

Fundamentos de ciencia de datos con R

Gema Fernández-Avilés y José-María Montero

2023-05-30

Índice general

Prefacio	5
¡Hola mundo!	5
¿Por qué este libro?	6
¿A quién va dirigido?	7
El paquete CDR	8
¿Por qué R?	8
Agradecimientos	9
I Modelización estadística	11
1. Modelización lineal	13
1.1. Modelización	13
1.2. Procedimiento de modelización	15
1.3. Procedimiento con R: la función lm()	17
1.4. Casos prácticos	18
1.5. Comentarios finales	26

Prefacio

¡Hola mundo!

El siglo XXI está siendo testigo de grandes cambios vertiginosos en el contexto social y tecnológico, entre otros. Los tiempos han cambiado, la sociedad se ha globalizado y “exige” respuestas inmediatas a problemas muy complejos. Vivimos en el mundo de la **información**, de los **datos**, o mejor, de las **bases de datos masivas**, y los ciudadanos y, sobre todo, las empresas y los gobiernos, dirigen su mirada hacia el mundo científico para que les ayude a “**oír las historias**” que cuentan esos datos acerca de la realidad de la que han sido extraídos. Y dado su enorme volumen y sofisticación (en el nuevo mundo las imágenes y los textos, por ejemplo, también son datos), exigen algoritmos de nueva generación en el campo del *machine learning*, o incluso del *deep learning*, para “oír las historias” que cuentan. No parecen mirar al “antiguo” investigador científico, sino al “nuevo” *científico de datos*.

Ello, inevitablemente, se traduce en la necesidad de profesionales con una gran capacidad de adaptación a este nuevo paradigma: los científicos de datos, también llamados por algunos los “nuevos hombres del Renacimiento”, para lo cual las Universidades y demás instituciones educativas especializada se apresuran a incluir el grado de Ciencia de Datos en su oferta educativa y a ofrecer seminarios de software estadístico de acceso abierto para sus estudiantes de primeros cursos.

Con la emergencia de la nueva sociedad, en la que el manejo de la ingente cantidad de información que genera se hace absolutamente necesario para circular por ella, la **Ciencia de Datos** ha venido para quedarse. Sin embargo, el mundo de la Ciencia de Datos es cualquier cosa menos sencillo. En él, cualquier ayuda, cualquier guía es bienvenida. Por ello, es muy recomendable que la persona que se quiera introducir en él, sea con fines de investigación o con fines profesionales, se agarre de la mano de un guía especializado que le lleve, de una manera amena, comprensible y eficiente, desde el planteamiento de su problema y la captura de la información necesaria para poderle dar una solución, hasta la redacción de las conclusiones finales que ha obtenido con los modernos informes reproducibles colaborativos. Y como en la parte central de ese camino tendrá que luchar con grandes gigantes (en la actualidad denominados técnicas estadísticas y algoritmos), el guía tendrá que explicarle, de manera sencilla y amena, en qué consiste la lucha (las técnicas y los algoritmos) y cómo llegar a la victoria lo más rápido posible, enseñándole a moverse por el mundo del software estadístico, en nuestro caso **R**, que le permitirá realizar los cálculos necesarios para vencer al problema planteado a una velocidad vertiginosa.

En resumen, la información masiva y el moderno tratamiento estadístico de la misma son la “mano invisible” que gobierna la sociedad del siglo XXI, y este manual pretende ser el guía anteriormente mencionado que le llevará de la mano cuando quiera caminar por ella.

¿Por qué este libro?

Lo dicho anteriormente ya justifica por sí solo la aparición de este manual. Afortunadamente, no es el primero en la materia, pues son ya bastantes los materiales de calidad publicados sobre Ciencia de Datos. Sin embargo, quizás, éste pueda ser considerado el más completo. Y ello por varias razones.

La primera es su **completitud**: este manual lleva de la mano al lector desde el planteamiento del problema hasta el informe que contiene la solución al mismo; o desde no saber qué hacer con la información de la que dispone, hasta ser capaz de transformar tales bases de datos masivas, y casi imposibles de manejar, en respuestas a problemas fundamentales de una empresa, institución o cualquier agente social.

La segunda es su **amplitud temática**:

- (I) Parte de las dos primeras preguntas que un neófito se puede hacer sobre esta temática: ¿qué es eso de la Ciencia de Datos que está en boca de todos? Y, ¿qué diablos es **R** y cómo funciona?
- (II) Enseña cómo moverse en la jungla del *Big Data* y de los “nuevos” tipos de datos, siempre bajo el paraguas de la ética de los datos y del buen gobierno de dichos datos.
- (III) Muestra al lector cómo obtener conocimiento de la oscuridad del enorme banco de información a su disposición, que no sabe cómo abordar ni manejar.
- (IV) No deja a nadie atrás, y de forma previa al contenido central del manual (las técnicas de Ciencia de Datos), incluye unas breves, pero magníficas, secciones sobre los rudimentos de la probabilidad, la inferencia estadística y el muestreo, para aquéllos no familiarizados con estas cuestiones.
- (V) Aborda una treintena de técnicas de Ciencia de Datos en el ámbito de la modelización, análisis de datos cualitativos, discriminación, *machine learning* supervisado y no supervisado, con especial incidencia en las tareas de clasificación y clusterización -así como, en el caso no supervisado, de reducción de la dimensionalidad, escalamiento multidimensional y análisis de correspondencias-, *deep learning*, análisis de datos textuales y de redes, y, finalmente, ciencia de datos espaciales (desde las perspectivas de la geoestadística, la econometría espacial y los procesos de punto).
- (VI) Hace especial hincapié en la reproducibilidad en tiempo real (o no) entre los distintos miembros de un equipo (sea universitario, empresarial, o del tipo que sea) y en la difusión de los resultados obtenidos, enseñando al lector cómo generar informes reproducibles mediante RMarkdown y documentos Quarto o en otros modernos formatos.
- (VII) Dedica un capítulo a la creación de aplicaciones web interactivas (con Shiny).

Índice general

7

- (viii) Para aquéllos con pasión por la codificación, y que quieran compartir código y colaborar con otros desarrolladores, este manual aborda la gestión rápida y eficaz de proyectos (del tamaño que sean) mediante Git, un sistema de control de versiones distribuido, gratuito y de código abierto, y GitHub, un servicio de alojamiento de repositorios Git del cual, aquellos no familiarizados con la cuestión de la codificación, o con aversión a ella, podrán tomar el código que necesitan.
- (ix) Muestra al lector los primeros pasos para iniciarse en el geoprocесamiento en la nube.
- (x) Y, finalmente, aborda más de una docena de casos de uso (en medicina, periodismo, economía, criminología, marketing, moda, demanda de electricidad, cambio climático, reconocimiento de patrones en la forma de tuitear...) que ilustran la puesta en práctica de todos los conocimientos anteriormente adquiridos.

La cuarta razón es que todo lo que el lector aprende en este manual lo puede reproducir y poner en práctica inmediatamente con **R**, puesto que el manual está trufado de *chunks* (o trozos de código **R**) que no tiene más que cortar y pegar para reproducir los ejemplos que se muestran en el libro, cuyos datos están en el paquete CDR; o utilizar dichas *chunks* para abordar el problema que le ocupa con los datos que tenga a su disposición. Una buena razón, sin duda. Por consiguiente, el manual es una buena combinación “teoría-práctica-software” que permite abordar cualquier problema que el científico de datos se plante en cualquier disciplina o situación empresarial, médica, periodística...

La quinta es su **variedad de perspectivas**. Son **más de 40 los participantes** en este manual. Algunos de ellos, prestigiosos profesores universitarios; otros, destacados miembros de instituciones públicas; otros, CEOs de empresas en la órbita de la ciencia de datos; otros, *big names* del mundo de **R** software... El manual es, sin duda, un magnífico ejemplo de colaboración Universidad-Empresa para buscar soluciones a los problemas de las sociedades modernas.

¿A quién va dirigido?

Fundamentos de ciencia de datos con R está dirigido a todos aquellos que desean desarrollar las habilidades necesarias para abordar proyectos complejos de Ciencia de Datos y “pensar con datos” (como lo acuñó Diane Lambert, de Google). El deseo de resolver problemas utilizando datos es su piedra angular. Por tanto, como se avanzó anteriormente, este manual no deja a nadie atrás, y lo único que requiere es “el deseo de resolver problemas utilizando datos”. No excluye ninguna disciplina, no excluye a las personas que no tengan un elevado nivel de análisis estadístico de datos, no excluye a nadie. Se ha procurado una combinación de rigor y sencillez, y de teoría y práctica, todo ello con sus correspondientes códigos en **R**, que satisfaga tanto a los más exigentes como a los principiantes.

También está destinado a todos aquellos que quieran sustituir la navegación por la web (la búsqueda del video, publicación de blog o tutorial *online* que solucione su problema –frustración tras frustración por la falta de consistencia, rigor e integridad de dichos materiales, así como por su sesgo hacia paquetes singulares para la implementación de las cuestiones que tratan–), por

una “**biblia de la ciencia de datos**” rigurosa pero sencilla, práctica y de aplicación inmediata sin ser ni un experto estadístico ni un experto informático.

Pero si a alguien está destinado especialmente, es a la comunidad hispano hablante. Este manual es un guiño a dicha comunidad, para que tenga a su disposición, en su lengua nativa, uno de los mejores manuales de Ciencia de Datos de la actualidad.

El paquete CDR



El paquete **CDR** contiene la mayoría de conjuntos de datos utilizados en este libro que no están disponibles en otros paquetes. Para instalarlo use la función `install_github()` del paquete `remotes`.

```
# este comando sólo necesita ser ejecutado una vez
# si el paquete remotes no está instalado, descomentar para instalarlo

# install.packages("remotes")
remotes::install_github("cdr-book/CDR")
```

La lista de todos los conjuntos de datos puede obtenerse haciendo `data()`.

```
library('CDR')
data(package = "CDR")
```

Este paquete ayudará al lector a reproducir todos los ejemplos del libro. De acuerdo con las mejores prácticas en **R**, el paquete **CDR** sólo contiene los datos utilizados en el libro.

¿Por qué R?

R es un lenguaje de código abierto para computación estadística que se ha consolidado entre la comunidad científica internacional, en las últimas dos décadas, como una herramienta de primer

Índice general

9

nivel, consolidándose como líder permanente en el ámbito de la implementación de metodologías estadísticas para el análisis de datos. La utilidad de **R** para la Ciencia de Datos deriva de un fantástico ecosistema de paquetes (activo y en crecimiento), así como de un buen elenco de otros excelentes recursos: libros, manuales, *blogs*, foros y *chats* interactivos en las redes sociales, y una gran comunidad dispuesta a colaborar, a orientar y a resolver diferentes cuestiones relacionadas con **R**.

Por otra parte, **R** es el lenguaje estadístico y de análisis de datos más utilizado en la mayoría de los entornos académicos y, cómo no, por una larga lista de importantes empresas, entre las que se cuentan Facebook (análisis de patrones de comportamientos relacionado con actualizaciones de estado e imágenes de perfil), Google (para la efectividad de la publicidad y la previsión económica), Twitter (visualización de datos y agrupación semántica), Microsoft (adquirió la empresa Revolution R), Uber (análisis estadístico), Airbnb (ciencia de datos), IBM (se unió al grupo del consorcio R), New York Times (visualización)...

La comunidad **R** también es particularmente generosa e inclusiva, y hay grupos increíbles, como *R-Ladies* y *Minority R Users*, diseñados para ayudar a garantizar que todos aprendan y usen las capacidades de **R**.

Agradecimientos

No queremos dar por finalizado este prefacio sin agradecer a los 44 autores participantes en esta obra su esfuerzo por condensar, en no más de 20 páginas, la teoría, práctica y tratamiento informático de la parte de la Ciencia de Datos que les fue encargada. Y no sólo eso; el “más difícil todavía” fue que debían dirigirse a un abanico de potenciales lectores tan grande como personas haya con “el deseo de resolver problemas utilizando datos”. Era misión imposible. Sin embargo, a la vista del resultado, ha sido misión cumplida. El esfuerzo mereció la pena.

Además, nos gustaría agradecer el apoyo incondicional recibido por (en orden alfabético): Itzcoatl Bueno, Ismael Caballero, Emilio L. Cano, Diego Henangómez, Ricardo Pérez, Manuel Vargas y Jorge Velasco.

También queremos poner de manifiesto que la edición de este texto ha sido financiada por diversos entes de la Universidad de Castilla-La Mancha. En su mayor parte, por el **Máster en Data Science y Business Analytics (con R software)** (a través de la orgánica: 02040M0280), pero también por la Facultad de Ciencias Jurídicas y Sociales de Toledo (a través de su contrato programa: orgánica 00440710), el Departamento de Economía Aplicada I (mediante sus fondos departamentales, DEAI 00421I126) y el Grupo de Investigación Economía Aplicada y Métodos Cuantitativos (que ha dedicado parte de sus fondos a la edición de esta obra, orgánica 01110G3044-2023-GRIN-34336).

A todos, eternamente agradecidos por ayudarnos en este reto de transformar la oscuridad en conocimiento, de convertir en una ciencia y en un arte la difícil tarea de sacar valor de los datos, el petróleo del futuro. Quizás en este momento no seamos conscientes de que hemos puesto nuestro granito de arena a la ciencia que, a buen seguro, juegue uno de los papeles más importantes de este siglo, caracterizado por el predominio de la información. Una ciencia, la Ciencia de Datos, que combina el análisis estadístico de datos, la algoritmia y el conocimiento del

negocio para sacar valor del bien más abundante de la sociedad en la que vivimos: la información. Una disciplina cuyo dominio caracteriza a los científicos de datos (también denominados los nuevos personajes del Renacimiento), profesión que ya fue calificada hace más de veinte años en la *Harvard Business Review* y en *The New York Times*, entre otros, como la “más sexy del siglo XXI”.

Nota

Este manual está publicado por [McGraw Hill](#). Las copias físicas están disponibles en [McGraw Hill](#). La versión *online* se puede leer de forma gratuita en <https://cdr-book.github.io/> y tiene la [licencia de Creative Commons Reconocimiento-NoComercial-SinObraDerivada 4.0 Internacional](#).

Si tiene algún comentario o sugerencia, no dude en contactar con los editores y los autores. ¡Gracias!

Parte I

Modelización estadística

Capítulo 1

Modelización lineal

Víctor Casero-Alonso^a y María Durbán^b

^aUniversidad de Castilla-La Mancha

^bUniversidad Carlos III de Madrid

1.1. Modelización

Se acude a los **modelos de regresión** para intentar explicar la relación entre dos o más variables. Para ello se predefine un modelo que pretende explicar el comportamiento de la variable **respuesta o dependiente**, denotada por Y , utilizando la información proporcionada por las **variables explicativas**, también llamadas independientes o predictoras, denotadas por X_1, \dots, X_k . Pero dichas variables pueden ser de distinto tipo. Si la variable respuesta es continua, más concretamente, si se puede asumir que sigue una **distribución de probabilidad Normal**, y al menos una de las variables explicativas es también continua, se puede acudir a la **modelización lineal** que se desarrolla en este capítulo. Sin embargo, si la variable respuesta fuese de otro tipo, por ejemplo, dicotómica, la modelización lineal no sería adecuada. En el Cap. ??, en el que se aborda el **modelo lineal generalizado**, quedará más clara esta distinción.

El primer paso en el proceso de modelización es intentar explicar una variable respuesta, que de aquí en adelante se supone continua y con distribución Normal, a partir de una sola de las variables explicativas, de forma *lineal* (**modelo lineal simple**). Dicho modelo probablemente no será “bueno”, no explicará bien el comportamiento de la variable respuesta si la realidad que se pretende explicar es compleja, pero podría ser *suficiente* para el propósito del estudio¹.

¹La capacidad de explicación la proporciona el coeficiente de **bondad de ajuste** o coeficiente de determinación lineal, R^2 (véase la Sec. 1.2.1).

Nota

Se entiende por **modelo lineal** aquel cuya relación entre las variables viene determinada por una combinación *lineal* de los parámetros, por ejemplo:

- $Y = \beta_0 + \beta_1 X + \epsilon.$
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon.$
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \epsilon.$

El último ejemplo refleja un modelo lineal en los parámetros, pero no lineal en las variables, por el término X_1^2 . Ejemplos de **modelos no lineales** en los parámetros son:

- $Y = \beta_0 \cdot e^{\beta_1 X_1} + \epsilon.$
- $Y = \beta_0 + X_1^{\beta_1} + \epsilon.$

Por ejemplo, se sabe que el peso de una persona está relacionado con muchos factores, pero uno de los más determinantes es la altura. Si se recogen datos de pesos y alturas de un conjunto de personas se puede ajustar el modelo y obtener una explicación *suficiente*, aunque parcial, del peso de una persona a partir de su altura. Es claro que la inclusión de otras variables en el modelo puede ayudar a *explicar* mejor la variable respuesta. Se llega así al denominado **modelo de regresión lineal múltiple** que se puede expresar matemáticamente como:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), i = 1, \dots, N. \quad (1.1)$$

donde:

- β_0 es el **término independiente o constante** del modelo,
- β_1, \dots, β_k son los **coeficientes de regresión o parámetros** del modelo, que se estimarán a partir de los datos observados $(x_{1i}, \dots, x_{ki}, y_i)$ y reflejan la magnitud del efecto *lineal* (constante) sobre la variable explicada Y de incrementos unitarios en las variables explicativas X_i .
- y ϵ_i es el **término de error** del modelo, la parte de Y que no es capaz de explicar la parte determinista del mismo ($\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki}$), que se supone sigue una distribución de probabilidad Normal, con media 0 y varianza constante σ^2 ;
- además, se asume que las **observaciones** son **independientes**.

Siguiendo con el ejemplo del peso, añadir alguna variable genética, el sexo u otras, ayudará a mejorar la “bondad” del modelo lineal. Otro ejemplo consiste en pretender explicar el salario en un determinado sector económico en función de los años de experiencia, la formación, la situación familiar, el sexo, etc., de los trabajadores. Nótese que entre las variables explicativas sí puede haber variables de distinto tipo, continuas, categóricas, etc.². Ahora bien, la interpretación de los coeficientes dependerá del tipo de variable al que van asociados, como se verá en los casos prácticos (Sec. 1.4).

²Conviene mencionar la estrecha relación entre regresión lineal múltiple y **ANOVA con varios factores** (véase la Sec. 1.4.6.1).

Un par de referencias para ampliar conocimientos sobre este tema utilizando **R** son [Faraway \(2002\)](#) y [James et al. \(2013\)](#).

1.2. Procedimiento de modelización

1.2.1. Estimación del modelo

Los datos recogidos u observados sirven para **especificar** la relación predefinida de antemano, mediante la **estimación** de los coeficientes β_i que mejor ajustan dicha relación, utilizando el método de **mínimos cuadrados**. Además, los correspondientes contrastes permiten decidir si cada coeficiente es **significativamente distinto de 0**, esto es, si tiene un *efecto* significativo sobre la respuesta,³ en cuyo caso tiene sentido mantener en el modelo la variable a la que va asociado. En la práctica, el coeficiente estimado es *significativo* si su **p-valor** (definido en la Sec. ??) asociado es suficientemente pequeño.

Nota

Se acepta, mayoritariamente, como "suficientemente pequeño" un p-valor inferior a 0.05, lo que supone un nivel de confianza en las estimaciones del 95 %. Pero dicho valor es arbitrario y podrían considerarse otros valores de referencia. Por ejemplo, en las salidas de **R** aparecen otros tres niveles de referencia: 0,1, 0,01 y 0,001. En general, cuanto menor sea el p-valor más confianza se tendrá en las conclusiones.

Como se avanzó anteriormente, si algún coeficiente no es significativo, procede eliminar del modelo la variable explicativa asociada. En tal caso, se vuelven a estimar los coeficientes de las variables que se mantienen hasta llegar a un modelo con todos los coeficientes significativos, iterando las veces necesarias⁴. Para facilitar esta labor, se han desarrollado métodos automáticos de selección de variables, basados en la comparación de la varianza residual (haciendo uso del test *F*), mediante el estadístico AIC (criterio de información de Akaike), etc.⁵ Junto con los contrastes, se pueden aportar los intervalos de confianza de los coeficientes, que, si son significativos, no contendrán el valor 0.

En la nota a pie de página 4 no entiendo lo de "Por ello, a pesar de la no significatividad estadística de algún coeficiente, en ocasiones, la variable asociada se mantiene en el modelo. ¿Te refieres a que son variables que, por ejemplo, están incluidas en la formulación teórica del modelo?

A la par del contraste de significación de cada coeficiente, se obtiene el **contraste de significación global del modelo**. La hipótesis nula es que todos los coeficientes β_1, \dots, β_k son 0. Dicho de otro modo, que el conocimiento de las variables X_1, \dots, X_k no aporta información alguna para explicar los valores de Y .

³Desde el punto de vista estadístico, la influencia/efecto sobre la respuesta no es fruto del azar.

⁴El proceso debe basarse en la relación entre las variables predefinidas de antemano. Por ello, a pesar de la no significatividad estadística de algún coeficiente, en ocasiones, la variable asociada se mantiene en el modelo.

⁵Consúltese el Cap. 10 de [Faraway \(2002\)](#).

También se ha de obtener la **bondad del ajuste** del modelo, normalmente medida por el **coeficiente de determinación lineal**, R^2 (adimensional, que toma valores entre 0 y 1). Para comparar entre diferentes modelos, se utiliza el **R^2 ajustado/corregido**, que tiene en cuenta la composición/complejidad del modelo (número de variables, etc.). Cuanto mejor ajuste el modelo los datos observados, más próximo a 1 será el valor de R^2 (1 indica una relación lineal perfecta entre la variable respuesta y las predictoras). Por el contrario, un R^2 cercano a 0 indica que el modelo estimado ajusta mal los datos.

Es habitual valorar conjuntamente la significación global del modelo, su bondad de ajuste y la significación de cada uno de los coeficientes, considerándose apropiados aquellos modelos que son globalmente *significativos* y tienen la suficiente “bondad”, aunque tengan coeficientes no significativos.

1.2.2. Validación del modelo

Aunque el modelo sea significativo se debe *validar*, es decir, se deben someter a contraste los supuestos estadísticos que subyacen al modelo. Para ello se utilizan los **residuos** del modelo, la parte de Y que no explica la regresión estimada o, en otros términos, la diferencia entre los valores observados y los estimados. Matemáticamente,

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \dots + \hat{\beta}_k x_{ki}), \quad i = 1, \dots, N.$$

En la expresión anterior, al tratarse de residuos, las X y las Y irían en minúscula, no? Son valores observados y valores estimados. Lo he cambiado a minúscula

Los supuestos a contrastar son:

- los residuos han de tener **varianza constante** (por definición tienen **media cero**).
- los residuos han de seguir la **distribución** de probabilidad **Normal**.
- las observaciones tienen que ser **independientes**.
- la **relación** entre la variable respuesta y las explicativas se supone **lineal**.
- las variables explicativas son linealmente independientes: ninguna puede ser explicada como combinación lineal de las otras. En caso contrario, se tendría el conocido problema de la **multicolinealidad** y debería quitarse del modelo la variable explicada por el resto.

1.2.3. Interpretación de los coeficientes

Una vez validado el modelo, se procede a la interpretación de los coeficientes significativos. Teniendo en cuenta la expresión del modelo de regresión lineal múltiple (1.1), la regla general de interpretación de cada uno de los coeficientes de regresión estimado $\hat{\beta}_i$ es simple y directa: el cambio/impacto *medio* en el valor de la variable respuesta Y ante un cambio unitario de una variable explicativa **continua** o ante un cambio de categoría (desde la que se toma como referencia) si la variable es categórica. Y ello *ceteris paribus*, esto es, manteniendo constante el valor de las demás variables explicativas.

1.3. Procedimiento con **R**: la función `lm()`

17

¿continua" no será quantitativa). Si fuera correcto continua, que pasa con las cuantitativas discretas?

Habrá que tener en cuenta la magnitud de cada variable, porque la influencia real en la respuesta podría ser de poca magnitud (quizá por las unidades o escala utilizada), pero significativa estadísticamente.

1.2.4. Predicción

La utilidad del modelo estimado (especificado) queda plasmada en su utilización para **predecir** nuevos valores, \hat{y}_i , a partir del conocimiento/asignación de nuevos valores de las variables explicativas, $\{x_{1i}, \dots, x_{ki}\}$. No obstante, dichas predicciones son valores *esperados (medios)*, pudiéndose construir sus correspondientes intervalos de confianza.

1.3. Procedimiento con R: la función `lm()`

R tiene implementada la función `lm()` para ajustar/estimar modelos de regresión lineal múltiple:

```
lm(formula, data = ..., ...)
```

El argumento mínimo necesario es **formula**, donde se predeterminará la relación entre las variables respuesta y explicativas de una forma bastante intuitiva:

- $Y \sim X$, es la fórmula a utilizar para definir un modelo simple donde Y denota la variable respuesta y X la variable explicativa: $Y = \beta_0 + \beta_1 X + \epsilon$.
- $Y \sim X_1 + X_2$, define un modelo lineal múltiple con 2 variables explicativas: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$.
- $Y \sim X_1 + X_2 + X_3 - 1$ elimina el término independiente, β_0 , del modelo lineal múltiple de 3 variables explicativas.
- ...

En el segundo argumento, **data**, se indica el conjunto de datos donde se encuentran las variables de trabajo. No es especificarlo si están en el **Environment**.

Hay que hacer notar que **R** considera, por defecto, las variables explicativas como cuantitativas. Si se tienen variables categóricas codificadas con números hay que indicarle que las trate como categóricas, usando la función `factor()`⁶. De no hacerlo, la función las consideraría numéricas, con el consecuente error de interpretación de los coeficientes asociados.

A partir de los datos disponibles de Y, X_1, \dots, X_k , la función `lm()` estima los coeficientes $\hat{\beta}_i$ asociados a cada variable X_i , mediante el **método de mínimos cuadrados**, y calcula sus errores estándar, con los que obtiene sus estadísticos de contraste (de la t de Student)⁷ y su

⁶Si no está ya definida como `factor()` en el conjunto de datos.

⁷Con los errores estándar también se pueden obtener los intervalos de confianza de los coeficientes.

significación. En el objeto `lm` que se genera también se almacenan los valores ajustados, residuos, etc., que se pueden mostrar a través de funciones genéricas disponibles en **R**. Algunas de ellas son:

- `print()`: muestra un breve resumen.
- `summary()`: proporciona un resumen completo.
- `coef()`: proporciona las estimaciones de los coeficientes del modelo.
- `confint()`: construye intervalos de confianza para los coeficientes.
- `fitted.values()`: muestra los valores ajustados del modelo (para cada observación del `data.frame`).
- `residuals()`: calcula los residuos del modelo (también para cada observación del `data.frame`).

1.4. Casos prácticos

En esta Sección se utilizan los datos `airquality`⁸, que consisten en 154 medidas (de 6 variables) de calidad del aire en Nueva York. Las variables consideradas aquí son las cuatro siguientes:

- `Ozone`: Concentración media de ozono en la atmósfera (en ppb, partes por billón).
- `Solar.R`: Radiación solar (en lang, Langleys).
- `Wind`: Velocidad media del viento (en mph, millas por hora).
- `Temp`: Temperatura máxima diaria (en grados Fahrenheit).

El objetivo es establecer la relación entre la concentración de ozono en la atmósfera, variable respuesta, y las variables meteorológicas `Solar.R`, `Wind` y `Temp`, variables explicativas. Los valores disponibles de las cuatro variables permiten considerarlas como variables continuas.

Antes de proceder con el ajuste múltiple se pueden realizar los ajustes simples, individuales. La Fig. 1.1 representa 3 regresiones lineales simples, de la variable respuesta `Ozone` sobre cada una de las 3 variables explicativas. Cada gráfico muestra un **diagrama de dispersión** de sólo dos variables, la explicativa en el eje *X* y la respuesta en el eje *Y*, obteniéndose la popularmente denominada **nube de puntos**. En tales diagramas se puede ver si entre las variables hay relación lineal, o no, y en caso de que la haya, si es positiva/directa (a mayores valores de *X*, mayores valores de *Y*) o negativa/inversa. En cada gráfico se ha añadido la correspondiente recta de regresión lineal (con su correspondiente intervalo de confianza), que podría no ser la más apropiada, como parece que ocurre en las regresiones de `Ozone` sobre `Wind` y sobre `Temp`. En ambos casos, la relación parece más bien no lineal, aunque podría ser suficiente (en función del interés del estudio) para explicar relativamente bien el comportamiento del nivel de concentración de ozono. El código para el obtener el primer gráfico sería:

```
library("ggplot2")
ggplot(airquality, aes(Solar.R, Ozone)) +
  geom_point() +
```

⁸Conjunto de datos incluido con la instalación `base` de **R**. Más información ejecutando `?airquality`.

1.4. Casos prácticos

19

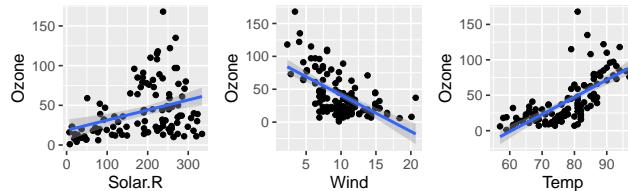


Figura 1.1: Gráficos de dispersión de las variables explicativas frente a la variable respuesta

```
theme(aspect.ratio = 1) +
geom_smooth(method = "lm")
```

En regresión lineal múltiple no es posible visualizar en un sólo gráfico la relación entre la variable respuesta y varias variables explicativas, salvo si son sólo 2, en cuyo caso se tendría un gráfico en 3 dimensiones, generalmente difícil de visualizar.

1.4.1. Estimación de los coeficientes

Se comienza ajustando el siguiente modelo lineal múltiple:⁹:

$$Ozone = \beta_0 + \beta_1 Solar.R + \beta_2 Wind + \beta_3 Temp + \epsilon$$

La definición en **R** del modelo se puede ver como primer argumento de la función `lm()`. El objeto que genera la función `lm()` se guarda bajo el nombre de `airq_lm` y, a continuación, se muestra su resumen con `summary()`:

```
airq_lm <- lm(Ozone ~ Solar.R + Wind + Temp, data = airquality)
summary(airq_lm)
#>
#> Call:
#> lm(formula = Ozone ~ Solar.R + Wind + Temp, data = airquality)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#>
```

⁹Más adelante se introducirá una variable categórica para enriquecer el análisis.

```
#> -40.485 -14.219 -3.551 10.097 95.619
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) -64.34208   23.05472 -2.791  0.00623 **
#> Solar.R       0.05982    0.02319  2.580  0.01124 *
#> Wind          -3.33359    0.65441 -5.094 1.52e-06 ***
#> Temp           1.65209    0.25353  6.516 2.42e-09 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 21.18 on 107 degrees of freedom
#>   (42 observations deleted due to missingness)
#> Multiple R-squared:  0.6059, Adjusted R-squared:  0.5948
#> F-statistic: 54.83 on 3 and 107 DF,  p-value: < 2.2e-16
```

La salida del `summary()` proporciona las estimaciones de los coeficientes del modelo (columna `Estimate`). El término independiente aparece como (`Intercept`) y toma el valor $\beta_0 = -64.3421$, el coeficiente asociado a `Solar.R` es $\beta_1 = 0.0598$, etc. También aparecen sus p-valores asociados (columna `Pr(>|t|)`), pudiéndose comprobar que los 4 coeficientes son significativos al 5 %. Según la leyenda `Signif. codes`, a mayor número de asteriscos mayor significación del coeficiente (menor p-valor). Así, los coeficientes de `Temp` y `Wind` son más significativos que el de `Solar.R`.

También se pueden apreciar (penúltima línea) dos medidas de la bondad del ajuste del modelo considerado: el R cuadrado múltiple y el R cuadrado múltiple ajustado. En el ejemplo, el R^2 (ajustado) es 0.5948, que se podría considerar “suficiente” o no en función del objetivo del estudio, aunque, en este caso, está claro que el modelo no explica suficientemente bien la concentración de ozono.

En la última linea de la salida aparece información sobre el contraste global del modelo: valor del estadístico F , grados de libertad y p-valor asociado. Como se aprecia, el modelo es globalmente significativo (p-valor del orden de 10^{-16}).

1.4.2. Validación

Lo anterior carece de *validez* si no se satisfacen las hipótesis del modelo mencionadas en la Sec. 1.2.2, principalmente las relativas a varianza constante (homocedasticidad) y normalidad. Para ello se realiza un análisis de residuos. La función `autoplot()` del paquete `ggfortify` proporciona los gráficos que se muestran en la Fig. 1.2.

```
library("ggfortify")
autoplot(airq_lm) +
  theme_minimal()
```

Por un lado, el gráfico de residuos frente a valores ajustados (fitted) muestra cierta heterocedasticidad (varianza cambiante con el valor en el eje X) y no linealidad (ya apreciable de

1.4. Casos prácticos

21

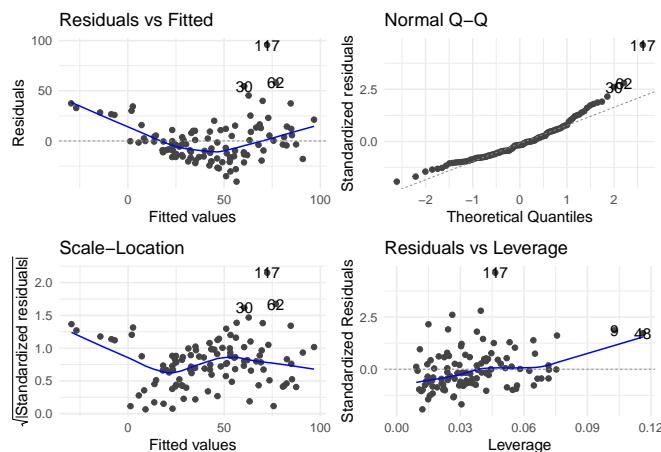


Figura 1.2: Gráficos de residuos

forma individual en la Fig. 1.1). Por su parte, el gráfico Normal Q-Q, que enfrenta los residuos estandarizados con los cuantiles de la distribución Normal, indica que los residuos presentan desviaciones de la normalidad en ambas colas.

Para completar el análisis gráfico se puede acudir a contrastes de hipótesis vistos en la Sec. ???. El más habitual para contrastar normalidad es el de **Shapiro-Wilk**, que se implementa en R con la función `shapiro.test()`¹⁰. Para contrastar la homocedasticidad se puede utilizar alguno de los tres tests implementados para tal fin en el paquete `lmtest()`: el de Breusch-Pagan `bptest()`, el de Goldfeld-Quandt `gqtest()` o el de Harrison-McCabe `hmctest`.

```
shapiro.test(airq_lm$residuals)
#>
#> Shapiro-Wilk normality test
#>
#> data: airq_lm$residuals
#> W = 0.91709, p-value = 3.618e-06
lmtest::bptest(airq_lm)
#>
#> studentized Breusch-Pagan test
#>
#> data: airq_lm
#> BP = 5.0554, df = 3, p-value = 0.1678
```

El contraste de homocedasticidad lleva a no rechazar tal supuesto ($p\text{-valor} > 0.05$), pero el contraste de Shapiro-Wilk confirma la falta de normalidad ($p\text{-valor} < 0.05$). A este respecto, en la Fig. 1.3 se muestra el histograma de la variable `Ozone`, apreciándose que los datos recogidos presentan asimetría incompatible con la normalidad, asumida por defecto para la variable `res`.

¹⁰Se pueden encontrar otros contrastes de normalidad en el paquete `nortest`.

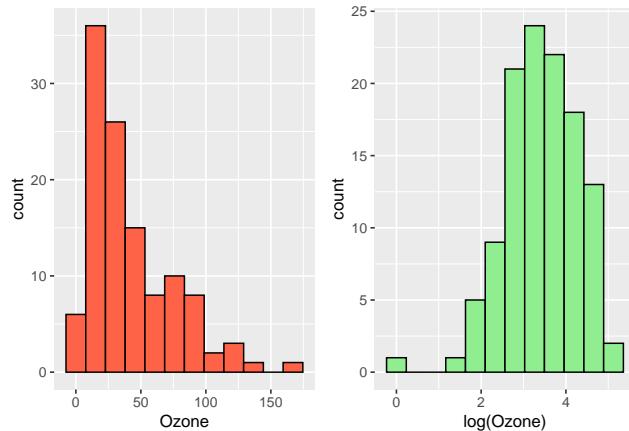


Figura 1.3: Histogramas de las variables ‘Ozone’ y ‘log(Ozone)’

puesta. Una posible solución sería el uso de una transformación logarítmica, que produce cierta simetría en la distribución de la variable, acercándola, por tanto, a la normalidad.

Para el análisis de colinealidad se pueden representar gráficos 2 a 2 de las variables explicativas, para comprobar si están o no correlacionadas (Fig. 1.4).

```
library("GGally")
ggpairs(airquality[, 2:4])
```

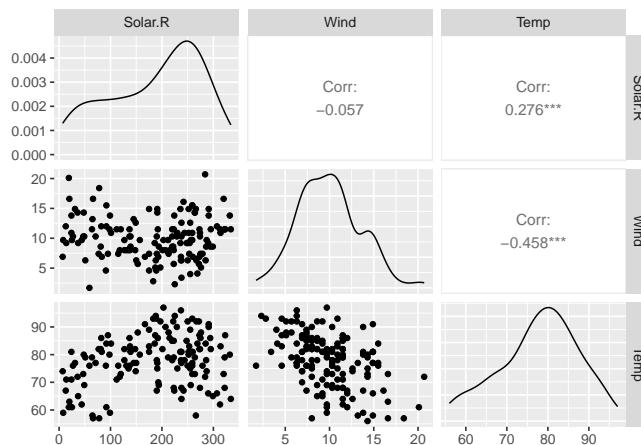


Figura 1.4: Gráfico de dispersión por pares de las variables explicativas

Pero este es un análisis parcial, puesto que una de las variables explicativas podría venir explicada por el resto o varias de ellas. Por si este fuera el caso, conviene calcular también los

factores de inflación de la varianza (VIF), que indican el incremento de la varianza estimada del coeficiente de regresión de una determinada variable explicativa como consecuencia de la colinealidad con las demás (para más detalle, véase el Cap. 3 de James et al. (2013)).

El mínimo valor de VIF es 1, no existiendo límite superior. Una regla general para interpretar los VIF es la siguiente: Si el VIF de una variable explicativa X_i es 1, no hay correlación entre ella y cualquier otra variable explicativa del modelo. Si está entre 1 y 5 la correlación es moderada y no provoca graves problemas. Si es mayor que 5 la correlación es fuerte y, probablemente, las estimaciones de los coeficientes y los p-valores resultantes de la estimación del modelo no sean confiables.

Los VIF se pueden obtener mediante la función `vif()` del paquete `car`:

```
car::vif(airq_lm)
#> Solar.R      Wind      Temp
#> 1.095253 1.329070 1.431367
```

En la Fig. 1.4 se aprecia que los gráficos de dispersión muestran ausencia de correlación entre `Solar.R` y `Wind`; sin embargo, la correlación entre `Wind` y `Temp` no parece despreciable. No obstante, todos los VIF son prácticamente unitarios, por lo que se puede concluir que el modelo no presenta multicolinealidad.

1.4.3. Interpretación de los coeficientes

De acuerdo con lo dicho en la Sec. 1.2.3 se tiene que:

- Un incremento en `Temp` de un grado Fahrenheit, manteniéndose constantes los valores de `Wind` y `Solar.R`¹¹, provoca un aumento (por ser positivo el coeficiente) promedio en el nivel de concentración de ozono en el aire de 1,6521 ppb.
- El coeficiente de `Wind` es negativo, por lo que un aumento en la variable `Wind`, *ceteris paribus*, reduce la concentración de ozono. En concreto, dicha reducción, es de 3,3336 ppb por cada milla por hora que se incremente la variable `Wind`.
- La influencia de `Solar.R` en el nivel de concentración de ozono en la atmósfera es positiva, como la de `Temp`, pero de mucha menor magnitud (por las unidades de una y otra). Concretamente, por cada langley (Ly) que se incremente `Solar.R` el nivel de concentración de ozono se eleva, *ceteris paribus*, en 0,0598 ppb.

Por tanto, el impacto promedio de un incremento unitario en la magnitud de las variables explicativas depende de la variable y de la magnitud de su coeficiente.

Conviene mencionar que las interpretaciones realizadas no deben extrapolarse a valores fuera del rango que toman las variables explicativas, porque en esas regiones podrían darse otros efectos distintos del lineal que presupone el modelo estimado.

¹¹Si estos cambian, tendrán su correspondiente impacto en el nivel de concentración de ozono.

1.4.4. Predicción

Aunque el modelo estimado no es adecuado, por la falta de normalidad, linealidad, etc., detectadas, a continuación se ilustra cómo obtener predicciones con la función `predict()`. Para ello, se asignan los valores de interés a las variables explicativas del modelo, con formato `data.frame`, obteniéndose predicciones del valor medio de la variable respuesta, junto con sus intervalos de confianza o predicción, según se proporcione al argumento `interval` los valores `confidence` o `prediction`, respectivamente. En el siguiente ejemplo se obtienen predicciones de niveles de concentración de ozono para un par de casos elegidos arbitrariamente (el primero corresponde a `Solar.R=50`, `Wind=5` y `Temp=62`):

```
nueva_meteo <- data.frame(
  Solar.R = c(50, 300),
  Wind = c(5, 17),
  Temp = c(62, 90)
)
predict(airq_lm, newdata = nueva_meteo, interval = "confidence")
#>       fit      lwr      upr
#> 1 24.41075 11.01412 37.80739
#> 2 45.62141 31.46838 59.77444
predict(airq_lm, newdata = nueva_meteo, interval = "prediction")
#>       fit      lwr      upr
#> 1 24.41075 -19.662967 68.48448
#> 2 45.62141   1.311914 89.93090
```

Como se puede observar, en ambos casos la predicción puntual (`fit`) es la misma y se obtiene sustituyendo en el modelo estimado los valores de las variables explicativas para los cuales se desea realizar la predicción. Sin embargo, los intervalos de confianza son distintos. Con `confidence` se obtienen intervalos de confianza para el valor medio de las predicciones correspondientes a los días en los que los valores de las variables predictoras sean unos dados. Con `prediction`, el intervalo de confianza es para la predicción de un valor individual, es decir, para la predicción de un día concreto con esas condiciones meteorológicas. Los intervalos de predicción consideran tanto la incertidumbre de la estimación de un valor (debida a la estimación de los parámetros desconocidos) como la variación aleatoria de los valores individuales muestreados (las observaciones muestrales son variables aleatorias). Esto significa que el intervalo de predicción es siempre más ancho que el intervalo de confianza.

1.4.5. Nuevo ajuste con `log(Ozone)`

Ante los problemas de falta de normalidad de la variable `Ozone`, se ajusta un nuevo modelo con la variable `log(Ozone)` como respuesta (véase Sec. ??). Se aprovecha para introducir una variable *dicotómica* para explicar su interpretación. Se define `Temp_f` dicotomizando `Temp` (tomando sólo dos valores): 1, si la temperatura está por encima de su mediana; y 0, si está por debajo.

1.4. Casos prácticos

25

```

mediana <- median(airquality$Temp)
Temp_f <- factor(as.numeric(airquality$Temp > mediana))
lairq_lm <- lm(log(Ozone) ~ Wind + Solar.R + Temp_f, data = airquality)
summary(lairq_lm)
#>
#> Call:
#> lm(formula = log(Ozone) ~ Wind + Solar.R + Temp_f, data = airquality)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -2.55347 -0.29689  0.02409  0.37171  1.18373
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 3.3879872  0.2232099 15.178 < 2e-16 ***
#> Wind        -0.0885666  0.0161038 -5.500 2.61e-07 ***
#> Solar.R      0.0030723  0.0005973  5.143 1.23e-06 ***
#> Temp_f1      0.6999123  0.1158384  6.042 2.25e-08 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.5572 on 107 degrees of freedom
#>   (42 observations deleted due to missingness)
#> Multiple R-squared:  0.5972, Adjusted R-squared:  0.5859
#> F-statistic: 52.89 on 3 and 107 DF,  p-value: < 2.2e-16

```

Al redefinirse la variable respuesta y una variable explicativa del modelo, las estimaciones de los coeficientes cambian respecto al modelo anterior. Todos los coeficientes siguen siendo significativos al 5 % (incluso al 0.1 %), el modelo global también es significativo y el R^2 (ajustado) es similar al del modelo anterior. En el Cap. ??, en el que se abordan los modelos aditivos generalizados, se verá cómo se puede modelar la relación entre `Ozone` y el resto de variables de una forma más satisfactoria. No obstante, este segundo modelo puede ser útil para ilustrar la relación entre las variables, sin olvidar que se ha de comprobar su validez. Para ello, se haría de nuevo el análisis de residuos (que se deja como tarea al lector, al obtenerse de manera idéntica al anterior). En los gráficos de residuos se observará mayor homocedasticidad, linealidad y normalidad que en el caso anterior.

1.4.6. Coeficientes de variables categóricas

A continuación, se aborda la interpretación de coeficientes asociados a variables categóricas. `Temp_f` toma los valores 0 y 1, según la temperatura sea menor o mayor que la mediana respectivamente. En la salida anterior de `R` sólo aparece `Temp_f1`. El 1 final indica que el coeficiente está asociado a la categoría 1 de `Temp_f`. Para los cálculos con estas variables categóricas, `R` toma una categoría como referencia¹² y proporciona un coeficiente para cada una de las restan-

¹²La primera “alfanuméricamente”, si no se especifica expresamente el orden con el argumento `levels` en la función `factor()`.

tes categorías, que representa el cambio *medio* al pasar desde la categoría de referencia a cada una de ellas (técnicamente utiliza variables *dummy*) o diferencia entre la media de la variable respuesta en las observaciones correspondientes a una categoría específica y a la categoría que sirve de referencia. La categoría de referencia está considerada en los cálculos del término independiente del modelo. Por lo tanto, el coeficiente de `Temp_f1` indica que, *ceteris paribus*, la concentración media de ozono de los días con temperaturas por encima de la mediana (categoría 1) es 0.6999 ppb mayor que la de los días con temperaturas inferiores a ella (categoría 0).

1.4.6.1. Comparativa: regresión frente a ANOVA

En la Fig. 1.1 se pueden apreciar regresiones simples *puras* (variable continua sobre variable continua). Si se regresa la variable `Ozone` sobre la variable categórica `Temp_f`, no se obtendrá un gráfico similar¹³. No obstante, el gráfico ayudará a comparar visualmente las medias de la variable respuesta en cada categoría. En realidad, al incluir un *factor* en el modelo se está realizando un contraste *t de Student* para averiguar si existen diferencias entre la media de la variable respuesta para cada categoría con respecto a la categoría de referencia. Técnicamente, tales contrastes dos a dos son equivalentes al contraste ANOVA (análisis de la varianza), aunque este permite comparar si las medias de la variable respuesta en todas las categorías son iguales o no. El ANOVA es un caso particular de regresión lineal en los parámetros, concretamente, cuando todas las variables explicativas son categóricas.

1.5. Comentarios finales

En capítulos posteriores se abordarán modelizaciones más complejas, como por ejemplo, los modelos lineales generalizados, GLM, (Cap. ??), los modelos aditivos generalizados, GAM, (Cap. ??) y los modelos mixtos (Cap. ??). También se verán modelos *sparse* y métodos penalizados de regresión (Cap. ??), como la regresión *ridge*, que permite manejar los problemas que genera la presencia de multicolinealidad.

Queda fuera de este capítulo la inclusión de variables de confusión, que permiten obtener resultados independientemente de los valores de tales variables. El ejemplo típico es la variable sexo: independientemente del sexo del individuo se quiere obtener la influencia de X_i en la respuesta Y . La solución ya ha salido: incluir la variable en el modelo para que, *ceteris paribus*, se pueda interpretar en el análisis.

Sería bueno que pudieseis poner las dos primeras líneas más intuitivas pues lo de obtener resultados es muy genérico. O simplemente plantear el problema ampliando el ejemplo del sexo como variable de confusión

También podrían haberse considerado interacciones entre variables, que se suelen interpretar como sinergias o antagonismos. Pero dada la limitación de espacio y el carácter introductorio de este capítulo no se ha considerado oportuno, pues, además, la interpretación de dichas interacciones suele ser compleja. En el Cap. 3 de James et al. (2013) puede encontrarse un ejemplo.

¹³`ggplot(airquality, aes(Temp_f, Ozone)) + geom_point()`.

Resumen

En este capítulo se introduce el modelo de regresión lineal. En particular:

- se presenta el modelo de regresión lineal múltiple indicando los pasos del análisis de regresión: estimación, validación, interpretación y predicción. La regresión lineal simple se plantea como un caso particular de la múltiple.
- se muestra el uso de **R** para el ajuste de este tipo de modelos.
- se presentan diversos casos prácticos para ilustrar la interpretación de los coeficientes de regresión, tanto asociados a variables continuas como a categóricas, la interpretación de las predicciones y el resto de análisis.
- se mencionan distintos problemas de modelización que el análisis ayuda a detectar, proponiendo a su vez soluciones para solventarlos.

Bibliografía

Faraway, J. J. (2002). *Practical regression and ANOVA using R.*, volume 168. University of Bath Bath.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer. <https://www.statlearning.com/>.

Índice alfabético

- ANOVA, 14, 26
- bondad del ajuste, 15
- coeficientes de regresión, 14
- contraste
 - de Breusch-Pagan, 21
 - de Shapiro-Wilk, 20
- diagrama
 - de dispersión, 18
 - distribución Normal, 13
- error
 - del modelo, 14
- lm, 17
- modelo
 - de regresión, 13
 - de regresión lineal múltiple, 14
 - lineal, 14
 - lineal generalizado, GLM, 13
 - no lineal, 14
- multicolinealidad, 16
- método
 - de mínimos cuadrados, 15, 17
- observaciones independientes, 14
- p-valor, 15
- predicción del modelo lineal, 17
- R cuadrado, 15
- residuos, 16
- significativo, 15
- término independiente, 14
- variable
 - explicativa, 13
 - respuesta, 13