupos relativamente distintos.

2.1.5 Problemas de Regresión frente a Problemas de Clasificación

Las variables pueden caracterizarse como cuantitativas o cualitativas (también denominadas conocidas como categóricas).

- Las variables cuantitativas toman valores numéricos.
- Las variables cualitativas toman valores de diferentes clases, o categorías.

La regresión lineal por mínimos cuadrados, se utiliza con una respuesta cuantitativa, mientras que la regresión logística, suele utilizarse con un29a respuesta cualitativa(binaria).

Algunos métodos estadísticos, como K-nearest neighbors y boosting, pueden utilizarse en el caso de respuestas cuantitativas o cualitativas, también se puede utilizar la regresión lineal cuando la respuesta es cuantitativa y la regresión logística cuando la respuesta es cualitativa.

2.2 Evaluación de la Precisión de los Modelos

En estadística, ningún método domina a todos los demás en todos los conjuntos de datos posibles, en un conjunto de datos concreto, un método específico puede funcionar pero otro método puede funcionar mejor en un conjunto de datos similar pero diferente.

2.2.1 Medición de la Calidad del Ajuste

Para evaluar el rendimiento de un método de aprendizaje estadístico en un conjunto de datos determinado, necesitamos algún modo de medir hasta qué punto sus predicciones se ajustan realmente a los datos observados.

En el ámbito de la regresión, la medida más utilizada es el error cuadrático medio (ECM), dado por:

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

El MSE será pequeño si las respuestas predichas están muy cerca de las respuestas verdaderas, y será grande si las respuestas predichas y verdaderas difieren sustancialmente.

Con el MSE lo que nos interesa es la precisión de las predicciones que obtenemos cuando aplicamos nuestro método a datos de prueba que no hemos visto previamente.

El MSE de prueba disminuye inicialmente a medida que aumenta el nivel de flexibilidad, y casi siempre esperamos que el MSE de entrenamiento sea menor que el MSE de prueba porque la mayoría de los métodos de aprendizaje estadístico, directa o indirectamente,buscan minimizar el MSE de entrenamiento.

Un método importante es la validación cruzada, que es un método para estimar el MSE de prueba utilizando los datos de entrenamiento.

Ejemplo

- Predecir el precio de mañana o el precio del próximo mes.
- Predecir con exactitud el riesgo de diabetes de futuros pacientes basándose en sus mediciones clínicas.

2.2.2 El Equilibrio entre Sesgo y Varianza - Off

La forma de U en las curvas de MSE resulta ser el resultado de dos propiedades contrapuestas de los métodos de aprendizaje estadístico.

Se puede decir que para minimizar el error de prueba esperado, se necesita seleccionar un método de aprendizaje estadístico que consiga simultáneamente baja varianza y bajo sesgo. Se tiene que tener en cuenta que la varianza es intrínsecamente una cantidad no negativa, y el sesgo al cuadrado es una cantidad no negativa. Por lo tanto, vemos que el MSE de prueba esperado nunca puede estar por debajo de Var( ).

¿Qué entendemos por varianza y sesgo de un método de aprendizaje estadístico?

La varianza se refiere a la cantidad en la que f^ cambiaría, utilizando un conjunto de datos de entrenamiento diferente. Pero lo ideal es que la estimación de f no varíe demasiado entre conjuntos de entrenamiento. Sin embargo, si un método tiene una alta varianza entonces pequeños cambios en los datos de entrenamiento pueden dar lugar a grandes cambios en f^.

- Los métodos estadísticos más flexibles tienen mayor varianza.

El sesgo se refiere al error que se introduce al aproximar un problema de la vida real, que puede ser extremadamente complicado, a un valor mucho mayor que el de un problema real.

Como regla general, a medida que utilicemos métodos más flexibles, la varianza aumentará y el sesgo disminuirá.

Un buen rendimiento del conjunto de prueba de un método de aprendizaje estadístico requiere una varianza baja, así como un sesgo al cuadrado bajo. El reto consiste en encontrar un método para el que tanto la varianza como el sesgo al cuadrado sean bajos.

La relación entre el sesgo, la varianza y el MSE del conjunto de prueba se conoce como el **trade-off.**

2.2.3 Entorno de Clasificación

Muchos de los conceptos que hemos encontrado, como el equilibrio entre sesgo y varianza, se trasladan al entorno de clasificación con sólo algunas modificaciones debido a que yi ya no es cuantitativo, ahora es cualitativos

El enfoque más común para cuantificar la precisión de nuestra estimación f^ es la tasa de error de entrenamiento, es decir calcula la fracción de clasificaciones incorrectas y se calcula a partir de los datos utilizados para entrenar nuestro clasificador.

Clasificador Bayes

Es la probabilidad de que Y = j, dado el vector predictor observado x0 . Pr (Y = j|X = x0)

En un problema de dos clases en el que hay Bayes, sólo existen dos posibles valores de respuesta, la clase 1 o la clase 2, el clasificador de Bayes corresponde a la predicción de la clase uno si Pr(Y = 1|X = x0) > 0,5, y la clase dos en caso contrario.

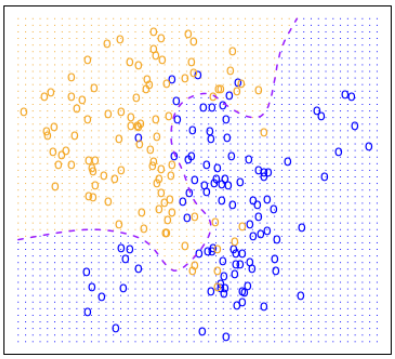


Figure 3: Clasificador Bayes

La figura 4 muestra un ejemplo en el que se utiliza un conjunto de datos simulados en un espacio bidimensional formado por los predictores X1 y X2.

Los círculos naranja y azules corresponden a observaciones de entrenamiento que pertenecen a dos clases diferentes. La región sombreada en naranja refleja el conjunto de puntos para los que Pr (Y = naranja|X) es superior al 50 %, mientras que la región sombreada en azul indica el conjunto de puntos cuya probabilidad es inferior al 50 %, la línea discontinua morada representa los puntos en los que la probabilidad es exactamente del 50%.

El clasificador de Bayes produce la tasa de error de prueba más baja posible, denominada tasa de error de Bayes, La tasa de error de Bayes es análoga al el error irreducible, **la tasa de error de Bayes es de 0,133.**

K– Nearest Neighbors

Dado un número entero positivo K y una observación de prueba x0 , el clasificador KNN identifica primero los K más cercanos a x0.

En la figura 5 se muestra un ejemplo ilustrativo del método KNN. En el panel de la izquierda, se observa un conjunto de datos de entrenamiento formado por seis observaciones azules y seis naranjas.

Nuestro objetivo es hacer una predicción para el punto marcado con la cruz negra. Supongamos que elegimos K = 3. Entonces KNN identificará primero las tres observaciones más cercanas a la cruz.

Por lo tanto, KNN predecirá que la cruz negra pertenece a la clase azul.

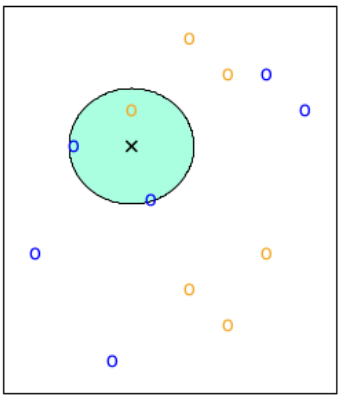


Figure 4: Conjunto de datos de entrenamiento

- A pesar de que se trata de un enfoque muy sencillo, KNN a menudo puede producir clasificadores sorprendentemente cercanos al clasificador óptimo de Bayes.
- La tasa de error utilizando KNN es de 0,1363, que se aproxima a la tasa de error de Bayes de 0,1304.
- La elección de K tiene un efecto drástico en el clasificador KNN obtenido.
- A medida que K aumenta, el método se vuelve menos flexible y produce una frontera de decisión cercana a la lineal. Esto corresponde a un clasificador de baja varianza pero alto sesgo.

Al igual que en el caso de la regresión, no existe una relación estrecha entre la tasa de error de entrenamiento y la tasa de error de prueba.En general, a medida que utilizamos métodos de clasificación más flexibles, el de error de prueba puede no serlo.

Tanto en el ámbito de la regresión como en el de la clasificación, elegir el nivel correcto de flexibilidad es fundamental para el éxito de cualquier método de aprendizaje estadístico.

CAPITULO 3

Regresión Lineal

Es una herramienta útil para predecir una respuesta cuantitativa. El enfoque de mínimos cuadrados se utiliza más comúnmente para ajustar este modelo.

3.1 Regresión Lineal Simple

Es un método para predecir una respuesta cuantitativa, Y a partir de una única variable de predicción, X. Supone que existe aproximadamente una relación lineal entre X e Y.

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

El signo " " se lee "se modela aproximadamente como". ß0 y ß1 son dos constantes desconocidas (coeficientes) que representan los términos de intersección y pendiente en el modelo lineal. Una vez que hayamos usado nuestros datos de entrenamiento para producir estimaciones:

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

donde ŷ indica una predicción de Y sobre la base de X = x. Se usa un símbolo de sombrero, para denotar el valor estimado de un parámetro desconocido o el valor predicho de la respuesta.

3.1.1 Estimación de los Coeficientes

En la práctica, ß0 y ß1 son desconocidos. Entonces, debemos usar datos para estimar los coeficientes.

$$(x_1, y_1),\ (x_2, y_2), \ldots,\ (x_n, y_n)$$

representan n pares de observación, cada uno de los cuales consta de una medida de X y una medida de Y. Nuestro objetivo es encontrar una intersección ˆ0 y una pendiente ˆ1 tal que la línea resultante esté lo más cerca posible de los n = datos. Esto con minimizar el criterio de mínimos cuadrados.

Esta es la diferencia entre el i-ésimo valor de respuesta observado y el i-ésimo valor de respuesta que es predicho por nuestro modelo lineal. Definimos la suma residual de cuadrados de dos formas:

Se define las estimaciones del coeficiente de mínimos cuadrados para la regresión lineal simple. Los minimizadores son:

3.1.2 Evaluación de la Precisión de las Estimaciones del Coeficiente

Asumimos que la verdadera relación entre X e Y toma la forma Y = f (X) + para alguna función desconocida f, donde es un término de error aleatorio de media cero. Si f se aproxima mediante una función lineal:

Aquí ß0 es el valor esperado de Y cuando X = 0 (término de intersección), y ß1 es la pendiente, el aumento promedio en Y asociado con un aumento de una unidad en X. Por lo general, asumimos que el término de error es independiente de X.

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

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

NO ENTENDI NADA DE AQUI ASI QUE ME SALTE AL SIGUIENTE PERO NO TERMINE

3.1.3 Evaluación de la Precisión del Modelo

Una vez rechazada la hipótesis nula a favor de la hipótesis alternativa, hay que cuantificar hasta qué punto el modelo se ajusta a los datos. La calidad de este ajuste se evalúa utilizando: el error estándar residual (RSE) y la estadística R2.

Error Estándar Residual

Debido a la presencia de términos de error, no se puede predecir perfectamente Y a partir de X. El RSE es una estimación de la desviación estándar de . Es decir, es la cantidad promedio que la respuesta se desviará de la verdadera línea de regresión. Se calcula usando la fórmula:

El RSE se considera una medida de la falta de ajuste del modelo de precisión de las estimaciones del coeficiente a los datos. Si las predicciones obtenidas con el modelo están muy cerca de los valores reales de los resultados, es decir, si ŷi yi para i = 1, . . . , n—entonces RSE será pequeño, es decir, el modelo se ajusta a los datos. Por otro lado, si ŷi está muy lejos de yi 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.

Estadística R^2

Proporciona una medida de ajuste alternativa. Toma la forma de una proporción, (varianza explicada), por lo que siempre toma un valor entre 0 y 1, y es independiente de la escala de Y. Fórmula:

- TSS es la suma total de los cuadrados, mide la varianza total en la respuesta Y, (cuadrados = cantidad de variabilidad inherente en la respuesta antes de que se realice la regresión).
- RSS mide la cantidad de variabilidad que queda sin explicar después de realizar la regresión.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

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

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

- TSS − RSS mide la cantidad de variabilidad en la respuesta que se explica (o elimina) al realizar la regresión
- R^2 mide la proporción de variabilidad en Y que se puede explicar usando X. Una estadística R^2 cercana a 1 indica que una gran proporción de la variabilidad en la respuesta se explica por la regresión. Un número cercano a 0 indica lo contrario; esto puede ocurrir porque el modelo lineal es incorrecto, o la varianza del error ^{2} es alta, o ambas cosas.

La estadística R^2 es una medida de la relación lineal entre X e Y, definida como:

3.2 Regresión Lineal Múltiple

Busca predecir una respuesta con más de un predictor. En lugar de ajustar un modelo de regresión lineal simple separado para cada predictor, es mejor extender el modelo de regresión lineal simple para que pueda acomodar varios predictores. Entonces el modelo de regresión lineal múltiple toma la forma:

donde Xj representa el j-ésimo predictor y ß`j`cuantifica la asociación entre esa variable y la respuesta. Interpretamos ß`j` como el efecto promedio sobre Y de un aumento de una unidad en Xj, manteniendo fijos todos los demás predictores.

3.2.1 Estimación de los Coeficientes de Regresión

Los coeficientes de regresión ß`0`, ß`1` . . . , ß`p` son desconocidos y deben estimarse. Se estian por:

Los parámetros se estiman utilizando el mismo enfoque de mínimos cuadrados, para minimizar la suma de los residuos al cuadrado:

Se puede usar cualquier paquete de software estadístico para calcular estas estimaciones de coeficientes.

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

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

$$\text{Cor}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

3.2.2 Algunas Preguntas Importantes

- ***¿Es útil al menos uno de los predictores en la predicción de la respuesta?***

¿Existe una relación entre la respuesta y los predictores?

Si 1 = 2 = · · · = p = 0, usamos una prueba de hipótesis para responder a esta pregunta. Probamos la hipótesis nula contra la alternativa:

Esta prueba de hipótesis se realiza calculando el estadístico F

Si los supuestos del modelo lineal son correctos, se puede demostrar que $E\{\text{RSS}/(n-p-1)\} = \sigma^2$ y que, siempre que H0 sea verdadera, $E\{(\text{TSS} - \text{RSS})/p\} = \sigma^2$.

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 Ha es verdadera, entonces $E\{(\text{TSS} - \text{RSS})/p\} > \sigma^2$, por lo que esperamos que F sea mayor que 1.

A veces queremos probar que un subconjunto particular de q de los coeficientes es cero. Esto corresponde a una hipótesis nula.

donde por conveniencia hemos puesto las variables elegidas para omisión al final de la lista. En este caso ajustamos un segundo modelo que utiliza todas las variables excepto la última q. Suponga que la suma residual de cuadrados para ese modelo es $\dot{\text{RSS}}_0$. Entonces el estadístico F apropiado es:

El enfoque de usar una estadística F para probar cualquier asociación entre los predictores y la respuesta funciona cuando p es relativamente pequeño (en comparación con n). Si p > n, entonces hay más coeficientes j para estimar que observaciones a partir de las cuales estimarlos. En este caso, ni siquiera podemos ajustar el modelo de regresión lineal múltiple usando mínimos cuadrados, por lo que no se puede usar el estadístico F.

- **¿Todos los predictores ayudan a explicar Y, o solo es útil un subconjunto de los predictores?**

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

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

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

Decidir sobre variables importantes

El primer paso es calcular el estadístico F y examinar el valor p asociado. Si concluimos sobre la base de ese valor p que al menos uno de los predictores está relacionado con la respuesta, entonces es natural preguntarse cuáles son los culpables. Si p es grande, es probable que hagan algunos descubrimientos falsos.

Es posible que todos los predictores estén asociados con la respuesta, pero es más frecuente que la respuesta solo esté asociada con un subconjunto de los predictores. La tarea de determinar qué predictores están asociados con la respuesta, para ajustar un solo modelo que involucre solo esos predictores, se conoce como *selección de variables.*

Idealmente, nos gustaría realizar la selección de variables probando muchos modelos diferentes, cada uno de los cuales contiene un subconjunto diferente de predictores. Determinamos qué modelo es el mejor mediante el Cp de Mallow, el criterio de información de Akaike (AIC), el criterio de información bayesiano (BIC) y el R^2 ajustado.

- *Selección de avance*: Modelo nulo, que contiene una intersección pero no predictores. Luego ajustamos p regresiones lineales simples y agregamos al modelo nulo la variable que resulta en el RSS más bajo. Luego agregamos a ese modelo la variable que resulta en el RSS más bajo para el nuevo modelo de dos variables. Este enfoque continúa hasta que se cumple alguna regla de parada.
- *Selección hacia atrás*: Comenzamos con todas las variables del modelo y eliminamos la variable con el valor p más grande. Se ajusta el nuevo modelo de variable $(p - 1)$. Este procedimiento continúa hasta que se alcanza una regla de parada.
- *Selección mixta*: Combinación de selección hacia adelante y hacia atrás. Comenzamos sin variables en el modelo y, agregamos la variable que proporciona el mejor ajuste. Continuamos agregando variables una por una. Si en algún punto el valor p de una de las variables supera cierto umbral, lo eliminamos. Continuamos realizando estos pasos hacia adelante y hacia atrás hasta que todas las variables en el modelo tengan un valor p lo suficientemente bajo.

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_a: \text{ at least one } \beta_j \text{ is non-zero.}$$

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

La selección hacia atrás no se puede usar si p > n, mientras que la selección hacia adelante siempre se puede usar. La selección mixta puede remediar esto.

- **¿Qué tan bien se ajusta el modelo a los datos?**

Ajuste del modelo

Dos de las medidas numéricas más comunes del ajuste del modelo son RSE y R^2, la fracción de varianza explicada.

En la regresión lineal múltiple R^2 es igual a Cor(Y, $\hat{Y}$)2 , el cuadrado de la correlación entre la respuesta y el modelo lineal ajustado. Un valor de R^2 cercano a 1 indica que el modelo explica una gran parte de la varianza en la variable de respuesta. En general, RSE se define como:

Los modelos con más variables pueden tener un RSE más alto si la disminución de RSS es pequeña en relación con el aumento de p.

- **Dado un conjunto de valores predictores, ¿qué valor de respuesta deberíamos predecir y qué tan precisa es nuestra predicción?**

Predicciones

Una vez que hemos ajustado el modelo de regresión múltiple, es sencillo aplicar al estimacion de coeficientes para predecir la respuesta Y sobre la base de un conjunto de valores para los predictores X1, X2, . . . , XP . Hay tres tipos de incertidumbre:

- El coeficiente ^0, ^1, . . . , ^p son estimaciones para 0 , 1 , . . , p. Es decir, el plano de mínimos cuadrados $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$ es solo una estimación para el verdadero plano de regresión de la población $f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$
- Suponer un modelo lineal para f (X) es una aproximación a la realidad, por lo que existe una fuente adicional de error potencialmente reducible que llamamos *sesgo del modelo.*
- Incluso si conociéramos f (X), valores verdaderos de 0, 1, . . . , p, el valor de respuesta no se puede predecir perfectamente debido al error aleatorio . Los intervalos de predicción son más amplios que los de confianza, porque incorporan tanto el error reducible como el error irreducible.

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

$$F = \frac{(\text{RSS}_0 - \text{RSS})/q}{\text{RSS}/(n-p-1)},$$

$$\text{RSE} = \sqrt{\frac{1}{n-p-1}\text{RSS}}.$$

3.3 Otras Consideraciones en el Modelo de Regresión

3.3.1 Predictores Cualitativos

Predictores con solo Dos Niveles

Si un predictor cualitativo (factor) solo tiene dos valores posibles, para incorporarlo en un modelo de regresión creamos una variable ficticia que toma dos posibles valores numéricos. Por ejemplo:

$$x_i = \begin{cases} 1 & \text{if } i\text{th person owns a house} \\ 0 & \text{if } i\text{th person does not own a house,} \end{cases}$$

como predictor en la ecuación de regresión:

En lugar de un esquema de codificación 0/1, podríamos crear una variable ficticia:

0 puede interpretarse como el saldo promedio general de la tarjeta de crédito (ignorando efecto de propiedad de la casa), y 1 es la cantidad por la cual los propietarios y no propietarios tienen saldos de tarjetas de crédito que están por encima y por debajo del promedio, respectivamente.

Predictores Cualitativos con Más de Dos Niveles

En este caso, una sola variable ficticia no puede representar todos los valores posibles. Entonces, se debe crear variables ficticias adicionales. Por ejemplo:

Se obtiene el modelo:

Ahora, 0 puede interpretarse como el saldo promedio de tarjetas de crédito para personas del Este, 1 puede interpretarse como la diferencia en el saldo promedio entre personas del Sur y del Este, y 2 puede interpretarse como la diferencia en el saldo promedio entre Oeste contra Este. Siempre habrá una variable ficticia menos que el número de niveles. El nivel sin variable ficticia se conoce como *línea de base*.

Los coeficientes y sus valores p dependen de la elección de la codificación de la variable ficticia. Podemos usar una prueba F $H_0 : \beta_1 = \beta_2 = 0$.

El uso de este enfoque de variable ficticia no presenta dificultades al incorporar predictores tanto cuantitativos como cualitativos.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person owns a house} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person does not.} \end{cases}$$

$$x_i = \begin{cases} 1 & \text{if } i\text{th person owns a house} \\ -1 & \text{if } i\text{th person does not own a house} \end{cases}$$

3.3.2 Extensiones del Modelo Lineal

El modelo de regresión lineal estándar proporciona resultados interpretables. Sin embargo, hace varias suposiciones altamente restrictivas. Establece que la relación entre los predictores y la respuesta es aditiva y lineal.

- Aditividad significa que la asociación entre un predictor Xj y la respuesta Y no depende de los valores de los otros predictores.
- Linealidad establece que el cambio en la respuesta Y asociado con un cambio de una unidad en Xj es constante, independientemente del valor de Xj.

Eliminación de la Suposición Aditiva

Considere el modelo de regresión lineal estándar con dos variables:

De acuerdo con este modelo, un aumento de una unidad en X1 está asociado con un aumento promedio en Y de 1 unidades, independientemente del valor de X2. Una forma de extender este modelo es incluir un tercer predictor, llamado *término de interacción*. Esto da como resultado el modelo:

Se puede reescribir como:

donde $\tilde{\beta}_1 = \beta_1 + \beta_3 X_2$. Dado que ˜1 ahora es una función de X2, la asociación entre X1 e Y ya no es constante: un cambio en el valor de X2 cambiará la asociación entre X1 e Y. Un argumento similar muestra que un cambio en el valor de X1 cambia la asociación entre X2 e Y.

Relaciones No Lineales

En algunos casos, la verdadera relación entre la respuesta y los predictores puede no ser lineal. Utilizando regresión polinomial, se puede extender el modelo lineal para acomodar las relaciones no lineales.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person owns a house} \\ \beta_0 - \beta_1 + \epsilon_i & \text{if } i\text{th person does not own a house.} \end{cases}$$

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is from the South} \\ 0 & \text{if } i\text{th person is not from the South} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is from the West} \\ 0 & \text{if } i\text{th person is not from the West.} \end{cases}$$

3.3.3 Problemas Potenciales

Los más comunes al ajustar un modelo son los siguientes:

- **No Linealidad de las Relaciones Respuesta-Predictor.**

Los diagramas de residuos son una herramienta gráfica útil para identificar la no linealidad.

Dado un modelo de regresión lineal simple, podemos trazar los residuos, ei = yi − ŷi , contra el predictor xi . En el caso de un modelo de regresión múltiple, dado que hay varios predictores, representamos los residuos frente a los valores ajustados ŷi. Idealmente, la gráfica residual no mostrará un patrón perceptible.

Si la gráfica residual indica que hay asociaciones no lineales en los datos, entonces un enfoque simple es usar transformaciones no lineales de los predictores, como log X, $\sqrt{X}$, and X^2, en el modelo de regresión.

- **Correlación de Términos de Error.**

Una suposición importante del modelo de regresión lineal es que los términos de error no están correlacionados. Si existe esta correlación, entonces los errores estándar estimados tenderán a subestimar los errores estándar verdaderos. Los valores de p asociados con el modelo serán más bajos de lo que deberían ser; esto podría llevarnos a concluir erróneamente que un parámetro es estadísticamente significativo.

Tales correlaciones ocurren con frecuencia en el contexto de datos de series de tiempo, que consisten en observaciones para las cuales se obtienen mediciones en puntos discretos en el tiempo.

En muchos casos, las observaciones que se obtienen en puntos de tiempo adyacentes tendrán errores positivamente correlacionados. Si los errores no están correlacionados, entonces no debería haber un patrón perceptible. Por otro lado, si los términos de error están positivamente correlacionados, entonces podemos ver un seguimiento en los residuos, es decir, los residuos adyacentes pueden tener valores similares.

- **Varianza No Constante de los Términos de Error.**

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is from the South} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is from the West} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is from the East.} \end{cases}$$

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

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

Otra supisición es que los términos de error tienen una varianza constante, $\mathrm{Var}(\epsilon_i) = \sigma^2$. Los errores estándar, intervalos de confianza y las pruebas de hipótesis asociadas con el modelo lineal se basan en esta suposición.

Desafortunadamente, las varianzas de los términos de error no siempre son constantes, pueden aumentar con el valor de la respuesta. Se pueden identificar (heteroscedasticidad), a partir de la presencia de una forma de embudo en el gráfico de residuos. Ante este problema, una posible solución es transformar la respuesta Y utilizando una función cóncava como *log* Y o $\sqrt{Y}$. Tal transformación da como resultado una mayor cantidad de reducción de las respuestas más grandes (reducción de la heterocedasticidad).

- **Valores Atípicos.**

Un valor atípico es un punto para el cual yi está lejos del valor predicho por el modelo. Los valores atípicos pueden surgir por una variedad de razones, como el registro incorrecto de una observación durante la recopilación de datos.

Es típico que un valor atípico que no tiene un valor de predictor inusual tenga poco efecto en el ajuste de mínimos cuadrados. Sin embargo, incluso si un valor atípico no tiene mucho efecto sobre el ajuste de mínimos cuadrados, puede causar otros problemas.

Los gráficos residuales se pueden utilizar para identificar valores atípicos. Si creemos que se ha producido un valor atípico debido a un error en la recopilación o el registro de datos, entonces una solución es simplemente eliminar la observación. Sin embargo, se debe tener cuidado, ya que un valor atípico puede indicar una deficiencia con el modelo, como un predictor faltante.

- **Puntos de Alto Apalancamiento.**

Las observaciones con alto apalancamiento tienen un valor inusual para xi. 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 solo un par de observaciones, porque cualquier problema con estos puntos puede invalidar todo el ajuste. Por esta razón, es importante identificar observaciones de alto 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

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\ &= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon \end{aligned}$$

valores de cada predictor individual, pero que sea inusual en términos del conjunto completo de predictores. Este problema es más pronunciado en configuraciones de regresión múltiple con más de dos predictores.

Para cuantificar el apalancamiento de una observación, calculamos la estadística de apalancamiento. Para una regresión lineal simple:

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

La estadística de apalancamiento hi siempre está entre 1/n y 1, y el apalancamiento promedio para todas las observaciones es siempre igual a (p + 1)/n.

- **Colinealidad.**

Se refiere a la situación en la que dos o más variables predictoras están estrechamente relacionadas entre sí.

Dado que la colinealidad reduce la precisión de las estimaciones de los coeficientes de regresión, hace que aumente el error estándar de ˆj. Recuerde que la estadística t para cada predictor se calcula dividiendo ˆj por su error estándar.

La colinealidad da como resultado una disminución de la potencia de la prueba de hipótesis, la probabilidad de detectar correctamente un coeficiente distinto de cero.

Para evitar tal situación, es deseable identificar y abordar los posibles problemas de colinealidad al ajustar el modelo.

Una forma sencilla de detectar la colinealidad es observar la matriz de correlación de los predictores. Un elemento de esta matriz que es grande en valor absoluto indica un par de variables altamente correlacionadas y, por lo tanto, un problema de colinealidad en los datos. Desafortunadamente, es posible que exista colinealidad entre tres o más variables (multicolinealidad) incluso si ningún par de variables tiene una correlación particularmente alta. Evaluar la multicolinealidad, calculando el factor de inflación de la varianza (VIF). El VIF es la relación de la varianza de ˆj cuando se ajusta el modelo completo dividida por la varianza de ˆj si se ajusta solo. El valor más pequeño posible para VIF es 1, l(ausencia total de colinealidad), un valor VIF que excede 5 o 10 indica una cantidad problemática de colinealidad.

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

3.5 Comparación de Regresión Lineal con K-Nearest Neighbors

Por el contrario, los métodos no paramétricos no asumen explícitamente una forma paramétrica para f (X) y, por lo tanto, brindan un enfoque alternativo y más flexible para realizar la regresión. En el método de regresión KNN, dado un valor para K y un punto de predicción

x0, la regresión KNN primero identifica las observaciones de entrenamiento K más cercanas a x0, representadas por N0. Luego estima f (x0) usando el promedio de todos las respuestas de entrenamiento en N0.

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in N_0} y_i$$

En general, el valor óptimo de K dependerá del compromiso sesgo-varianza. 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 completamente de una sola observación. Por el contrario, los valores más grandes de K proporcionan un ajuste más suave y menos variable; la predicción en una región es un promedio de varios puntos, por lo que cambiar una observación tiene un efecto menor. Sin embargo, el suavizado puede causar un sesgo al enmascarar parte de la estructura en f (X).

El enfoque paramétrico superará al enfoque no paramétrico si la forma paramétrica que se ha seleccionado está cerca de la verdadera forma de f .

Los métodos paramétricos tenderán a superar a los enfoques no paramétricos cuando haya un pequeño número de observaciones por predictor. Incluso cuando la dimensión es pequeña, podríamos preferir la regresión lineal a KNN desde el punto de vista de la interpretabilidad. Si el MSE de prueba de KNN es solo un poco más bajo que el de la regresión lineal, podríamos estar dispuestos a renunciar a un poco de precisión de predicción en aras de un modelo simple que pueda describirse en términos de solo unos pocos coeficientes, y para el cual los valores de p están disponibles.

3.6 Laboratorio

3.6.1 Librerias

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

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

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

    Boston
```

3.6.2 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")
```

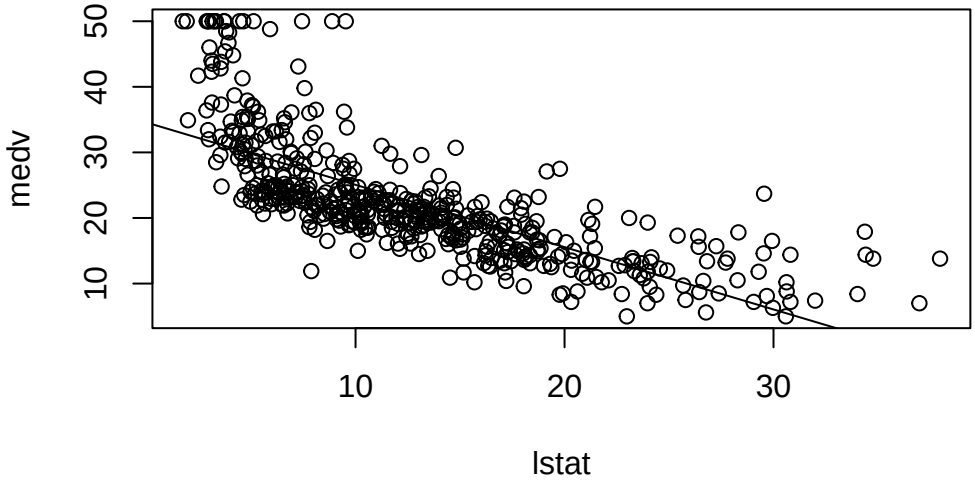
```
       fit      lwr      upr
1 29.80359 29.00741 30.59978
2 25.05335 24.47413 25.63256
3 20.30310 19.73159 20.87461
```

```
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
```

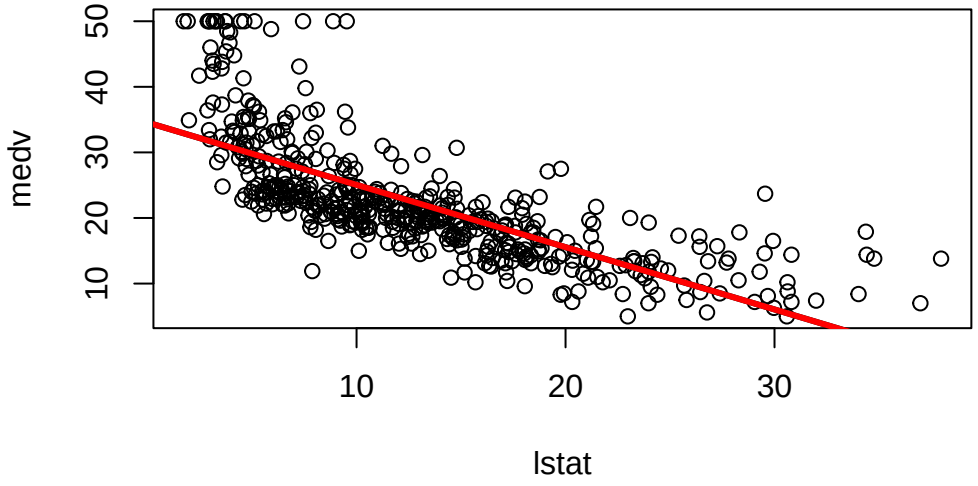
Ahora trazaremos `medv()` y `lstat()` junto con la línea de regresión por mínimos cuadrados mediante las funciones `plkot()` y `abline()`.

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

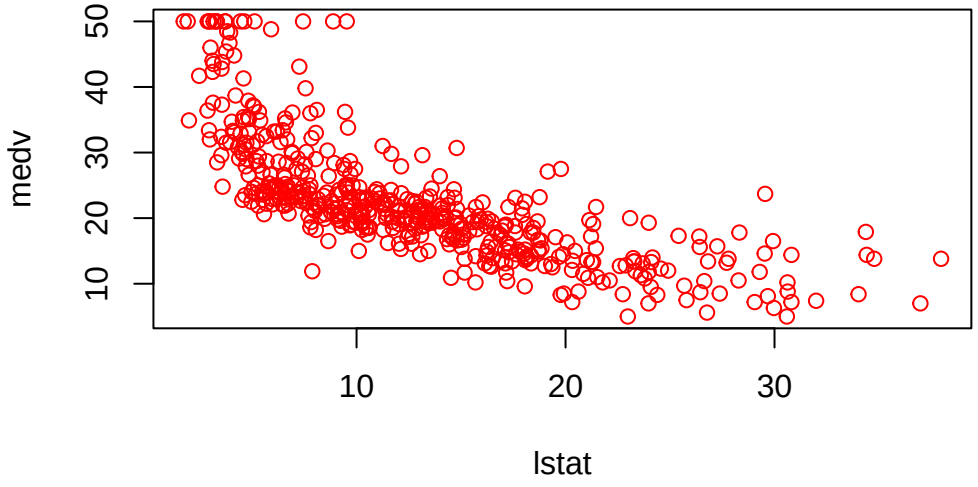


- La función `abline()` sirve para dibujar cualquier línea, no sólo la línea de regresión por mínimos cuadrados. Podemos dibujar una recta con intercepto a y pendiente b, escribimos abline(a, b).
- El comando `lwd=3` hace que la anchura de la línea de regresión se incremente en un factor de 3.
- La funcion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

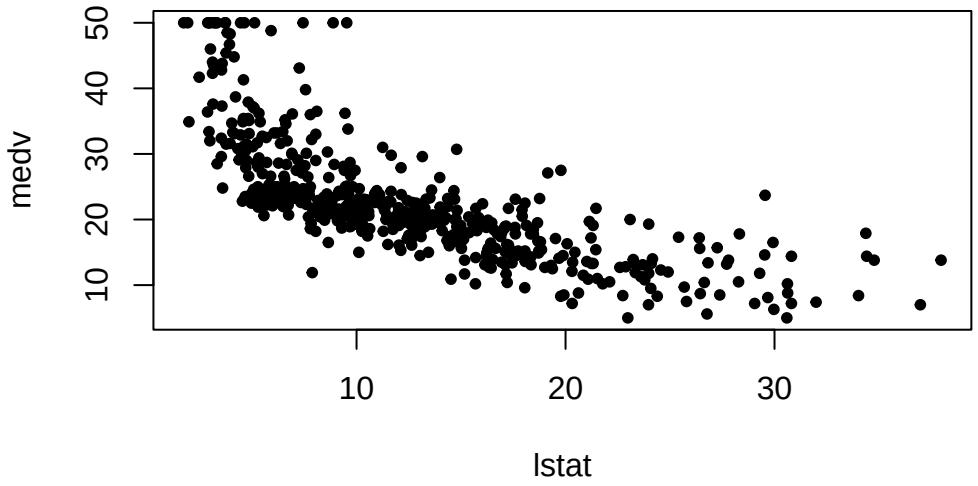
```
plot(lstat , medv )
abline(lm.fit, lwd = 3)
abline(lm.fit, lwd= 3 , col = " red ")
```



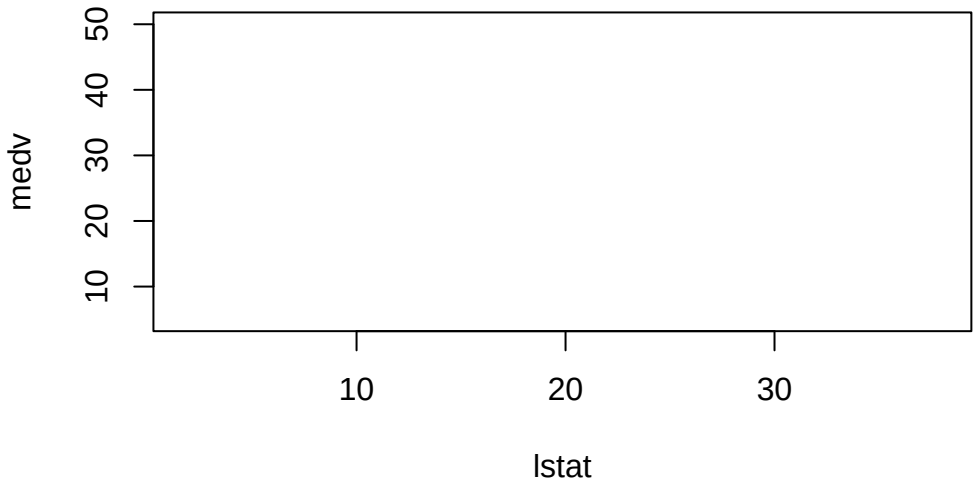

```
plot(lstat, medv , col = " red ")
```



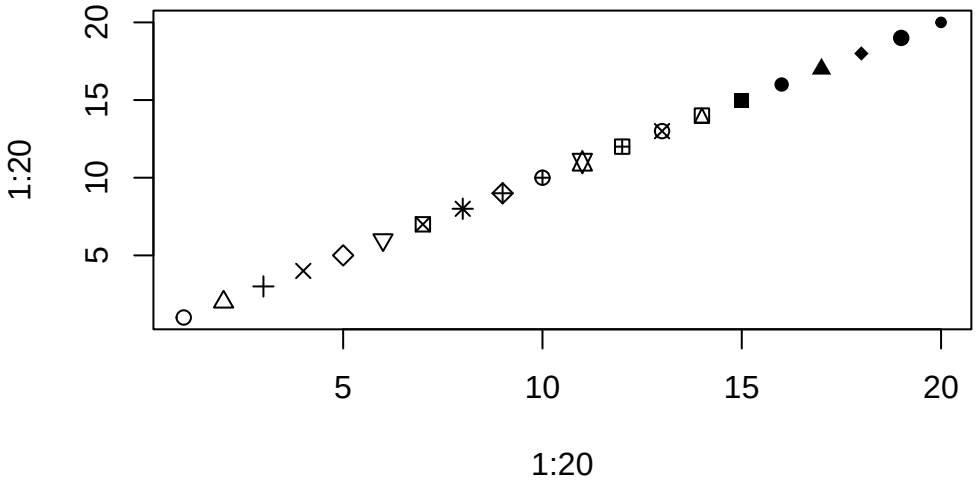
```
plot(lstat, medv , pch = 20)
```




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

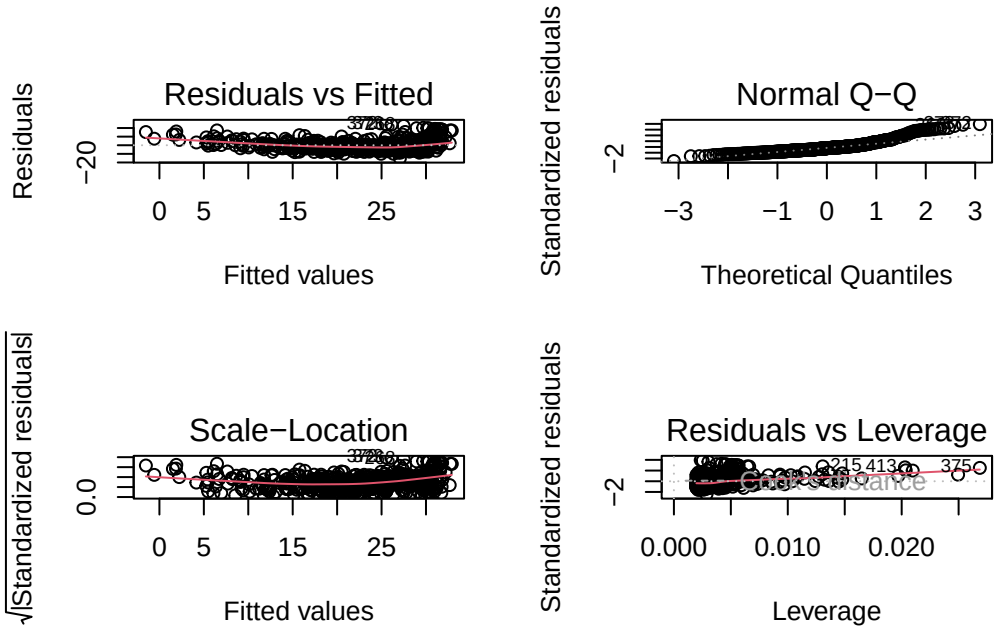



```
plot(1:20, 1:20 , pch = 1:20)
```



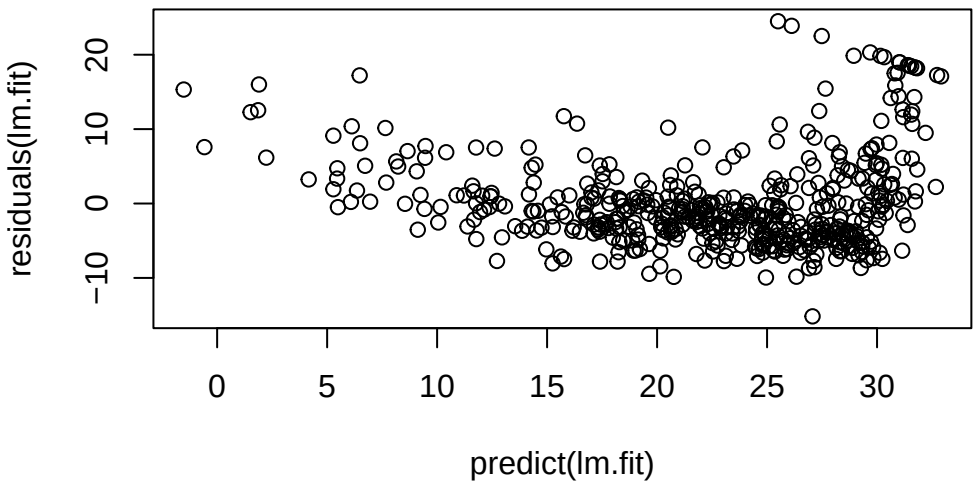
- Cuatro diagramas de diagnóstico se producen al aplicar la función `plot()` directamente a la salida de `lm()`. Este comando producirá un gráfico a la vez, y al presionar Enter se generará el siguiente gráfico. Sin embargo, es conveniente ver los cuatro gráficos juntos.
- Las funciones `par()` y `mfrow()`,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

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

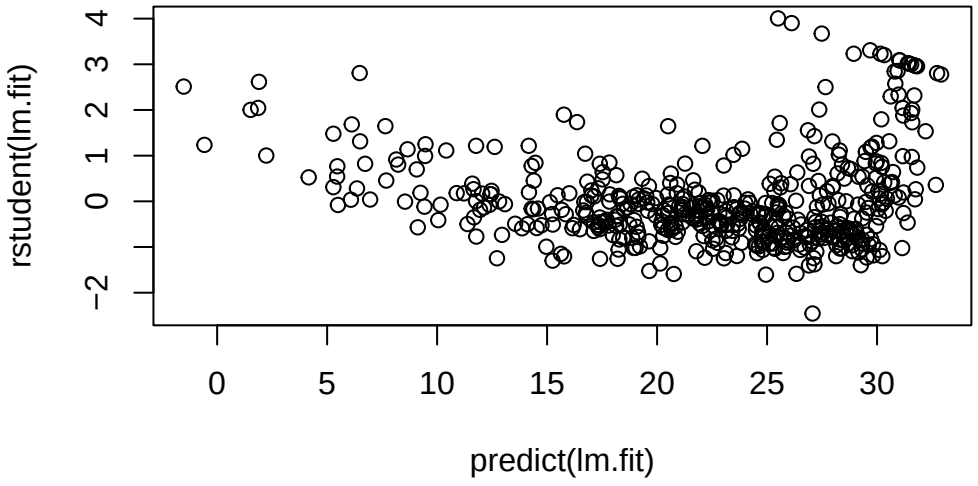


- La función `residuals()` permite calcular los residuos de un ajuste de regresión lineal.
- La función `rstudent()` devolverá los residuos estudentizados, y podemos usarla para graficar los residuos contra los valores ajustados.

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

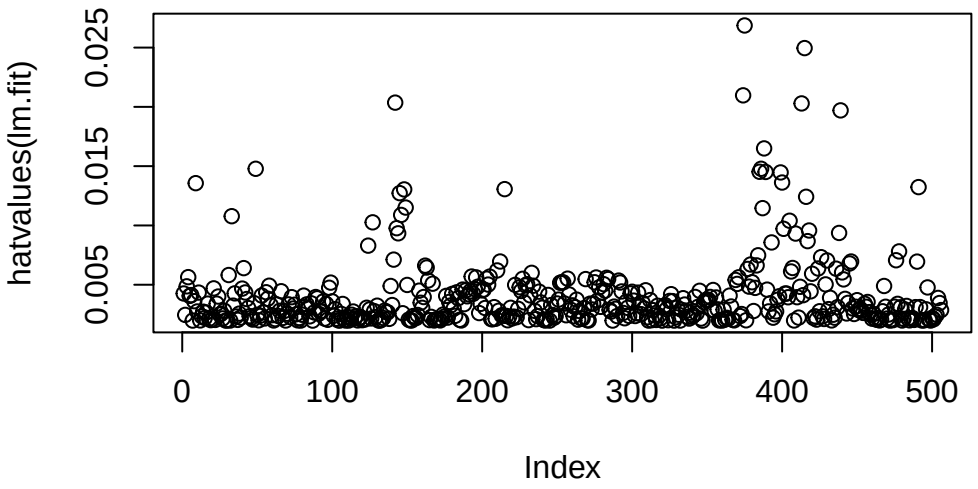



```
plot(predict(lm.fit), rstudent(lm.fit))
```



- 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).

3.6.3 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)
```

3.6.4 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
```

3.6.5 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)
```

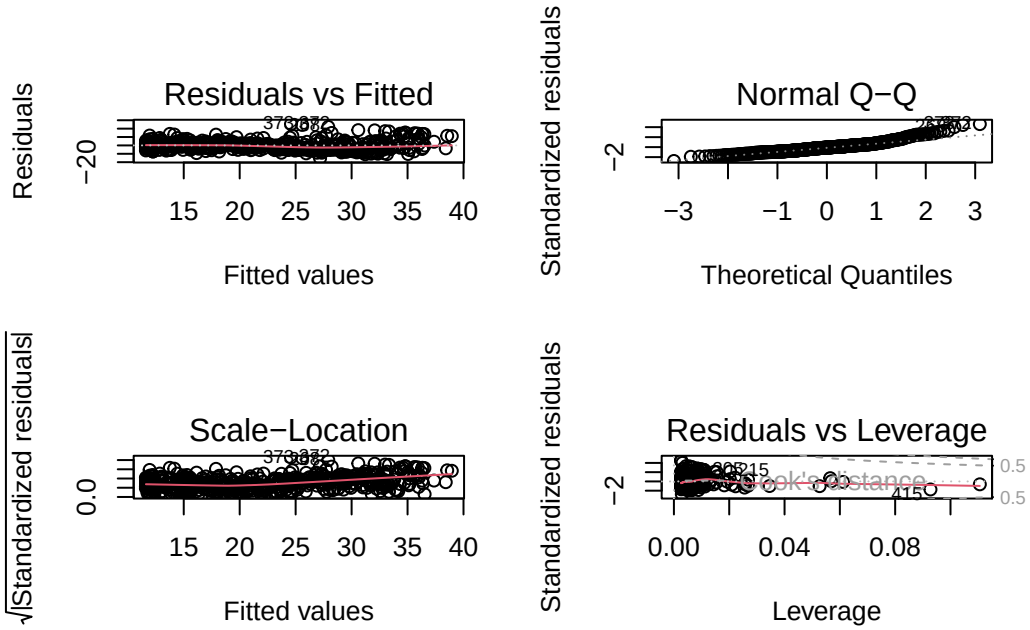
```
Analysis of Variance Table

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
```

3.6.6 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.

```
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.

```
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.

3.6.7 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.

```
{r}
LoadLibraries
```

```
Error: object 'LoadLibraries' not found
```

```
{r}
LoadLibraries()
```

```
Error in LoadLibraries() : 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."
```