académica

Estadísta aplicada a la valoración
Modelos multivariantes

Francisco Guijarro Martínez

EDITORIAL
UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Estadística aplicada a la valoración

modelos multivariantes

Francisco Guijarro Martínez

Editorial Universitat Politècnica de València

Primera edición, 2013

Distribución: Telf. 963 877 012/ http://www.lalibreria.upv.es / Ref 6140_01_01_01

ISBN: 978-84-9048-119-6 (Versión impresa)

Impreso bajo demanda

Índice

Capítulo 1. Introducción

Aunque no muy frecuentemente, desde algunos foros del campo profesional de la tasación se escuchan voces críticas en referencia a la incorporación de los métodos y técnicas estadísticas en el ámbito de la tasación. Entre las razones esgrimidas se encuentra la excesiva complejidad de estas técnicas cuando la tasación siempre se ha considerado una profesión sencilla en sus métodos, o el excesivo tiempo que su aplicación implicaría en la práctica para el tasador. Además, también en su aplicación es usual encontrarse con la habitual aversión que los cambios, por pequeños que sean, producen en cualquier organización.

Ciertamente la aplicación de estas técnicas puede suponer una complejidad superior a la de otros métodos, como el de homogeneización o corrección. Sin embargo, precisamente este método de homogeneización, tan extendido en la práctica profesional de la valoración inmobiliaria, no resulta ser otra cosa que una simplificación de un método por comparación como es el análisis de regresión. La diferencia fundamental entre ambos enfoques, aunque no la única, es la forma en que se determina el peso o importancia de las variables que intervienen en la valoración. En el método de homogeneización es el propio tasador quien de forma subjetiva establece estos pesos, mientras que en el análisis de regresión es un proceso estadístico quien realiza la ponderación de una manera objetiva y única. Esta diferencia supone que distintos tasadores, en un mismo instante de tiempo, puedan llegar a un resultado bien distinto en función de cómo han interpretado la importancia de cada una de estas variables; o de qué testigos o comparables han utilizado en la tasación. Y como es sabido, está en el espíritu de la normativa internacional de valoración que los métodos, técnicas y procesos aplicados se rijan por la máxima objetividad y transparencia posible.

Lógicamente, no debe ser el tasador quien individualmente y sin ningún apoyo soporte el rigor procedimental de los métodos estadísticos. Debe buscar la colaboración de dos pilares fundamentales para poder llevar a cabo su labor de forma eficaz y eficiente: software específico que permita aplicar los métodos estadísticos de forma rigurosa y con celeridad; una potente base de datos de comparables, proporcionada por una asociación de tasadores, sociedad de tasación, o cualquier otro organismo que pueda desarrollar el doble papel de proveedor de datos y herramientas de tasación, y controlador de la actividad tasadora.

En definitiva, aunque sean comprensibles los temores de parte de la profesión por la introducción de estos métodos y técnicas estadísticas en el ámbito de la valoración, no es menos cierto que ya en la actualidad existen herramientas que permiten su utilización por parte de los profesionales y empresas del sector.

1.1 Valoración y Estadística

El avance de la valoración como práctica profesional en los últimos años ha venido reforzado por el uso de los métodos y técnicas estadísticas. Cuando en la valoración inmobiliaria se nombran expresiones como "valoraciones masivas", se hace referencia a la aplicación de diferentes técnicas provenientes del ámbito científico y desarrolladas por investigadores en estadística.

Si bien en los primeros tiempos la aplicación de estas técnicas era residual, a menudo dificultada por la escasez de programas informáticos y la lentitud en sus procesos, hoy día cuentan con el apoyo de las grandes compañías de tasación, que dedican departamento completos al estudio, análisis, desarrollo e implantación de dichas técnicas sobre áreas muy concretas de la valoración profesional. Incluso también han servido para diversificar las actividades de dichas sociedades, ampliando el abanico de servicios que ofrecen a sus clientes.

No sólo estamos hablando sobre cómo describir una muestra de observaciones o testigos cuando estamos redactando un informe de tasación. Nos referimos, más bien, a cómo poder inferir el precio de las cosas a través de la relación que esta variable guarda con las características que definen los objetos de valoración, y de otros comparables que encontramos en el entorno que sirven para construir dichos modelos estadísticos. En definitiva, la estadística nos proporciona métodos y técnicas que no sólo describen una muestra de observaciones y resumen el comportamiento de sus variables, sino que además son un apoyo fundamental a la hora de construir modelos predictivos, que sean capaces de estimar el precio de mercado alertando, además, del error que puede cometerse en dicha estimación.

Aunque en las tasaciones actuales sólo se informa del valor de mercado más probable, es de esperar que en el futuro los clientes también quieran conocer el "rango" de valores de mercado más probables, como una medida del riesgo asociado a la valoración. Esto es, que no se conformen con un informe en el que se diga que un inmueble es tasado en 300.000€, sino que el valor de dicho inmueble puede fluctuar entre 280.000€ y 320.000€ con un nivel de confianza estadística del 95%. Puede parecer, en estos momentos, una quimera plantear este tipo de situaciones, puesto que hasta ahora no es habitual en las tasaciones hipotecarias. Sin embargo, y a modo de ejemplo, cuando alguien

quiere realizar una inversión financiera no sólo pregunta la rentabilidad promedio de la misma, sino también su nivel de riesgo: sobre qué valores puede eventualmente oscilar dicha rentabilidad. Lo mismo será aplicable, en un futuro, a los informes de tasación.

Gracias al desarrollo de las tecnologías de la información, en paralelo con la mayor capacidad de procesamiento y velocidad de los ordenadores de sobremesa y portátiles, tabletas y *smartphones*, los valoradores pueden apoyarse en todo este conocimiento estadístico para ofrecer un valor añadido y diferenciador en sus informes de tasación.

Pero la estadística no está sólo del lado del valorador como profesional individual, sino que adicionalmente sirve como herramienta de control para las sociedades de tasación. En la mayor parte de los países, la profesión de tasación se encuentra organizada alrededor de estas sociedades, que dan soporte en el día a día a los profesionales del sector. Actualmente, la mayor parte del informe de tasación se desarrolla internamente por estas sociedades, con lo que el tasador se puede concentrar en los detalles técnicos de su trabajo. Cuando se valora un inmueble, la descripción de la zona en que se encuentra, el precio y características de los inmuebles que sirven como comparables, y la organización del propio informe de valoración, es llevada a cabo de forma automática por el software que la sociedad de tasación pone a disposición de sus profesionales. De esta forma, un tasador puede realizar diferentes tasaciones sin presentarse físicamente en la oficina, con la única asistencia de un teléfono inteligente y un software de apoyo que le permita acceder a toda la información de su sociedad de tasación.

Estas nuevas tecnologías también permiten, en este caso a la sociedad de tasación, llevar un control del trabajo efectuado por su plantilla de tasadores. Dicha supervisión es llevada a cabo por los denominados tasadores de control. Estos se encargan de revisar el trabajo de los profesionales, comprobando que la metodología se ajusta a la normativa dictada por el organismo regulador en cada país, y que los resultados concuerdan con los obtenidos en promedio por el resto de valoradores. Así, es fácil detectar cuando un tasador está estimando valores por encima del valor de mercado, o los está infravalorando. Es precisamente en este ámbito donde la estadística juega un papel fundamental. Si un tasador emite informes con precios superiores a los de mercado en un porcentaje que supera un umbral predeterminado, el tasador de control alertará sobre dicha situación, exigiendo al tasador de campo que justifique los valores aportados. ¿Cómo determinar el umbral? Ahí es, precisamente, donde entran en juego los métodos y técnicas estadísticas.

1.2 Población y muestra

Entrando en materia, dos conceptos que un tasador debe ser capaz de distinguir de forma meridiana si quiere aplicar cualquier técnica estadística inferencial son los de población y muestra.

La población está compuesta por todos y cada uno de los elementos que intervienen en un problema, mientras que la muestra será un subconjunto escogido de dicha población. Supongamos, a modo de ejemplo, que a través de un observatorio nuestra sociedad de tasación está llevando a cabo un estudio sobre el precio medio de los apartamentos en la ciudad de Bogotá, Colombia. Para quienes no conozcan dicha ciudad, deben saber que se trata de una urbe de más de 7 millones de habitantes, distribuida en 20 localidades para una mejor gestión administrativa de la misma.

Si para conocer el precio medio de los apartamentos se encuestara a todas y cada una de sus viviendas, diríamos que el análisis ha sido poblacional, pues incluye a todos los elementos (apartamentos) de la ciudad. Además de ser un proceso realmente largo y costoso.

Sin embargo, si los recursos y paciencia de nuestra sociedad de tasación son finitos, resulta plausible intentar estimar el precio medio de los apartamentos a través de un número limitado de viviendas, cuidadosamente escogidas para que sean representativas de la población en su conjunto. Dicho subconjunto de viviendas configuraría la muestra objeto de estudio.

Siguiendo con el mismo ejemplo, si tuviéramos que recopilar el precio considerando únicamente las viviendas que han sido objeto de transacción en el último año, también aquí contaríamos con dos posibilidades. La primera, recopilar la información de todas y cada una de las viviendas que han sido compradas (vendidas) en este último año: población. Lo que también puede convertirse en una tarea ciertamente costosa. O una segunda opción, considerando sólo un subconjunto para intentar ahorrar tiempo y dinero: muestra.

Un último ejemplo, extremo por su lejanía con el ámbito de la valoración y por su planteamiento, pero que dejará clara la diferencia entre población y muestra. Si usted visita al médico por un dolor que le viene afectando durante los últimos días, y el médico solicita que le practiquen una analítica de sangre, ¿qué preferiría, que le tomaran una muestra o la población completa de sangre? Espero que haya escogido la primera opción y se encuentre en condiciones de seguir leyendo el resto del manual.

1.3 Tipos de datos

Como veremos más adelante, resulta fundamental conocer con qué tipo de datos estamos trabajando antes de plantear uno u otro análisis estadístico. En función de cómo sean nuestros datos, deberemos escoger entre las diferentes posibilidades que nos ofrece la estadística. La pregunta que debemos hacernos es, ¿cómo se miden nuestros datos?

1.3.1 Datos nominales o categóricos

Son variables que tienen acotado su rango de valores a un número determinado de posibilidades y que, por tanto, podemos nombrar y enumerar. Si tratamos un conjunto de viviendas y estamos recopilando información sobre las mismas, un típico ejemplo de variable nominal sería el distrito postal en que se encuentran. Si nuestra ciudad tiene 30 distritos postales diferentes, entonces tenemos 30 categorías. No importa que los códigos postales sean números, de igual forma consideraríamos la variables nominal o categórica. Lo mismo ocurriría con la variable orientación, donde podríamos definir cuatro niveles básicos: norte, este, sur, oeste.

Si quisiéramos valorar los derechos de traspaso de un jugador de fútbol profesional, y para ello necesitáramos conocer la posición que ocupa dentro del campo de juego, podríamos definir una variable "posición" con 4 niveles: portero, defensa, centrocampista y delantero. Veremos que, finalmente, necesitamos transformar estas categorías en números para poder tratarlos estadísticamente, pero ello no supone que la variable original deje de ser categórica.

1.3.2 Datos ordinales

Vienen representados por diferentes categorías pero, a diferencia de los anteriores, entre ellos existe un orden. Por ejemplo, si definimos el entorno comercial de una vivienda con tres niveles -Muy bueno, Bueno y Deficiente- es evidente que entre ellos existe una prelación. Supondremos que el mejor entorno comercial es el de las viviendas etiquetadas con un "Muy bueno", seguidas de las "Bueno" y terminando por las "Deficiente".

Al igual que con las variables categóricas, tendremos que pensar en qué forma se incorporan en nuestros análisis estadísticos, ya que cualquiera de las técnicas que utilizaremos en el ámbito de la valoración asume que se trabaja con datos numéricos. En otro capítulo examinaremos en qué forma debe llevarse a cabo esta transformación.

1.3.3 Datos numéricos

Los dos tipos de datos anteriores constituyen el grupo de datos o variables cualitativos. Los datos numéricos en su origen se incluyen en el grupo de datos cuantitativos, y sí permiten su consideración directa en los análisis estadísticos.

Probablemente sean los más comunes, sobre todo cuando hablamos del campo de la valoración. En la valoración inmobiliaria, servirían de ejemplo la superficie (medida en metros cuadrados), el número de dormitorios, el número de baños, la planta en la que se sitúa la vivienda, la distancia al centro de la ciudad, el ancho de la calle, etc. Podemos trabajar tanto con datos enteros como continuos.

Sin embargo, y aún tratándose todas ellas de variables claramente numéricas, el enfoque con que se tratarán en los modelos estadísticos de valoración puede ser muy diferente. Veremos cómo la variable número de dormitorios puede ser tratada de forma distinta a la variable superficie. Por ejemplo, en un modelo de regresión el coeficiente asociado a la variable número de dormitorios informará sobre el incremento medio en el precio de una vivienda por dormitorio adicional. Sin embargo, podemos pensar que no estaríamos dispuestos a pagar la misma cantidad por pasar de 2 a 3 dormitorios, que por pasar de 3 a 4. El incremento marginal puede –debe- ser decreciente, y de alguna manera habrá que incluir dicha premisa en el modelo estadístico. Con la superficie, sin embargo, este efecto puede ser prácticamente insignificante, de forma que el incremento en el precio de pasar de 100 a 101 metros cuadrados pueda considerarse muy similar (o el mismo) que de pasar de 101 a 102 metros cuadrados. En cualquier caso, será el propio modelo estadístico el que nos permita contrastar cada una de estas hipótesis.

1.4 Distribución

La principal forma de describir el comportamiento de una variable es a través de la distribución de sus datos. La forma de su distribución nos informa de cuál es el valor medio o tendencia central de la variable, así como la dispersión o heterogeneidad que podemos encontrar alrededor de ese valor medio.

La distribución más conocida en el ámbito estadístico es la distribución normal, ya que la mayoría de las variables continuas siguen esta distribución. Su forma es la de una clásica campana de Gauss, y se emplea habitualmente en cualquier proceso de inferencia estadística. Esta distribución viene totalmente caracterizada por dos estadísticos: media y desviación típica. La distribución normal estandarizada se distingue del resto por tener media cero y desviación típica uno.

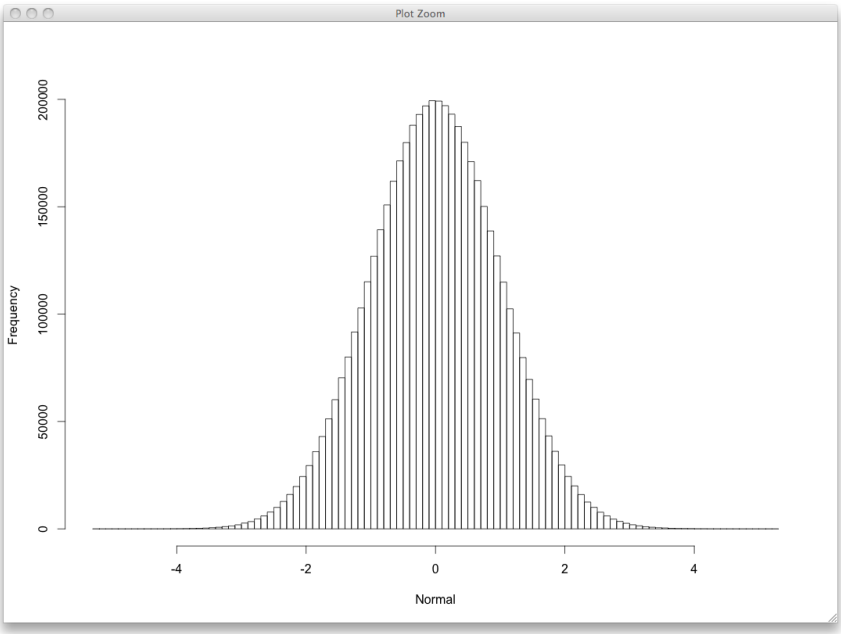


Figura 1. Histograma de una distribución normal estandarizada

1.5 Software estadístico

Sin pretender ser exhaustivos en este punto, nos limitaremos a enunciar algunos de los programas que pueden ser utilizados por el valorador particular a la hora de realizar los cálculos estadísticos de su informe de tasación, o los que emplearía la propia sociedad de tasación si quisiera facilitar esta tarea al profesional o a sus tasadores de control.

1.5.1 Hoja de cálculo Microsoft Excel

Sin duda, la principal ventaja de este software es su elevado grado de implantación. Si hiciéramos una encuesta preguntando si se tiene instalada esta aplicación en sus equipos informáticos, la mayoría del sí sería abrumadora. Además de infinidad de tareas diarias que pueden resolverse con la hoja de cálculo, en la actualidad trae incorporadas una serie de librerías y opciones que permiten llevar a cabo análisis estadísticos muy básicos, pero de gran importancia para el tasador. A lo largo del libro se expondrán diferentes

ejemplos realizados con esta hoja de cálculo, para poder sacarle el mayor partido posible.

Como inconveniente, al menos hasta la versión disponible en el momento de escribir este libro, es la escasa calidad gráfica de sus tablas y figuras, sobre todo si se comparan con otros programas específicos de estadística. Debemos pensar que la parte de estadística de datos es un extra que se ha incorporado a un programa que no está pensado en sí mismo para ser un paquete estadístico. La hoja de cálculo puede servir para muchos propósitos, pero es evidente que en su origen no estaba, por ejemplo, la idea de realizar un análisis de regresión paso a paso; o medir la multicolinealidad en un modelo de regresión; etc.

1.5.2 SPSS (Statistical Package for Social Sciences)

Se trata de una aplicación que incorpora gran cantidad de técnicas estadísticas vinculadas del ámbito de las ciencias sociales. Inicialmente se desarrolló para el tratamiento estadístico de grandes muestras producto de encuestas, si bien con el tiempo ha ido incorporando otras técnicas propias de otras áreas y, sobre todo, añadiendo una mayor capacidad gráfica. Sin duda, destaca entre los programas estadísticos por su facilidad de manejo, y por la alta calidad de sus tablas y gráficos, que se pueden incorporar fácilmente a cualquier informe de tasación. Es muy habitual, incluso, identificar tablas o figuras obtenidas con este programa en artículos científicos de reconocidas publicaciones internacionales.

Para el tasador o sociedad de tasación que quiera llevar a cabo análisis de mayor complejidad que los ofrecidos por una hoja de cálculo, el programa SPSS estaría entre sus imprescindibles. Como inconveniente, al menos el más referido entre los tasadores, su elevado precio.

URL: www.ibm.com/software/es/analytics/spss, donde se puede descargar una versión de evaluación.

1.5.3 R

El paquete R es uno de los de mayor aceptación entre la comunidad estadística y da lugar al lenguaje que lleva su mismo nombre, R, que guarda gran similitud con el lenguaje de programación C. Se trata de un software libre, en el que desarrolladores de todo el mundo van incorporando nuevas capacidades y análisis, lo que facilita la labor del usuario final. Al ser gratuito, su accesibilidad es mucho mayor que la de otros programas.

El principal inconveniente lo encontramos es que no es un programa de tipo ventanas, sino que funciona a base de comandos. De esta manera, si el usuario quiere llevar a cabo un análisis de regresión sobre determinada base de datos, primero tendrá que ejecutar una función que permita leer el fichero de datos, y luego lanzar la función encargada de la regresión. Algo como:

```
> datos <- read.csv("Fichero ejemplo.csv")

> regr <- glm(precio ~ superficie + n.dormit + antig, data = datos)
```

Donde "Fichero ejemplo.csv" sería el archivo donde se encuentran los datos de las viviendas, y la regresión lanzada con el comando "glm" explicaría el precio en función de la superficie, el número de dormitorios y la antigüedad de la vivienda.

Como es de esperar, serán pocos los tasadores que tengan tiempo y conocimiento suficientes para manejar programas de este tipo. Sin embargo, las sociedades de tasación sí pueden permitirse contar en su equipo técnico de desarrollo con programadores o estadísticos que dominen este tipo de programas, y que presten el necesario apoyo técnico a los tasadores de campo y a los tasadores de control.

URL: www.r-project.org.

Capítulo 2. Descripción de los datos

2.1 Introducción

La descripción de los datos permitirá conocer cómo se distribuyen los mismos de una forma rápida, y así tener una primera impresión sobre el comportamiento del precio o cualquier otra variable que estemos analizando. No extraeremos ningún modelo de valoración de la descripción de los datos, pero sí nos ayudará a tomar decisiones sobre qué técnicas estadísticas podemos aplicar para desarrollar un buen modelo de tasación.

Esta descripción se llevará a cabo de forma distinta, según se analicen datos cualitativos (nominales y ordinales) o cuantitativos (numéricos).

2.2 Descripción de datos cualitativos

Estos datos pueden describirse numérica o gráficamente, según interese más en cada caso. Para la descripción numérica puede utilizarse una tabla de frecuencias, mientras que para la gráfica podemos optar por un gráfico de barras. En ambos casos se trata de identificar el número de observaciones registradas para cada uno de los niveles de la variable.

Supongamos que queremos tener una visión rápida de la distribución de la variable Entorno comercial en una muestra compuesta por 107 viviendas. A continuación se representan los valores de las 6 primeras viviendas que componen esta pequeña muestra, para las variables Precio, Superficie, Entorno comercial y Precio por metro cuadrado:

```
> head(datos)
  Precio Superficie Entorno.comercial Precio.m2
1  76000        100        Deficiente  760.0000
2 106000        125         Muy bueno  848.0000
3  93000        103        Deficiente  902.9126
4 111000        120        Deficiente  925.0000
5 100000        108         Muy bueno  925.9259
6  86000         91        Deficiente  945.0549
```

Esta submuestra ha sido seleccionada mediante el comando *head* del programa R[1]. Para estos primeros capítulos utilizaremos algunas funciones de este paquete, si bien la mayoría de los resultados que aparecen en el resto del libro se han obtenido con Excel y SPSS.

La tabla de frecuencias nos informa del número de viviendas que se encuentra en cada una de las tres categorías que antes mencionábamos:

```
> table(datos$Entorno.comercial)

     Bueno Deficiente  Muy bueno

        29         13         65
```

De esta forma, vemos cómo en nuestra muestra la mayoría de las viviendas (65 sobre las 107 totales) se encuentran en una zona de la ciudad con un entorno comercial que habríamos definido como "Muy bueno". El número de viviendas con un entorno comercial "Bueno" asciende a 29, mientras que el nivel menos frecuente es el de viviendas con entorno comercial "Deficiente" (13).

Esta misma información la podríamos representar a través del gráfico de barras, lo que resulta especialmente aconsejable conforme aumenta el número de niveles en la variable. De esta forma, de un simple vistazo se pueden comparar los diferentes niveles que componen la variable. Precisamente en la figura 2 se presenta un gráfico de barras para representar la variable Entorno comercial. Dicho gráfico se ha obtenido con el programa R, y su apariencia es muy similar a la que obtendríamos con otros paquetes.

[1] Debemos resaltar que el programa R utiliza el punto como separador decimal.

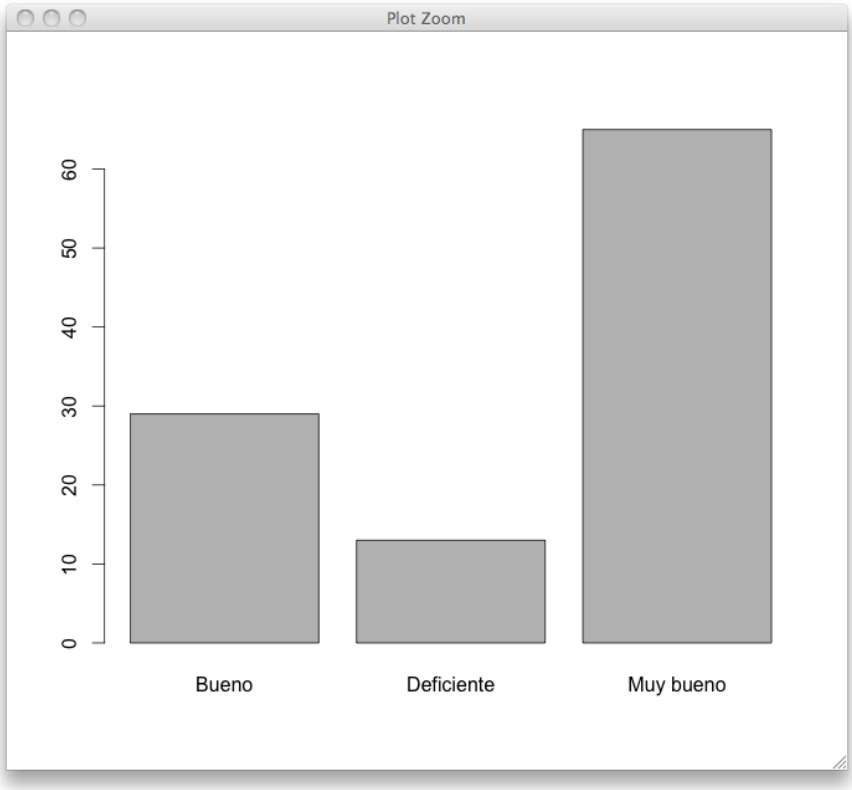


Figura 2. Gráfico de barras de la variable Entorno comercial

2.3 Descripción de datos cuantitativos

Al igual que en el caso de la información cualitativa, los datos cuantitativos pueden ser descritos gráfica y numéricamente. De esta forma tenemos un resumen o visión rápida de los mismos, sin tener que examinarlos uno a uno.

Cuando tratamos de resumir una variable cuantitativa normalmente hacemos uso de los denominados estadísticos descriptivos, que nos informan precisamente de la forma en que se distribuyen los datos. Tenemos dos clases de estadísticos descriptivos: de posición y de dispersión.

2.3.1 Estadísticos de posición

Los estadísticos de posición nos informan acerca de la tendencia central de los datos, siendo los más habituales la media, mediana y moda.

2.3.1.1 Media

La media se obtiene como el promedio de los datos. Así, dada una variable X de la que se dispone de 100 observaciones o registros, la media se calcularía como:

$$Media\ X\ = X = \frac{\sum_{i=1}^{n} X_i}{n}$$

donde X_i se correspondería con el valor de la variable X en la observación i-ésima. A modo de ejemplo, supongamos que queremos calcular la media en el precio de las 6 primeras observaciones de una muestra de datos:

$$Media\ Precio\ = X$$
$$= \frac{76.000 + 106.000 + 93.000 + 111.000 + 100.000 + 86.000}{6}$$
$$= 95.333{,}33€$$

Si quisiéramos calcular la media del precio para toda la muestra, compuesta por $n = 107$ observaciones, tendríamos que incluir el precio de todas ellas en el numerador:

$$Media\ Precio\ = X = \frac{76.000 + 106.000 + \cdots + 252.400}{107} = 136.472{,}10€$$

Luego el precio medio de las viviendas en esta ciudad, según se concluye de la muestra de viviendas analizada, es de 136.472,10€. Esto nos da una idea de la tendencia central de los precios en dicha ciudad, y nos puede permitir compararla con otras ciudades del entorno, o analizar la evolución en el tiempo de los precios medios.

En ocasiones el empleo de la media como estadístico de posición central de los datos puede incurrir en algunos problemas. Es lo que ocurre cuando, por error, hemos introducido un valor que no se corresponde con el real, simplemente porque nos hemos equivocado de tecla al introducirlo en el ordenador.

Supongamos que para el cálculo de la media del precio en la muestra, por error hemos añadido un cero de más al final del último precio considerado,

pasando de 252.400€ a 2.524.000€. Lógicamente este error desvirtuará el cálculo de la media, que ahora será:

$$Media\ Precio\ = X = \frac{76.000 + 106.000 + \cdots + 2.524.000}{107}$$
$$= 157.702{,}00€$$

Esto es, se ha producido un incremento en la media de algo más de 20.000€. En términos relativos, supone un aumento entorno al 15%. Y la distorsión aún habría sido mayor si se disminuyera el número de observaciones en la muestra.

Para evitar este tipo de situaciones, es habitual el empleo de la mediana como estadístico de posición en lugar de la media.

2.3.1.2 Mediana

La mediana es un estadístico de posición de menor sensibilidad que la media a la presencia de casos extremos. Su valor se obtiene como aquél que deja por encima de sí a la mitad de la muestra, y por debajo a la otra mitad. La posición de dicho valor dentro del conjunto de la muestra se obtendrá como:

$$Mediana = \frac{n+1}{2}$$

Para el caso del precio en nuestra muestra de 107 observaciones, la media ocupará la posición:

$$Mediana = \frac{107+1}{2} = 54$$

En cualquier caso, entendemos que previamente hemos ordenado la muestra de observaciones de menor a mayor precio. Los paquetes estadísticos calculan la mediana directamente, de forma que la ordenación por el precio la realizan de forma automática. Sin embargo, si quisiéramos obtener esta ordenación podríamos ejecutar un comando como el siguiente (programa R):

```
> datos$Precio[order(datos$Precio)]

  [1]  67500.0  72000.0  76000.0  84000.0  86000.0  90000.0  93000.0
95000.0  95700.0  95900.0  95900.0

 [12]  96000.0  96720.0  97200.0 100000.0 100000.0 100000.0 100000.0
100200.0 100800.0 101924.6 102000.0

 [23] 102000.0 102000.0 102180.0 102500.0 105120.0 105200.0 106000.0
111000.0 111187.0 111800.0 112000.0

 [34] 113000.0 113000.0 114000.0 114000.0 114000.0 114000.0 114000.0
114000.0 115000.0 115000.0 118320.0

 [45] 118620.0 118800.0 120000.0 120000.0 121000.0 122000.0 122400.0
126000.0 126000.0 126212.5 128000.0

 [56] 128000.0 129600.0 132000.0 132500.0 135000.0 136080.0 136900.0
138000.0 138000.0 138232.0 139000.0

 [67] 139500.0 140000.0 141000.0 143472.0 144000.0 145275.0 150000.0
150000.0 150000.0 150460.0 150600.0

 [78] 150800.0 153000.0 156000.0 159000.0 159900.0 159950.0 160340.0
161500.0 162273.3 162750.0 168000.0

 [89] 168150.0 170000.0 171000.0 171900.0 173520.0 174000.0 175310.0
175680.0 176700.0 180303.6 182970.0

[100] 190260.0 191000.0 192000.0 210000.0 212000.0 240000.0 252400.0
530000.0
```

En el resultado anterior hemos resaltado en negrita el precio que ocupa la posición 54. De esta forma, el valor de la mediana es de 126.212,5.

En los casos en que el número de observaciones en la muestra sea par, el valor de la mediana se calcula como el promedio entre los que ocupan las posiciones $n/2$ y $(n/2)+1$. En el ejemplo donde analizábamos el precio de las 6 primeras viviendas de la muestra la mediana se obtendría como:

```
> datos$Precio[order(head(datos$Precio))]

[1]  76000  86000  93000 100000 106000 111000

> median(datos$Precio[order(head(datos$Precio))])

[1] 96500
```

Hemos resaltado de nuevo en negrita los dos valores que en este caso nos sirven para calcular la mediana. Al tener 6 observaciones, los registros que deberíamos considerar serían los que ocupan las posiciones $6/2 = 3$ y $(6/2) + 1 = 4$. La vivienda que aparece en tercera posición tiene un precio de 93.000€, mientras que la que ocupa la cuarta posición registra un precio de 100.000€. El promedio de ambos valores, 96.500€, es la mediana del precio en esta submuestra.

2.3.1.3 Moda

La moda se define como aquél valor que aparece con más frecuencia entre los datos. Se utiliza fundamentalmente en el caso de variables tipo ordinal.

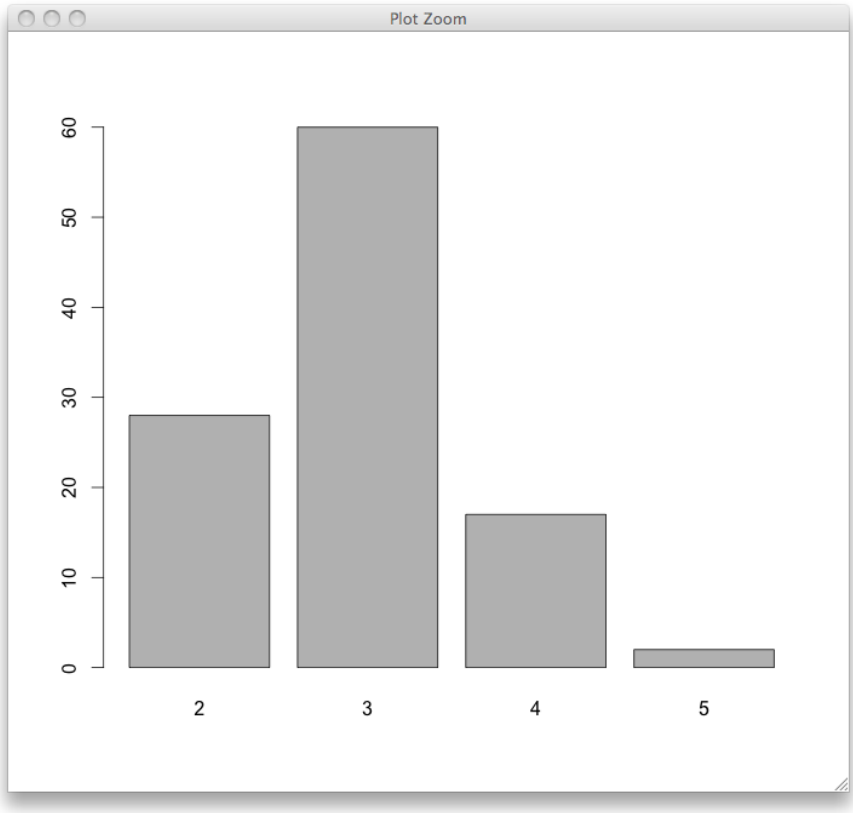


Figura 3. Gráfico de barras de la variable Número de habitaciones

Pongamos como ejemplo la variable número de habitaciones de la muestra compuesta por 107 viviendas. En este conjunto aparecen 28 viviendas con 2 habitaciones, 60 con 3 habitaciones, 17 con 4 habitaciones y únicamente 2 viviendas con 5 habitaciones. Claramente, la categoría más frecuente es la de viviendas con 3 habitaciones, luego la moda de la variable número de habitaciones será 3.

Lógicamente, la moda no será un buen representante de la tendencia central de los datos cuando la variable analizada sea una variable numérica con un amplio rango de valores, como por ejemplo la superficie en metros cuadrados, o la distancia al centro de la ciudad. En esos casos el número de valores repetidos es demasiado escaso como para que la moda pueda considerarse un estadístico de posición representativo de la tendencia central de los datos.

2.3.2 Estadísticos de dispersión

Los estadísticos de dispersión nos informan sobre la variabilidad o heterogeneidad en la distribución de los datos. En algunos casos, la medición de la dispersión es tan importante o más que la tendencia central de los datos. Para los inversores bursátiles, por ejemplo, es indispensable no sólo conocer la rentabilidad de los activos financieros, sino también su grado de volatilidad. Sin ambas mediciones, no parece adecuado tomar una decisión de inversión desde un punto de vista racional.

En el caso de los valoradores, conocer la dispersión de los datos también es de gran relevancia. Supongamos que un valorador tiene los datos medios sobre el precio en vivienda de dos municipios cercanos, y que dichos valores son prácticamente iguales. Podría entonces pensar que se trata de municipios muy homogéneos, y que la función de valoración que utilice en uno de ellos es perfectamente extrapolable al otro. Sin embargo, puede que al analizar con mayor detalle la distribución de los precios descubra que la dispersión de los mismos en un municipio es mucho mayor que en el otro. Ello le podría estar indicando que existen viviendas con calidades muy diferentes, o que existe una gran heterogeneidad en la antigüedad de las viviendas, etc. Esto es, que muy probablemente deba buscar variables que le permitan explicar la diferencia de precios entre las viviendas, y enriquecer el modelo de valoración frente al del municipio con precios más homogéneos.

2.3.2.1 Rango

El rango mide la diferencia entre el mayor y el menor valor de una variable, siendo la medida de dispersión más sencilla de aplicar.

Supongamos que queremos conocer el rango de precios en nuestro municipio de ejemplo, para conocer cuán homogéneas o heterogéneas son las viviendas. Si la vivienda más económica tiene un precio de 55.000€ y la más costosa de 425.000€, el rango se obtendría como la diferencia entre estos dos valores:

$$Rango = 425.000 - 55.000 = 370.000€$$

Pese a su simplicidad, el rango no suele ser el estadístico de dispersión más empleado. El principal motivo de su escasa utilización es que sólo emplea la información de dos observaciones, las más extremas, sin considerar qué ocurre con el resto de la muestra.

2.3.2.2 Varianza

La varianza se define a través de la siguiente expresión:

$$Varianza\ X = \sigma_X^2 = \frac{\sum_{i=1}^{n} X_i - X^{\ 2}}{n}$$

Su utilización es generalizada en el caso de datos numéricos.

Tomemos como ejemplo de cálculo del precio de la vivienda con los siguientes valores, correspondientes a un pequeño distrito en un casco urbano:

Número de vivienda i	Precio (€) X_i	$X_i - X$	$X_i - X\ ^2$
1	258.000	72.189,5	5.211.323.910,3
2	128.000	-57.810,5	3.342.053.910,3
3	133.500	-52.310,5	2.736.388.410,3
4	171.300	-14.510,5	210.554.610,3
5	135.500	-50.310,5	2.531.146.410,3
6	140.000	-45.810,5	2.098.601.910,3
7	138.000	-47.810,5	2.285.843.910,3
8	222.500	36.689,5	1.346.119.410,3
9	113.000	-72.810,5	5.301.368.910,3
10	108.000	-77.810,5	6.054.473.910,3
11	145.000	-40.810,5	1.665.496.910,3
12	204.350	18.539,5	343.713.060,3
13	191.900	6.089,5	37.082.010,3
14	181.500	-4.310,5	18.580.410,3
15	258.500	72.689,5	5.283.763.410,3
16	283.000	97.189,5	9.445.798.910,3
17	240.000	54.189,5	2.936.501.910,3
18	234.940	49.129,5	2.413.707.770,3
19	245.220	59.409,5	3.529.488.690,3
20	184.000	-1.810,5	3.277.910,3

En la muestra se han recogido un total de 20 inmuebles, con los precios que aparecen en la segunda columna. La media en el precio toma el siguiente valor:

$$X = \frac{258.000 + 128.000 + \cdots + 184.000}{20} = 185.810{,}5€$$

La tercera columna refleja la diferencia entre el precio de cada inmueble respecto de la media ($X_i - X$), mientras que la cuarta y última columna eleva al cuadrado los valores de la columna anterior $X_i - X\ ^2$.

De esta forma, la varianza se obtendría como el promedio de los valores de la cuarta columna.

$$Varianza\ X = \sigma_X^2 = \frac{\sum_{i=1}^{n} X_i - X^{\,2}}{n}$$
$$= \frac{5.211.323.910{,}3 + 3.342.053.910{,}3 + \cdots + 3.277.910{,}3}{20}$$
$$= 2.839.764.314{,}8$$

Uno de los problemas que habitualmente se atribuye a la varianza es el hecho de que su unidad de medida no es la misma que la unidad de medida de la variable sobre la que se calcula. Esto es, si el precio se mide en euros, la varianza del precio no se mide en euros. Realmente, la unidad de medida empleada en la varianza para este caso serían euros al cuadrado, ya que la varianza se ha construido precisamente como un promedio de diferencias cuadráticas entre precios. Esto hace que los valores obtenidos puedan parecer exageradamente altos y, sobre todo, no aportar información valiosa para el valorador.

Es por ello que a la hora de describir la dispersión de una variable resulta más habitual el empleo de la desviación típica.

2.3.2.3 Desviación típica

La desviación típica, también denominada desviación estándar (del inglés *standard deviation*), se define como la raíz cuadrada de la varianza:

$$Desviación\ típica\ X = \sigma_X = \sqrt{Varianza\ X} = \sqrt{\frac{\sum_{i=1}^{n} X_i - X^{\,2}}{n}}$$

Su principal virtud es que mantiene la misma unidad de medida que la variable para la que se aplica. En el caso del precio de las viviendas, si dicho precio viene expresado en euros, entonces también la desviación típica del precio viene expresada en euros.

$$Desviación\ típica\ X = \sigma_X = \sqrt{2.839.764.314{,}8} = 53.289{,}44$$

Además de mantener la misma unidad de medida que la variable de referencia, otra razón por la que la desviación típica es la medida de dispersión más ampliamente utilizada es su uso en inferencia estadística. Por inferencia se

entiende la inducción que, a partir de los datos recogidos en una muestra, podemos realizar sobre una población. En estos casos es donde la desviación típica, junto con la media de los datos, juega un papel fundamental.

Supongamos que los precios en el núcleo urbano que estamos analizando siguen una distribución normal, y que la muestra es representativa del comportamiento de los precios en el resto de la ciudad. En este caso, podríamos inferir que la media del precio de la vivienda en toda la ciudad coincide con la media de nuestra muestra:

$$Media\ Ciudad\ = Media\ Muestra\ = 185.810{,}5€$$

Pero, además, también podríamos acotar entre qué rango de valores se mueve un porcentaje significativo de viviendas. Así, podremos afirmar que el precio de las viviendas estará dentro de los siguientes rangos para los niveles de confianza del 90%, 95% y 99%:

Nivel de confianza del 90%: $X \pm 1{,}645\sigma_X = X - 1{,}645\sigma_X, X + 1{,}645\sigma_X$

Nivel de confianza del 95%: $X \pm 1{,}96\sigma_X = X - 1{,}96\sigma_X, X + 1{,}96\sigma_X$

Nivel de confianza del 99%: $X \pm 2{,}576\sigma_X = X - 2{,}576\sigma_X, X + 2{,}576\sigma_X$

Por lo tanto, y debiendo reiterar que estos cálculos serán válidos si 1) los precios siguen una distribución normal, y 2) la muestra es representativa del conjunto de viviendas de toda la ciudad, podremos afirmar que la media de los precios se mueve dentro de los siguientes intervalos:

Nivel de confianza del 90%:

$$X \pm 1{,}645\sigma_X =$$
$$185.810{,}5 - 1{,}645 \times 53.289{,}44 \quad 185.810{,}5 + 1{,}645 \times 53.289{,}44 \ =$$
$$98.149{,}37 \quad 273.471{,}63$$

Nivel de confianza del 95%:

$$X \pm 1{,}96\sigma_X =$$
$$185.810{,}5 - 1{,}96 \times 53.289{,}44 \quad 185.810{,}5 + 1{,}96 \times 53.289{,}44 \ =$$
$$81.363{,}20 \quad 290.257{,}80$$

Nivel de confianza del 99%:

$$X \pm 2{,}576\sigma_X =$$
$$185.810{,}5 - 2{,}576 \times 53.289{,}44 \quad 185.810{,}5 + 2{,}576 \times 53.289{,}44 =$$
$$48.536{,}90 \quad 323.048{,}10$$

Esto nos asegura que si, por ejemplo, extraemos otra muestra de la misma ciudad, el precio medio de las viviendas estará dentro del intervalo $81.363{,}20 \quad 290.257{,}80$ con una probabilidad o nivel de confianza del 95%. Esto es, no todas las muestras tendrán los precios de sus viviendas dentro de este rango, pero prevemos que en el 95% de los casos sí ocurrirá.

Lógicamente, el interés de todo valorador será poder realizar un análisis en el que la media esté lo más acotada posible y, por lo tanto, el anterior intervalo sea cuanto más estrecho mejor.

En el histograma de la siguiente figura puede observarse la distribución del precio por metro cuadrado en un barrio de una pequeña capital española. La muestra está compuesta por 101 viviendas. A simple vista puede observarse que la distribución no se corresponde exactamente con la de una distribución normal, ya que la cola de la derecha está algo más extendida que la de la izquierda. Esto ocurre cuando en la muestra aparecen viviendas, como es el caso, que tienen un precio por metro cuadrado considerablemente superior al resto. Se trata de valores que podríamos considerar *outliers*, o fuera de lo normal.

La media del precio por metro cuadrado en esta muestra se calcularía como el promedio de los precios unitarios:

$$X = \frac{925 + 925{,}92 + \cdots + 2.127{,}66}{101} = 1.452{,}40€$$

Y la desviación típica se computaría como la raíz cuadrada de la varianza:

$$\sigma_X = \overline{\sigma_X^2} = \overline{\frac{\sum_{i=1}^{n} X_i - X^{\,2}}{n}}$$

$$= \overline{\frac{925 - 1.452{,}40^{\,2} + 925{,}92 - 1.452{,}40^{\,2} + \cdots + 925 - 1.452{,}40^{\,2}}{101}}$$

$$= 273{,}33$$

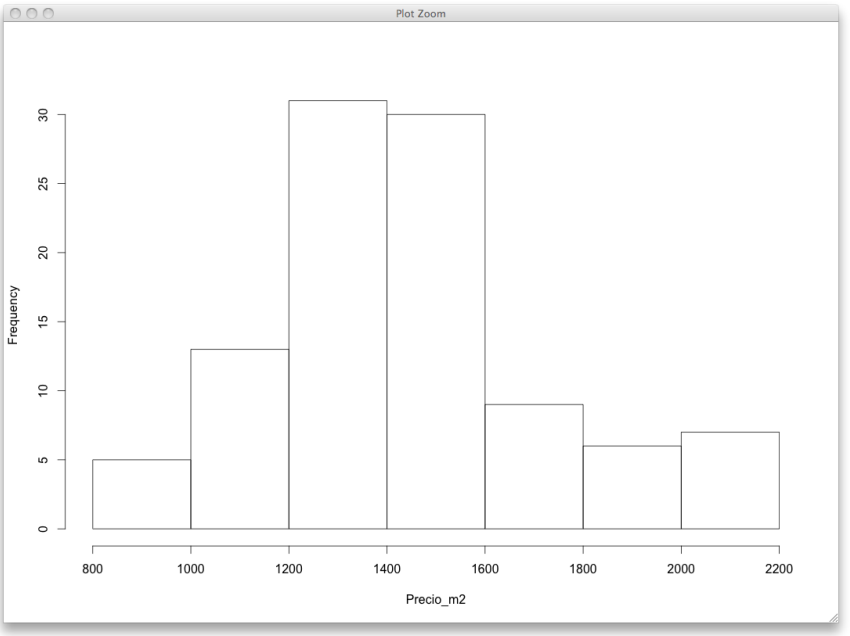


Figura 4. Histograma de frecuencias para la variable precio por metro cuadrado

Con lo que los intervalos de confianza se estimarían a partir de estos dos estadísticos de la siguiente forma:

Nivel de confianza del 90%:

$$X \pm 1{,}645\sigma_X =$$
$$1.452{,}40 - 1{,}645 \times 273{,}33 \quad 1.452{,}40 + 1{,}645 \times 273{,}33 =$$
$$1.002{,}77 \quad 1.902{,}03$$

Nivel de confianza del 95%:

$$X \pm 1{,}96\sigma_X =$$
$$1.452{,}40 - 1{,}96 \times 273{,}33 \quad 1.452{,}40 + 1{,}96 \times 273{,}33 =$$
$$916{,}67 \quad 1.988{,}13$$

Nivel de confianza del 99%:

$$X \pm 2{,}576\sigma_X =$$
$$1.452{,}40 - 2{,}576 \times 273{,}33 \quad 1.452{,}40 + 2{,}576 \times 273{,}33 =$$
$$748{,}30 \quad 2.156{,}50$$

Los precios en esta muestra se encuentran dentro del rango [925, 2.127,66], con lo que se trata de una muestra altamente heterogénea (el valor más grande dobla al valor más pequeño). Vemos cómo la parte inferior del rango es respetada en dos de las tres estimaciones realizadas (95% y 99%). Sin embargo, no ocurre lo mismo con la cota superior, que sólo está dentro del intervalo marcado por el 99%. La explicación está en el hecho ya comentado de la presencia de viviendas con precios por metro cuadrado excesivamente altos en relación con el resto, que podríamos considerar como no normales.

2.4 Relación entre variables: coeficiente de correlación

Hasta ahora únicamente se ha analizado el comportamiento de las variables de forma aislada. Los estadísticos de posición y de dispersión permiten conocer cuál es el comportamiento individual de las variables. Sin embargo, lo que resulta interesante en la mayor parte de los casos, y especialmente en el ámbito de la valoración, es conocer cuál es la relación entre dos o más variables. Esto es, cómo se relacionan por ejemplo el precio de una vivienda y su superficie. ¿A mayor superficie, mayor precio? O cómo se relacionan el número de goles marcado por un delantero con el valor del futbolista. ¿A mayor número de goles mayor valor? O el grado de relación entre el beneficio de una empresa con el valor de la misma. ¿A mayor beneficio, mayor valor? Además del signo, positivo o negativo, de estas relaciones, debemos determinar el grado de relación, traduciéndolo a un número que permita su cuantificación.

Antes de intentar cuantificar el grado de relación entre dos variables, examinemos el siguiente ejemplo gráfico donde se representa el precio de las viviendas en el eje de ordenadas frente a su superficie en el eje de abscisas. ¿Qué respondería usted si se le preguntara sobre la posible existencia de relación entre ambas variables?

Parece bastante claro, de la simple observación del gráfico, que ambas variables están relacionadas. Además, dicha relación es claramente positiva: a mayor superficie, mayor precio.

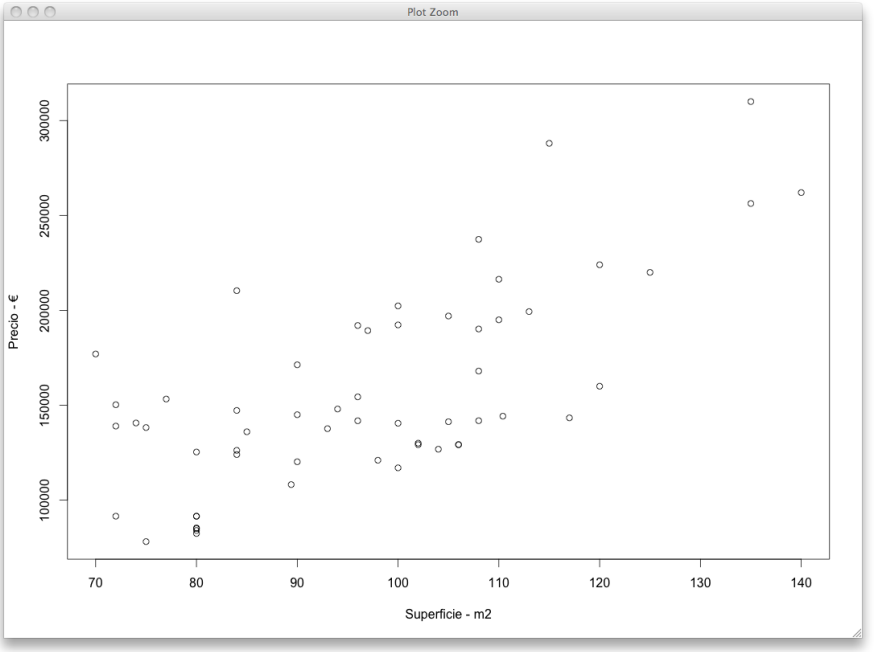


Figura 5. Gráfico de dispersión entre el precio y la superficie de las viviendas

También podemos encontrar ejemplos de relación negativa entre variables. En el caso de las viviendas, la antigüedad del edificio suele guardar una relación negativa con el precio, de forma que las viviendas de más reciente construcción suelen, en promedio, presentar mayores precios que las viviendas de mayor antigüedad. Y recalcamos la expresión "en promedio" ya que, evidentemente, existirán casos en los que esta relación no se aprecie. Puede ocurrir que una vivienda tenga un precio superior a otra aún siendo mucho más antigua, por ejemplo por encontrarse en una mejor situación en la ciudad, por tener más metros cuadrados de superficie útil, por tratarse de un edificio con especial valor arquitectónico o artístico, etc.

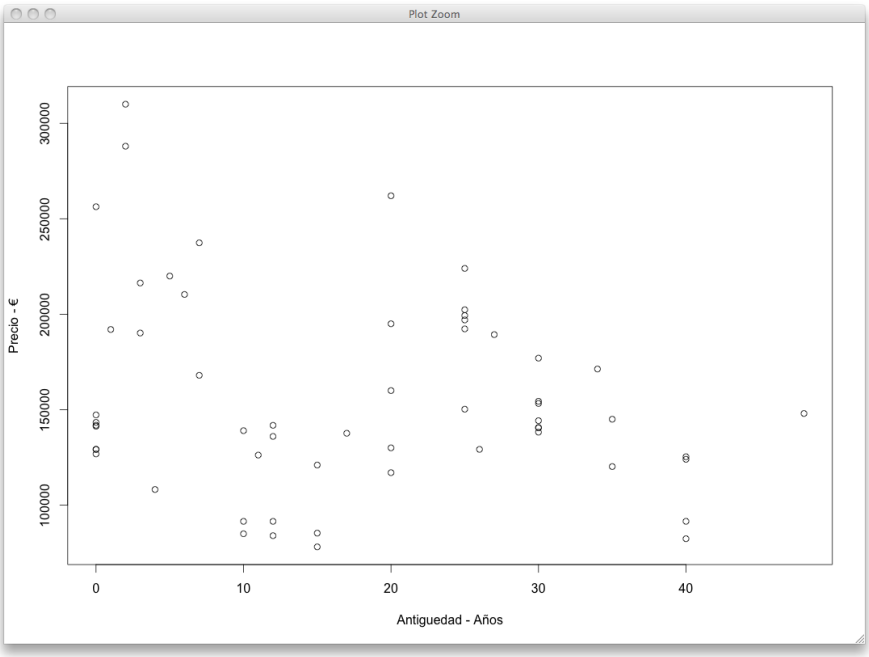


Figura 6. Gráfico de dispersión entre el precio y la antigüedad de las viviendas

Además de poder constatar de un modo gráfico el tipo de relación entre las variables, lo que para el valorador debe resultar realmente interesante es poder cuantificar el grado de relación entre las mismas. Y para ello puede utilizarse el coeficiente de correlación.

Sin embargo, veremos en los siguientes epígrafes que el cálculo de dicho coeficiente es distinto según tratemos con variables numéricas o con variables ordinales.

2.4.1 Coeficiente de correlación para variables numéricas

El coeficiente de correlación (ρ_{XY}) entre dos variables numéricas X e Y, también denominado coeficiente de correlación de Pearson, se obtiene a partir de la siguiente expresión:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

El numerador se corresponde con la covarianza (σ_{XY}) entre las variables X e Y, mientras que en el denominador aparece el producto de las desviaciones típicas de ambas variables.

La covarianza entre X e Y se calcula con la siguiente expresión:

$$\sigma_{XY} = \frac{\sum_{i=1}^{n} X_i - X \quad Y_i - Y}{n}$$

De manera que una forma alternativa de escribir la expresión del coeficiente de correlación entre X e Y sería:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\frac{\sum_{i=1}^{n} X_i - X \quad Y_i - Y}{n}}{\sqrt{\frac{\sum_{i=1}^{n} X_i - X \ ^2}{n}} \sqrt{\frac{\sum_{i=1}^{n} Y_i - Y \ ^2}{n}}} = \frac{\sum_{i=1}^{n} X_i - X \quad Y_i - Y}{\sqrt{\sum_{i=1}^{n} X_i - X \ ^2} \sqrt{\sum_{i=1}^{n} Y_i - Y \ ^2}}$$

A priori puede parecer complicado explicar la relación entre un par de variables a través de una fórmula como la anterior, que desde luego dista mucho de ser evidente en un primer vistazo. Conviene desglosar sus componentes para entender mejor el significado de la misma.

En el numerador del coeficiente de correlación, como se ha señalado anteriormente, aparece la covarianza σ_{XY}. Veamos cómo calcularla con un ejemplo ilustrativo para después interpretar el resultado.

Supongamos que tenemos 10 observaciones de dos variables X e Y, tal y como aparecen en la siguiente tabla. En la última fila de las columnas X e Y aparece el promedio de cada una de las variables para las 10 observaciones consideradas.

La columna $X_i - X$ recoge la diferencia entre cada observación de la variable X y la media X (ídem para $Y_i - Y$). Por último, la columna encabezada con $X_i - X \quad Y_i - Y$ recoge el producto de las dos columnas anteriores.

Si analizamos el signo de los valores en la última columna, lo primero que nos llama la atención es que casi todos son positivos (la única excepción se

produce con la observación número 7). A modo de explicación, examinemos lo ocurre con las observaciones 1 y 10.

En la primera observación, las variables toman los valores $X_1 = 2$ y $Y_1 = 4$. Ambos se sitúan por debajo de la media de sus respectivas variables: $X = 5{,}6$ y $Y = 6{,}7$. De esta forma, el resultado de $X_i - X \quad Y_i - Y$ es el producto de dos números negativos, que siempre será positivo. Esto mismo se repite para otras observaciones, luego parece que cuando una variable se sitúa por debajo de su media la otra también se coloca por debajo de su promedio.

En la última observación los valores son $X_{10} = 10$ y $Y_{10} = 11$. Ocurre entonces lo contrario que en el caso anterior: ambas observaciones están por encima de las medias de sus variables. De esta forma las columnas $X_i - X$ e $Y_i - Y$ toman valores positivos para la observación número 10, y su producto $X_i - X \quad Y_i - Y$ también. Como con la observación número 1, parece que X e Y vuelven a guardar una relación directa: cuando el valor de una variable se sitúa por encima de su media, la otra también aparece por encima de la suya.

Observación	X	Y	$X_i - X$	$Y_i - Y$	$X_i - X \quad Y_i - Y$
1	2	4	-3,6	-2,7	9,72
2	3	5	-2,6	-1,7	4,42
3	4	5	-1,6	-1,7	2,72
4	8	10	2,4	3,3	7,92
5	0	0	-5,6	-6,7	37,52
6	7	7	1,4	0,3	0,42
7	6	6	0,4	-0,7	-0,28
8	10	11	4,4	4,3	18,92
9	6	8	0,4	1,3	0,52
10	10	11	4,4	4,3	18,92
Media	*5,6*	*6,7*	*0*	*0*	*10,08*

En la última fila de la columna $X_i - X \quad Y_i - Y$ tenemos la media de los productos; esto es, el valor de la covarianza entre X e Y: 10,08.

En general, atendiendo al signo de la covarianza entre dos variables, diremos que:

- Si el signo es positivo, la relación entre las variables es positiva. Cuando una variable toma valores altos respecto de su promedio, la otra también lo hace. Y cuando una toma valores bajos respecto de

su promedio, la otra actúa de igual modo. ¿Qué signo cree que tomaría la covarianza entre el precio de una vivienda y su superficie?

- Si el signo es negativo, la relación entre las variables es negativa. Tal sería el caso de variables en las que cuando una toma valores altos respecto de su media, la otra los toma bajos. Y viceversa. Como ejemplo, podríamos considerar la covarianza entre el precio de una vivienda y su antigüedad, generalmente negativa.

El problema que subyace en esta clasificación es en qué grado se entiende por valores altos o bajos respecto de su promedio. Y ello debido a la diferente unidad en que se expresan las variables X e Y respecto de la covarianza entre ellas.

Supongamos que la unidad de medida de ambas variables fuera los $metros$, porque por ejemplo midieran una distancia entre dos localizaciones. En ese caso, la covarianza vendría expresada en $metros \times metros$ (o metros al cuadrado), lo que no es comparable con la unidad de media de las variables. En definitiva, nos encontramos con el mismo problema a la hora de medir la dispersión en una variable a través de la varianza.

Es por ello que la covarianza no se emplea habitualmente para medir el grado de relación entre variables, y sí el coeficiente de correlación que supera este inconveniente al venir acotado su rango de posibles valores. En concreto, el coeficiente de correlación sólo puede tomar valores entre -1 y +1:

$$\rho_{XY} \in -1, +1$$

de forma que:

- Cuanto más próximo está el coeficiente de correlación a +1, mayor es el grado de relación positiva entre las variables.
- Cuanto más próximo está el coeficiente de correlación a -1, mayor es el grado de relación negativa entre las variables.
- Valores del coeficiente próximos a 0, indican escasa o nula relación entre las variables.

Para el ejemplo anterior el coeficiente de correlación se calcularía como:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{10{,}08}{3{,}17 \times 3{,}29} = 0{,}968 = 96{,}8\%$$

Al ser un valor muy próximo a +1, podemos concluir que existe una fuerte relación positiva entre ambas variables, cosa que ya podíamos intuir también gráficamente:

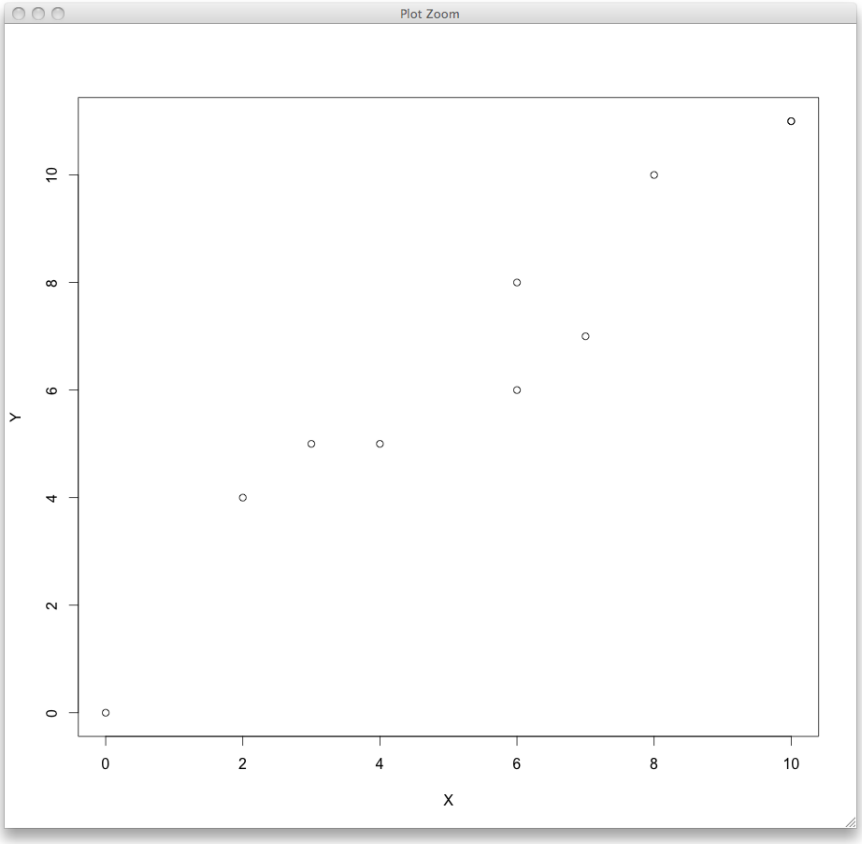


Figura 7. Gráfico de dispersión entre las variables X e Y

Veamos ahora un ejemplo más próximo al ámbito de la valoración. Supongamos que queremos conocer el grado de relación entre el precio de las viviendas y la superficie de las mismas. Ya pudimos constatar en una anterior figura que ambas variables están relacionadas de forma positiva entre sí: a mayor superficie, mayor precio, y viceversa.

En la siguiente tabla aparecen el precio y superficie de 30 viviendas. También la diferencia de cada variable respecto de sus medias, y el producto de dichas diferencias. Como en el ejemplo anterior, se ha reservado la última fila para la media de las 30 observaciones. De esta forma, la media de la última columna representa la covarianza entre las variables.

Vivienda	*Precio (Y)*	*Superficie (X)*	$Y_i - Y$	$X_i - X$	$X_i - X \quad Y_i - Y$
1	310.000	135	164.275,33	37,21	6.112.137,57
2	108.182	89,4	-37.542,67	-8,39	315.108,12
3	129.218	106	-16.506,67	8,21	-135.464,71
4	143.341	117	-2.383,67	19,21	-45.782,29
5	126.814	104	-18.910,67	6,21	-117.372,20
6	129.218	106	-16.506,67	8,21	-135.464,71
7	147.248	84	1.523,33	-13,79	-21.011,84
8	148.000	94	2.275,33	-3,79	-8.631,10
9	85.343	80	-60.381,67	-17,79	1.074.391,12
10	78.131	75	-67.593,67	-22,79	1.540.684,98
11	82.400	80	-63.324,67	-17,79	1.126.756,90
12	192.000	96	46.275,33	-1,79	-82.987,10
13	168.000	108	22.275,33	10,21	227.356,90
14	256.303	135	110.578,33	37,21	4.114.251,19
15	220.000	125	74.275,33	27,21	2.020.784,24
16	141.800	96	-3.924,67	-1,79	7.038,24
17	136.000	85	-9.724,67	-12,79	124.410,90
18	124.000	84	-21.724,67	-13,79	299.655,57
19	145.000	90	-724,67	-7,79	5.647,57
20	120.202	90	-25.522,67	-7,79	198.906,65
21	144.237	110,4	-1.487,67	12,61	-18.754,52
22	84.000	80	-61.724,67	-17,79	1.098.287,57
23	85.000	80	-60.724,67	-17,79	1.080.494,24
24	91.500	80	-54.224,67	-17,79	964.837,57
25	91.500	80	-54.224,67	-17,79	964.837,57
26	121.000	98	-24.724,67	0,21	-5.109,76
27	137.640	93	-8.084,67	-4,79	38.752,50
28	224.000	120	78.275,33	22,21	1.738.234,24
29	199.329	113	53.604,33	15,21	815.143,23
30	202.334	100	56.609,33	2,21	124.917,93
Media	*145.724,67*	*97,79*	*0,00*	*0,00*	*780.735,22*

A continuación aparecen los diferentes estadísticos necesarios para calcular el coeficiente de correlación entre el precio y la superficie de las viviendas:

$$\sigma_{XY} = \frac{\sum_{i=1}^{n} X_i - X \quad Y_i - Y}{n} = \frac{6.112.137{,}57 + 315.108{,}12 + \cdots + 124.917.93}{30} = 780.735{,}22$$

$$\sigma_Y = \sqrt{\frac{\sum_{i=1}^{n} Y_i - Y \ ^2}{n}} = \sqrt{\frac{164.275{,}33 \ ^2 + \ -37.542{,}67 \ ^2 + \cdots + \ 56.609{,}33 \ ^2}{30}} = 54.593{,}74$$

$$\sigma_X = \sqrt{\frac{\sum_{i=1}^{n} X_i - X \ ^2}{n}} = \sqrt{\frac{37{,}21 \ ^2 + \ -8{,}39 \ ^2 + \cdots + \ 2{,}21 \ ^2}{30}} = 16{,}58$$

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{780.735{,}22}{16{,}58 \times 54.593{,}74} = 0{,}863 = 86{,}3\%$$

El coeficiente de correlación indica una clara relación positiva entre ambas variables, próxima al 90%.

En la siguiente figura se ha representado el precio frente a la superficie de estas 30 viviendas, pudiendo constatar la clara relación positiva entre ambas variables. En el mismo gráfico aparece una recta que atraviesa la nube de puntos, y que identificaremos más adelante como la recta de regresión.

Un inconveniente del coeficiente de correlación es que puede verse gravemente afectado por la presencia de datos atípicos o anómalos (*outliers*). Este tipo de datos aparecen cuando por error se ha introducido algún valor que no se corresponde con el valor real, o se trata de una observación que presenta unas características muy diferentes a las del resto de la muestra.

Supongamos que en nuestro ejemplo anterior hubiéramos introducido, por error, una superficie de 80 metros cuadrados para la vivienda número 1, que en realidad tiene una superficie de 135 metros cuadrados. El nuevo gráfico de dispersión aparece en la figura 9, donde destaca por anómala la posición de esta vivienda.

Si recalculáramos el coeficiente de correlación, habría pasado del 86,3% inicial a un valor muy inferior: 57,2%. Y el único cambio producido ha sido el de una observación de la muestra.

De ahí que, en ocasiones, podamos obtener coeficientes de correlación muy bajos entre variables que a priori parezcan guardar una relación significativa. Debemos limpiar nuestra muestra de elementos que afecten negativamente a la representatividad de los datos o, como veremos más adelante, justificar esas anomalías por las especiales características del inmueble.

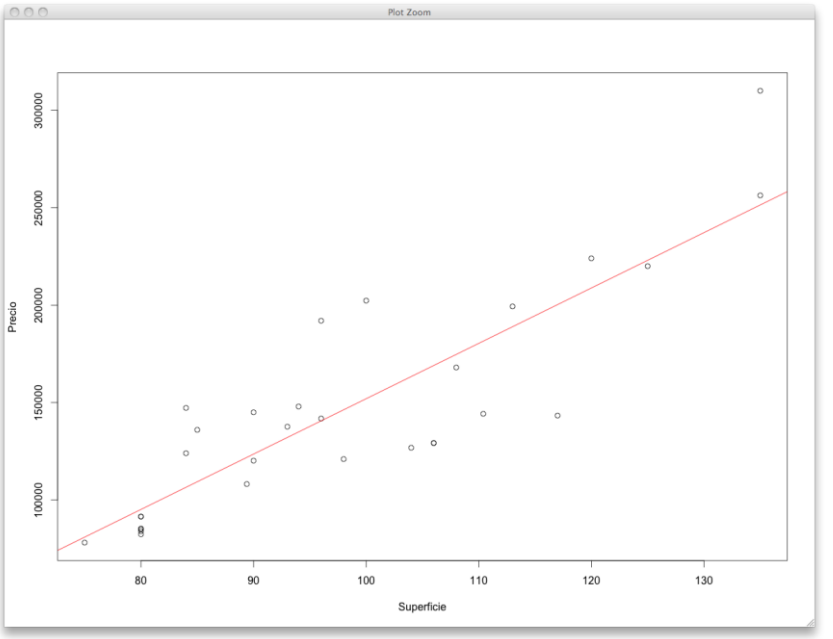


Figura 8. Gráfico de dispersión entre el precio y la superficie de las viviendas

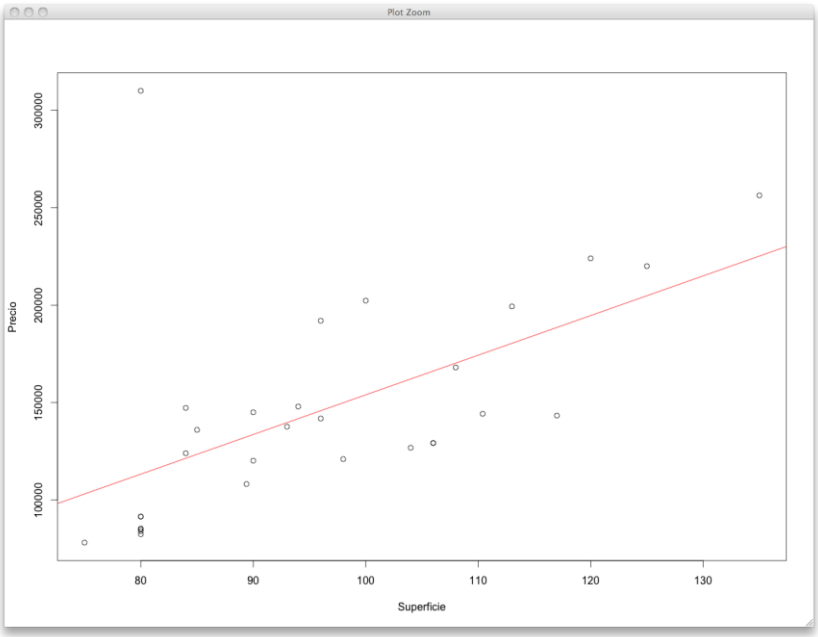


Figura 9. Gráfico de dispersión entre el precio y la superficie de las viviendas, donde una de ellas tiene una superficie errónea

2.4.2 Coeficiente de correlación para variables ordinales

El coeficiente de correlación introducido en el anterior epígrafe no resulta adecuado cuando al menos una de las variables es ordinal. En estos casos resulta recomendable utilizar el coeficiente de correlación de Spearman.

Veamos en primer lugar la definición de este coeficiente de correlación junto con algún ejemplo, para posteriormente comparar los resultados con los obtenidos mediante el coeficiente de correlación de Pearson.

La expresión del coeficiente de correlación de Spearman entre dos variables X e Y es la siguiente:

$$\rho_{XY} = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n\ n^2 - 1}$$

El parámetro n se corresponde con el número de observaciones, como en casos anteriores, mientras que d mide la diferencia en posición entre las observaciones.

Veamos un ejemplo ilustrativo utilizando nuevamente una pequeña muestra de 10 viviendas. Supongamos que estamos interesados en conocer si existe relación entre la superficie (variable numérica) y el número de dormitorios (variable ordinal). En la segunda y tercera columna de la tabla aparecen las variables para las que queremos calcular la correlación de Spearman. Vemos cómo el número de dormitorios fluctúa entre 1 y 3, pudiendo ser considerada una variable ordinal. En las dos siguientes columnas tenemos el ranking o posicionamiento ocupado por cada una de las observaciones en las dos variables.

Para el caso del precio la ordenación es sencilla. Ocupa el número 1 la vivienda de menor precio (120.202€), seguida en la posición 2 por la vivienda con el segundo precio más bajo (124.000€). Y así hasta llegar a la vivienda de mayor precio (256.303€) que ocupa la posición décima.

Para el ranking de la variable número de dormitorios seguimos el mismo procedimiento. Sin embargo, ahora tenemos observaciones que comparten una misma posición en el ranking. Por ejemplo, las viviendas con un sólo dormitorio ocupan las 3 primeras posiciones. En este caso, el cálculo de su ranking se haría promediando las posiciones que ocupan, para que de esta forma todas compartan el mismo ranking:

$$d_{1,2,3} = \frac{d_1 + d_2 + d_3}{3} = \frac{1 + 2 + 3}{3} = 2$$

Algo similar ocurre con las viviendas de 2 dormitorios, que ocuparían las posiciones 4, 5 y 6 en el ranking. Para calcular el posicionamiento utilizamos la expresión:

$$d_{4,5,6} = \frac{d_4 + d_5 + d_6}{3} = \frac{4 + 5 + 6}{3} = 5$$

Y por último el caso de las viviendas con 3 dormitorios:

$$d_{7,8,9,10} = \frac{d_7 + d_8 + d_9 + d_{10}}{4} = \frac{7 + 8 + 9 + 10}{4} = 8{,}5$$

Observación	*Precio (Y)*	*Número dormitorios (X)*	d_Y	d_X	$d_Y - d_X\ ^2$
1	120.202	2	1	5	16
2	124.000	1	2	2	0
3	136.000	1	3	2	1
4	141.800	1	4	2	4
5	144.237	3	5	8,5	12,25
6	145.000	2	6	5	1
7	168.000	3	7	8,5	2,25
8	192.000	2	8	5	9
9	220.000	3	9	8,5	0,25
10	256.303	3	10	8,5	2,25

A partir de estos valores, obtenemos el coeficiente de correlación de Spearman:

$$\rho_{XY} = 1 - \frac{6\ \sum_{i=1}^{n} d_i^2}{n\ n^2 - 1} = 1 - \frac{6\ \sum_{i=1}^{n}\ 16^2 + 0^2 + \cdots + 2{,}25^2}{10\ \ 10^2 - 1} = 0{,}709$$
$$= 70{,}9\%$$

que aún indicando el mismo tipo de relación positiva entre ambas variables que el coeficiente de correlación de Pearson (63,0%), difiere claramente en su valor numérico.

Probablemente el lector se habrá percatado de una característica muy relevante al calcular el coeficiente de correlación de Spearman cuando se combina una variable ordinal con una numérica continua, como es el precio. A la hora de realizar el ranking de una variable continua (o asimilable a una continua), la posición ocupada por cada uno de los valores no tiene en cuenta las diferencia en valor absoluto entre los valores originales. Por ejemplo, para el caso de la vivienda con menor precio (120.202€), su posición sería la misma para cualquier valor inferior a la cantidad actual. Es decir, podría aparecer con un valor de 100.000€, 80.000€, o cualquiera inferior a estos: su posición seguiría siendo la primera, y por lo tanto la correlación de Spearman no cambiaría. Sí que lo haría, y según el caso de manera muy significativa, la correlación de Pearson que, como vimos en un apartado anterior, es muy sensible a la presencia de datos anómalos.

Capítulo 3. El modelo de regresión simple

3.1 Introducción

El modelo de regresión fue ideado por el polifacético investigador Sir Francis Galton, quien publicó su trabajo *Natural Inheritance* en el año 1889. En su manual analizó la relación existente entre la altura física de padres e hijos, evidenciando que existía una relación positiva entre ambos: los padres altos solían tener hijos altos, mientras que los padres de menor estatura también solían tener hijos con una altura por debajo de la media. Sin embargo, también pudo observar que en ambos casos existía lo que denominó una regresión a la media, de forma que los hijos de padres con estatura superior a la media heredaban una altura también superior a la media, pero más próxima al promedio general que la de sus padres. De igual forma, los hijos de padres de estatura inferior a la media también eran bajos, pero menos que sus padres.

Para ligar ambas variables, altura de padres e hijos, ideó el análisis de regresión. En su versión más sencilla, conocida como análisis de regresión simple, se relaciona linealmente una variable Y con una variable X, de forma que Y recibe la denominación de variable dependiente o endógena, y X es denominada variable independiente, exógena o explicativa:

$$Y = f\ X\ = a + bX$$

Diremos entonces que Y es función de X, o que Y depende de X. En su forma funcional más sencilla el modelo de regresión simple adopta la forma de una recta, y viene totalmente determinado por la constante o término independiente a y la pendiente b asociada a la variable explicativa X (función afín).

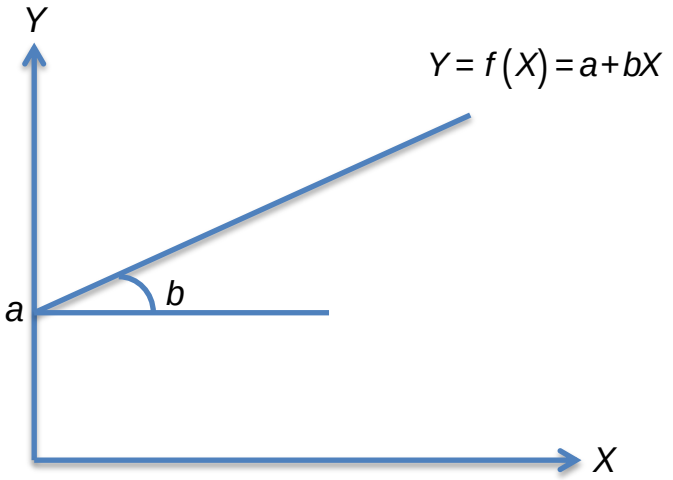


Figura 10. Recta del modelo de regresión lineal simple

La interpretación de los coeficientes a y b es sencilla. Supongamos que queremos establecer el precio (Y) en función de la superficie (X). El coeficiente a nos informa del precio que tendría una vivienda con una superficie de 0 metros cuadrados. Este dato no es muy significativo, ya que difícilmente vamos a encontrar una vivienda con esas características. El coeficiente b indica la variación del precio por cada variación unitaria de la superficie. Esto es, si la superficie de la vivienda a valorar se incrementa en un metro cuadrado, el precio estimado para la misma aumentará en b unidades monetarias. Y al contrario, si la superficie disminuye en un metro cuadrado, el precio también se reduce en b unidades monetarias.

3.2 Estimación del modelo de regresión lineal simple

Como veremos a continuación, los coeficientes a y b del modelo de regresión son calculados a partir de los valores de X e Y. Para ilustrar la forma en que se obtienen dichos coeficientes utilizaremos un ejemplo en el que se explica el precio de las viviendas a partir de su superficie. Supongamos que trabajamos únicamente con 10 observaciones (un número a todas luces insuficiente, pero que sólo persigue un fin estrictamente académico), y que las viviendas se distribuyen tal y como aparece en la siguiente figura.

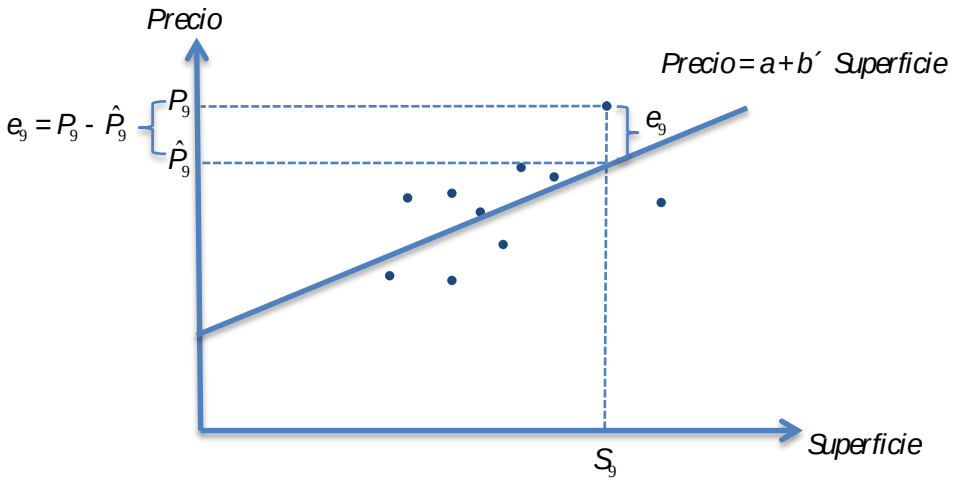


Figura 11. Recta regresión entre Precio y Superficie

En el gráfico aparecen en conjunto de 10 observaciones con sus respectivos precios y superficies. Al realizar la regresión, se obtiene la recta:

$$Precio = a + bSuperficie$$

que servirá para realizar las predicciones del precio de las viviendas según su superficie. Esto significa que, cuando queramos valorar una vivienda con superficie S_{11}, su valor pronosticado o estimado, P_{11}, será:

$$P_{11} = a + bS_{11}$$

Como se aprecia en la propia figura, podríamos obtener el precio pronosticado de las 10 observaciones a partir de la recta de regresión. Pongamos como ejemplo el caso de la vivienda número 9, con precio P_9 y superficie S_9. Al aplicarle la expresión de la recta, su precio pronosticado sería:

$$P_9 = a + bS_9$$

que no coincide con el precio real u observado P_9. De hecho, se observa con claridad que el precio estimado, que reposa sobre la recta de regresión, es inferior al precio observado. La diferencia entre ambos precios recibe el nombre de residuo u error:

$$e_9 = P_9 - P_9$$

Lógicamente, cualquier tasador estaría encantado de que todos los residuos de su muestra fueran 0. Eso significaría que los precios estimados coinciden con los observados, y nuestras funciones de valoración serían perfectas –al menos para las observaciones contenidas en la muestra-. Pero, como ya puede suponerse, la realidad dista mucho de coincidir con esta situación ideal.

El lector podría dibujar el resto de errores cometidos por la función de regresión en la estimación del precio. Para ello simplemente debería proyectar verticalmente cada uno de los puntos sobre la recta de regresión, y la distancia entre las observaciones originales y sus proyecciones sobre la recta serían los correspondientes residuos.

¿Cómo se obtiene entonces la función de regresión? O equivalentemente, ¿cómo se calculan los parámetros a y b? Es evidente que el objetivo final debe ser que los puntos (viviendas) estén lo más próximos a la recta de regresión. Esto indicaría que los errores cometidos en la estimación son pequeños, y por lo tanto nuestro modelo de valoración bastante ajustado.

Pues bien, la "distancia" de los puntos a la recta, que recuerden es la medida que queremos minimizar, se calculará no como los simples residuos sino como los residuos al cuadrado. De ahí la denominación de regresión por mínimos cuadráticos o mínimo-cuadrática con la que en ocasiones se refiere al análisis de regresión. De esta forma, los parámetros a y b son aquellos que hacen que la suma de los residuos al cuadrado sea lo más pequeña posible. El modelo de regresión puede representarse entonces como un modelo de optimización, con una función objetivo y tantas restricciones como observaciones tengamos en la muestra:

$$\underset{i=1}{\overset{n}{Min}} \; e_i^2$$

$$e_i = P_i - P_i = P_i - \; a + bS_i$$

O volviendo a nuestro modelo más general en el que intentamos explicar una variable Y a partir de otra variable X:

$$Min \sum_{i=1}^{n} e_i^2$$

$$e_i = Y_i - (a + bX_i)$$

La solución del anterior problema viene dada por las siguientes expresiones de b y a:

$$b = \frac{\sigma_{XY}}{\sigma_X^2}$$

$$a = Y - bX$$

donde σ_{XY} representa la covarianza entre las variables X e Y, σ_X^2 es la varianza de la variable X, y X e Y son los valores medios de las variables X e Y, respectivamente.

Veamos la aplicación de los resultados anteriores sobre un pequeño ejemplo.

3.3 Ejemplo ilustrativo de modelo de regresión lineal simple entre Precio y Superficie

En la siguiente tabla aparecen el precio y superficie de 10 viviendas, que nos servirán de ejemplo para ilustrar el cálculo de la recta de regresión. La primera columna sirve para enumerar el conjunto de viviendas, mientras que las dos siguientes recogen el precio y superficie de las mismas. En la última fila aparece el promedio de las variables.

Vivienda	Precio	Superficie	Precio estimado	Residuo
1	234.500	86	206.920,69	27.579,31
2	264.000	115	264.741,75	-741,75
3	204.000	95	224.865,15	-20.865,15
4	333.000	101	236.828,13	96.171,87
5	244.500	105,8	246.398,52	-1.898,52
6	330.000	127	288.667,70	41.332,30
7	124.000	66	167.044,09	-43.044,09
8	225.000	116	266.735,58	-41.735,58
9	246.000	130	294.649,19	-48.649,19
10	286.500	130	294.649,19	-8.149,19
Media	249.150	107,18	249.150,00	0,00

El modelo de valoración que buscamos relaciona el precio con la superficie a través de la regresión simple:

$$Precio = a + bSuperficie$$

Para el cálculo del coeficiente b, se necesita previamente estimar la covarianza entre el precio y la superficie, así como la varianza de la superficie:

$$\sigma_{XY} = \sigma_{SP} = \sum_{i=1}^{10} S_i - S \quad P_i - P \quad n = 708.011,82$$

$$\sigma_X^2 = \sigma_S^2 = \sum_{i=1}^{10} S_i - S \ ^2 \quad n = 355,10$$

$$b = \frac{\sigma_{XY}}{\sigma_X^2} = \frac{\sigma_{SP}}{\sigma_S^2} = \frac{708.011,82}{355,10} = 1.993,83 \ \frac{€}{m^2}$$

$$a = Y - bX = P - bS = 249.150 - 1.933{,}83 \times 107{,}18 = 35.451{,}34€$$

Observemos que la unidad del coeficiente b es euros por metro, mientras que la unidad del coeficiente a son euros. De esta forma, la función de regresión quedaría como:

$$Precio = a + bSuperficie = 35.451{,}34 + 1.993{,}83 \times Superficie$$

A partir de la expresión anterior podemos estimar los precios pronosticados, y calcular los residuos como la diferencia entre los precios observados y los precios estimados. En la cuarta columna de la tabla aparece el precio estimado para cada vivienda, y en la quinta el residuo. Vemos cómo el mayor error, en valor absoluto, lo cometemos con la vivienda número 4: su precio real u observado P_4 es de 333.000, mientras que el precio estimado por la función de regresión es de 236.828,13:

$$P_4 = a + bS_4 = 35.451{,}34 + 1.993{,}83 \times 101 = 236.828{,}13€$$

Esto hace que el residuo, o error cometido en la predicción, sea de 96.171,87:

$$e_4 = P_4 - P_4 = 333.000 - 236.828{,}13 = 96.171{,}87€$$

Una importante característica de los modelos de regresión es que la media de los precios observados coincide con la media de los precios pronosticados. En nuestro ejemplo, el precio medio de las 10 viviendas es de 249.150€, que se corresponde de manera exacta con el promedio de los 10 precios pronosticados (última fila de la tabla, columna Precio estimado). Esto es así porque los residuos se compensan unos con otros, de forma que la suma de los mismos es siempre cero.

Recuerdo que en una ocasión realizamos un trabajo de valoración para una de las sociedades de tasación más importantes del país que consistía en actualizar el valor de todos los inmuebles que determinada entidad financiera tenía en su Balance (valoración masiva). Una primera aproximación fue realizar un modelo de regresión (obviamente de mayor complejidad que el que de momento hemos descrito). El equipo técnico de tasación de la sociedad quiso

revisar los resultados, y para ello seleccionó una amplia muestra de viviendas a la que aplicó el modelo que habíamos desarrollado. Puesto que la entidad financiera estaba interesada en conocer el valor conjunto de todo su parque de viviendas, y no en el valor concreto de cada una de ellas, para evaluar nuestro trabajo midieron las diferencias entre los precios observados y los pronosticados por nuestro modelo. Al sumar las desviaciones (residuos), se dieron cuenta de que el valor era prácticamente cero, por lo que concluyeron que el modelo era excelente.

La explicación de este hecho, como ya habrá imaginado el lector, está en la propiedad que sobre los residuos acabamos de describir. Si los técnicos hubieran escogido toda la muestra para realizar su análisis, en lugar de una submuestra, ¡la suma de los residuos habría sido exactamente cero!

Eso no implica que el modelo sea excelente, ni mucho menos. Más adelante veremos cómo analizar la bondad y adecuación de nuestro modelo de regresión.

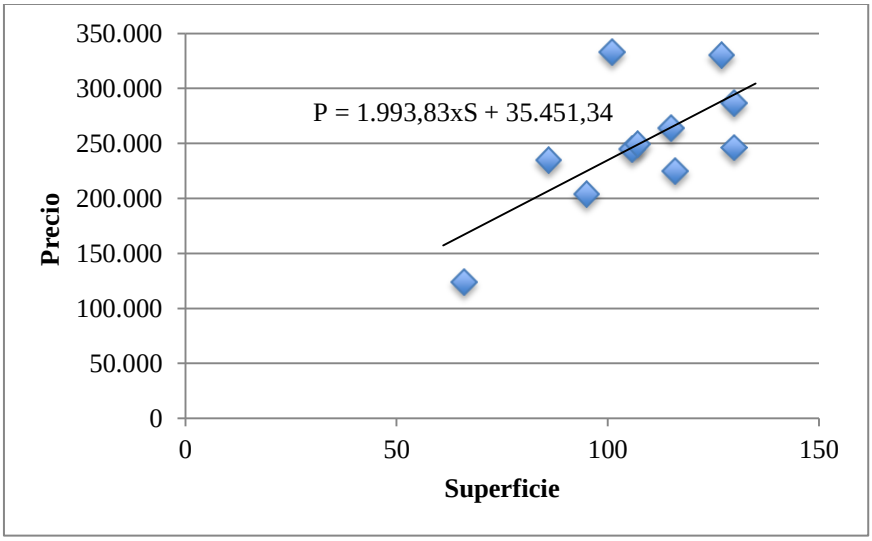


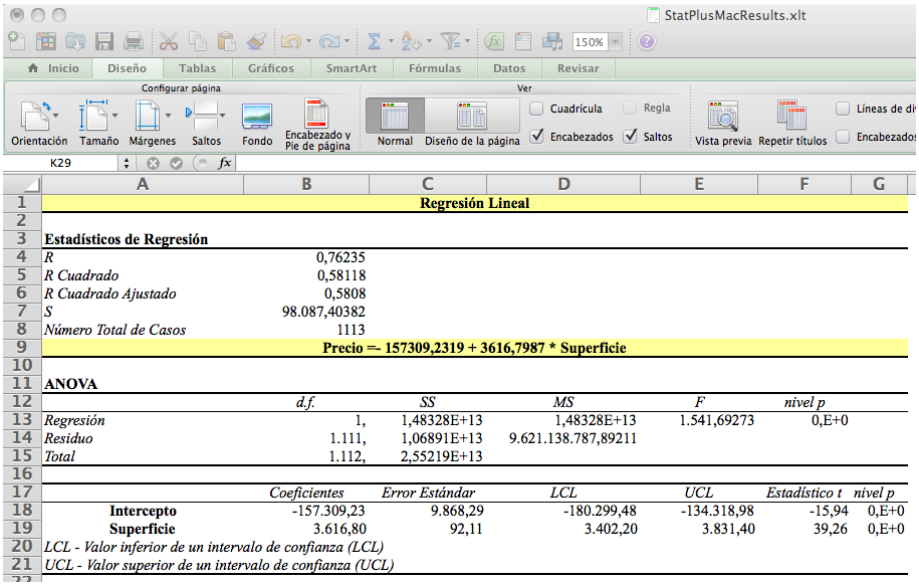
Figura 12. Recta regresión entre Precio y Superficie, junto con la estimación de los coeficientes a y b

Cierto. Seguro que el lector se pregunta qué ocurrió con nuestro modelo de valoración y con la sociedad de tasación que nos contrató. El modelo que proporcionamos a la sociedad tenía una bondad en el ajuste próxima al 90% (veremos más adelante que en valoración inmobiliaria éste es un porcentaje

próximo al ideal). Y respecto de la sociedad de tasación, ya se imaginará qué ocurrió con ella tras el pinchazo de la burbuja del ladrillo en España.

3.4 Significación estadística del modelo de regresión

Hasta el momento únicamente nos hemos dedicado a estimar los coeficientes de la recta de regresión, pues con ellos podemos aplicar nuestra función de valoración sobre un activo y predecir su precio. Pero hemos descuidado una posibilidad: que la función de valoración no sea significativa. Sería lo normal si, por ejemplo, intentáramos estimar el precio a partir del número de bombillas que hay en la vivienda. Desde luego podríamos calcular los coeficientes *a* y *b*, pero no deberíamos plantearnos su utilización en la práctica profesional.



	A	B	C	D	E	F	G
1	**Regresión Lineal**						
2							
3	**Estadísticos de Regresión**						
4	*R*	0,76235					
5	*R Cuadrado*	0,58118					
6	*R Cuadrado Ajustado*	0,5808					
7	*S*	98.087,40382					
8	*Número Total de Casos*	1113					
9	**Precio =- 157309,2319 + 3616,7987 * Superficie**						
10							
11	**ANOVA**						
12		*d.f.*	*SS*	*MS*	*F*	*nivel p*	
13	*Regresión*	1,	1,48328E+13	1,48328E+13	1.541,69273	0,E+0	
14	*Residuo*	1.111,	1,06891E+13	9.621.138.787,89211			
15	*Total*	1.112,	2,55219E+13				
16							
17		*Coeficientes*	*Error Estándar*	*LCL*	*UCL*	*Estadístico t*	*nivel p*
18	**Intercepto**	-157.309,23	9.868,29	-180.299,48	-134.318,98	-15,94	0,E+0
19	**Superficie**	3.616,80	92,11	3.402,20	3.831,40	39,26	0,E+0
20	*LCL - Valor inferior de un intervalo de confianza (LCL)*						
21	*UCL - Valor superior de un intervalo de confianza (UCL)*						

Figura 13. Resultado del análisis de regresión entre el Precio y la Superficie en la hoja de cálculo Microsoft Excel

La significación (estadística) de un análisis de regresión se lleva a cabo a través de la tabla ANOVA (Análisis de la varianza, del inglés *ANalysis Of VAriance*). Supongamos que tenemos una muestra de 1.113 viviendas, todas ellas de una misma ciudad, y de las que conocemos su precio de traspaso reciente junto con algunas características que nos parece pueden ser relevantes a la hora de estimar su precio. Entre estas variables incluimos la superfi-

cie, y realizando la regresión entre Precio y Superficie obtenemos los siguientes resultados[2] (figura 13).

En el estudio de la significación estadística del modelo de regresión únicamente nos fijaremos, de momento, en la tabla en cuyo encabezado aparece la palabra ANOVA. Aunque podríamos detallar el significado de sus filas y columnas, el resultado que realmente nos interesa desde un punto de vista práctico es el que aparece en la última columna: nivel p. Este coeficiente nos informa sobre la significación estadística del modelo, de forma que si el valor que observamos está por debajo del 5% diremos que el modelo es en su conjunto significativo desde un punto de vista estadístico. En el caso del ejemplo se puede apreciar un valor 0 (a falta de decimales), con lo que podemos afirmar que estadísticamente el modelo sí es significativo.

Si el valor hubiera estado por encima del umbral del 5%, tendríamos que haber descartado el modelo por no poder considerarlo significativo.

¿Significa esto que si el modelo obtiene un valor p (o *p-value*) por debajo del 5% podemos aplicar con total tranquilidad nuestro modelo en una valoración? La respuesta es no. El valor p de la tabla ANOVA nos permite descartar inicialmente cualquier modelo que no llegue al 5%, pero sin asumir que el modelo es válido para su aplicación en la práctica profesional. Para eso tendremos que examinar otros parámetros.

¿Y por qué un p valor del 5%? En la inmensa mayoría de modelos estadísticos las conclusiones siempre tienen asociado un nivel de confianza estadística determinado. Recuerde el lector que esto es estadística, y no matemáticas. En matemáticas 2 + 2 siempre son 4. En estadística diríamos que 2 + 2 tienen una *alta probabilidad* de ser 4. El nivel del 5% en el valor p está asociado, precisamente, a la alta probabilidad con la que queremos dotar a nuestros modelos estadísticos. Siempre que infiramos un modelo a partir de unos datos, las conclusiones se asociarán a un nivel de confianza determinado: el 95% es el valor más extendido en la práctica. De esta forma, el valor p se obtiene como la diferencia entre la unidad y el nivel de confianza que hemos seleccionado:

$$valor\ p = 1 - nivel\ de\ confianza$$

Si queremos que nuestros modelos tengan un nivel de confianza elevado, de al menos el 95%, entonces el máximo valor p admisible será del 5%.

[2] Estos valores se han obtenido con la hoja de cálculo Microsoft Excel.

3.5 Significación estadística de los coeficientes en el modelo de regresión

Supongamos que nuestro modelo ha resultado ser, en su conjunto, estadísticamente significativo. Que lo sea en su conjunto no significa que lo sea individualmente para cada uno de los coeficientes estimados.

En nuestra versión más sencilla del modelo de regresión únicamente estimamos dos coeficientes: el intercepto *a* y la pendiente *b*. Pues bien, puede ser que siendo el modelo estadísticamente significativo en su conjunto, alguno de estos dos coeficientes no lo sean.

Pero antes de indicar en dónde debemos fijarnos para evaluar la significación estadística de los coeficientes, ¿qué significa que un coeficiente no sea estadísticamente significativo?

En el análisis de regresión se lleva a cabo un test de significación estadística sobre cada uno de los coeficientes estimados. En concreto, se analiza si dichos coeficientes son o no distintos de cero. Por lo tanto, es como si el técnico estadístico de turno se preguntara: "bien, he obtenido dos coeficientes que me permiten aplicar esta función sobre datos fuera de la muestra, y espero que fuera de ella. Pero los valores de los coeficientes, ¿podrían considerarse distintos de cero si hubiera aplicado el análisis sobre otra muestra similar?". Desde luego el lector puede pensar que dicho problema es bien sencillo de resolver incluso para un estudiante de primaria: si un coeficiente tiene un valor de 10, pongamos como ejemplo, pues claramente es distinto de cero. Pero, ¿y si el coeficiente tiene un valor de 1? ¿O de 0,1? ¿Y si fuera 0,000000001? La respuesta más probable que daría a todos los casos es "sí, el coeficiente es distinto de cero. Porque o se es cero, o se es distinto de cero, no hay más posibilidades". Pues bien, ciertamente eso es lo que ocurre en matemáticas, pero recuerde que estamos en estadística.

El problema viene dado porque cuando estimamos una función de regresión lo hacemos a partir de una muestra de datos, y no de toda la población. Podemos estar seguros de que si ampliáramos nuestra muestra a toda la población los coeficientes obtenidos sufrirían variaciones respecto de los iniciales. De ahí que si obtenemos una constante *a* con un valor de 5 en nuestra muestra, puede que al modificar la muestra su valor estuviera mucho más próximo a cero, o que fuera incluso negativo. De alguna forma tenemos que poder asegurar que el valor obtenido es distinto de cero. Y muy importante, para un nivel de confianza determinado: pongamos el 95%.

Volviendo al ejemplo de la figura 13, se han estimado dos coeficientes:

$$Precio = -157.309{,}23 + 3.616{,}80 x Superficie$$

El signo negativo de la constante le tiene que haber llamado la atención, pero ahora no lo vamos a comentar.

Parece que ambos coeficientes son claramente distintos de cero, que nadie podría dudar de que se trata de valores diferentes a cero. Para constatarlo desde un punto de vista estadístico utilizamos la columna del valor p –sí, como hicimos al analizar el modelo en su conjunto-, que aparece en último lugar de la tabla de coeficientes. Y en ambos casos el valor p es cero, luego podemos concluir que los valores obtenidos para el intercepto y la pendiente son estadísticamente distintos de cero (con un nivel de confianza del 95%).

También deben comentarse los valores del Valor inferior de un intervalo de confianza y del Valor superior de un intervalo de confianza. Nos informan del rango de valores en los que podrían variar los coeficientes para un nivel de confianza determinado. Por así decirlo, es como si en el análisis nos estuviera diciendo: "En promedio, el coeficiente asociado a la superficie es de 3.616,80. Pero si repitiéramos el análisis sobre otras muestras similares a la actual, es posible que el coeficiente fluctuara entre un valor mínimo de 3.402,20 y un valor máximo de 3.831,40". El nivel de confianza para dichos extremos es del 95%, con lo que sólo en un 5% de los casos podríamos encontrarnos, en promedio, con coeficientes asociados a la superficie fuera de dicho intervalo.

Observe como en el intervalo [3.402,20 ‘ 3.831,40] no está el cero. Eso significa que podemos descartar que el coeficiente asociado a la superficie sea nulo.

Porque, ¿qué ocurre si llegamos a la conclusión de que el coeficiente que hemos obtenido no es estadísticamente distinto de cero? Pregúntese para qué quiere un coeficiente cero en su función de valoración. Si ni suma ni resta, entonces elimínelo y eso que se ahorra.

Supongamos que el coeficiente asociado a la superficie no hubiera sido estadísticamente distinto de cero, para un nivel de confianza del 95%. Significaría entonces que la superficie de la vivienda no influye en su precio, y por lo tanto tendríamos que descartar dicho coeficiente de la función de valoración. Mantenerlo no mejoraría nada, puesto que se ha demostrado que no influye en los precios de las viviendas. Piense que el objetivo del valorador es, finalmente, obtener modelos que sean lo más parsimoniosos posible. Esto es, que estén compuestos por el menor número de variables.

Cuanto mayor sea la sencillez del modelo, más atractivo será para el valorador. Podemos apuntar dos razones para ello: la primera, que un modelo sencillo siempre es preferible a uno complicado. Así ocurre en los modelos que se aplican en física, química o matemáticas. La explicación más sencilla de las cosas es siempre preferible a la explicación más compleja. Y en segundo lugar, como valorador preferirá un modelo que emplee el menor número de variables explicativas por el elevado coste que supone la obtención de las mismas. No es lo mismo construir un modelo de valoración que utilice 5 variables, que otro que emplee 20. En el segundo caso el valorador tendrá que hacer un sobreesfuerzo por medir 20 variables en cada una de las viviendas que compongan la muestra, mientras que en el primer caso con 5 únicas variables podrá desarrollar un modelo completo de valoración.

Más adelante, cuando tratemos el modelo de regresión múltiple con más de una variable explicativa, analizaremos el modo en que debemos eliminar aquellas variables que puedan haber resultados no significativas, o irrelevantes, en nuestro modelo de valoración.

Para completar este epígrafe, volvamos a un punto que habíamos resaltado anteriormente: el signo negativo del intercepto o constante. Dicho coeficiente ha obtenido un valor de -157.309,23. Esto puede hacer pensar al valorador que alguna vivienda podría ser valorada ¡negativamente!. Bueno, por muchas crisis del ladrillo que se sufra, quizá esto pueda antojarse algo objetivamente difícil de ser alcanzado.

Es posible que alguna vivienda se pudiera valorar con un precio negativo, en concreto todas aquellas que tuvieran una superficie inferior a los 43 metros cuadrados. Eso no implica que nuestro modelo de valoración sea un mal modelo de valoración. De hecho, en la muestra compuesta por 1.112 viviendas, ninguna tenía una superficie inferior a los 45 metros cuadrados, con lo que al menos en la muestra nunca alcanzaríamos una valoración negativa. Lo que en ningún caso deberíamos pretender es valorar una vivienda que tuviera menos de esos 45 metros cuadrados determinados por la vivienda de menor superficie en la muestra, como veremos más adelante.

Tampoco debemos pensar en los -157.309,23 como un precio "base". Dicho coeficiente debe interpretarse como el precio que tendría una vivienda con una superficie de 0 metros cuadrados, cosa que en la práctica sabemos no tiene sentido, así que no le demos más vueltas.

Otro error muy común es el de querer eliminar aquellos coeficientes que el valorador considera que no tienen sentido, como podría ser en este caso el del intercepto. Claramente su eliminación por este motivo sería un error, ya que el coeficiente ha resultado ser estadísticamente significativo. Si se eliminara, estaríamos empeorando nuestra función de valoración. Pensemos que prescindir del intercepto es, en realidad, añadir una restricción al modelo: que la

recta de regresión tiene que pasar por el origen del eje de coordenadas. Cualquier modelo al que se añada una restricción adicional ha de obtener, necesariamente, una solución peor (salvo que la restricción sea redundante).

En la siguiente figura se comparan las rectas de regresión con intercepto positivo (línea continua) y sin intercepto (línea discontinua). La primera se ha obtenido mediante regresión lineal, y en la segunda se ha restringido la anterior añadiendo una constante nula. Claramente en el segundo caso se empeora la solución, ya que se puede observar cómo algunos puntos quedan más alejados de la recta de regresión que en el primer caso.

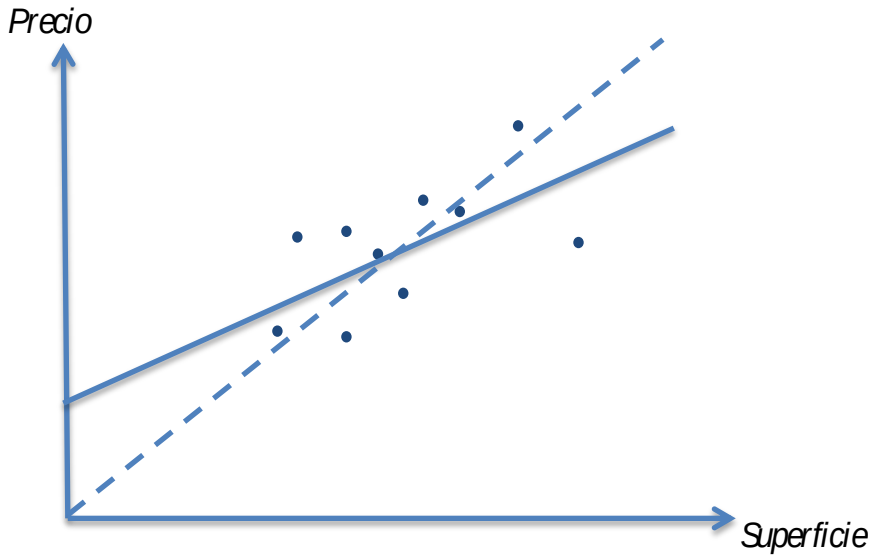


Figura 14. Comparativa de una recta de regresión con intercepto frente a otra recta de regresión sin intercepto

3.6 Qué hacer si la constante no es estadísticamente distinta de cero

Si en nuestro modelo de regresión el coeficiente de la constante no ha resultado ser estadísticamente significativo (valor p por encima del 5%), entonces lo razonable es plantear su eliminación del modelo. Esto es lo que debemos hacer con cualquier coeficiente que resulte ser no significativo, porque de esta manera conseguimos que nuestro modelo sea más parsimonioso. Y esta fue la respuesta que, en mis tiempos de estudiante, me dio mi profesora de modelos multivariantes al plantearle la cuestión.

Sin embargo, y siendo cierto lo anterior cuando analizamos cualquier coeficiente asociado a una variable explicativa, eliminar la constante en la función de valoración no plantea ninguna ventaja comparativa. Ciertamente, si podemos valorar una vivienda prescindiendo de una variable significa que el valorador tendrá que recopilar menos información en sus valoraciones. Y esto es claramente positivo para él. Pero, ¿qué ocurre si prescindimos de la constante? ¿También será un ahorro de tiempo para el valorador? Pues la respuesta no es tan clara, ya que incluir la constante en el modelo no supone tener que recopilar más información sobre las viviendas para poder obtener una función de valoración. Por lo tanto incluirla o no es, en principio, indiferente para el valorador.

Por otra parte, si el modelo nos dice que no es estadísticamente distinta de cero y decidimos eliminarla, los precios que estimemos sin ella no serán muy distintos a los que obtendríamos si la incluyéramos en la función de valoración. Pero eso sí, si la descartamos estaremos incluyendo una restricción en el modelo: que la recta de regresión deba pasar por el origen de coordenadas.

En resumen, eliminar una constante que no es estadísticamente significativa 1) no supone un ahorro en coste y tiempo para el valorador, y 2) añadimos una restricción adicional sobre nuestro modelo, lo que no hará precisamente que mejoremos los resultados aunque dicha restricción se haya demostrado redundante.

Además, cuando en un epígrafe posterior analicemos la bondad de los modelos de regresión mediante el coeficiente R cuadrado, veremos que eliminar la constante hace que el coeficiente R cuadrado no sea comparable con el de otros modelos con constante, por lo que definitivamente les propongo que cuando se encuentren con una constante estadísticamente no significativa no la eliminen de su modelo de valoración.

3.7 El estadístico R cuadrado: cómo analizar la bondad de mi modelo de regresión

En los epígrafes anteriores hemos analizado cuándo un modelo es significativo en su conjunto, y cuándo lo son de manera individual los coeficientes a y b de la recta de regresión. A continuación analizaremos el estadístico R cuadrado (R^2), que permite estudiar la bondad en el ajuste de un modelo de regresión y conocer cuál es su capacidad explicativa. De esta forma, podremos saber cuán bien o mal explica el precio una función de valoración, y si tiene sentido aplicarla en la práctica profesional.

Para entender cómo se construye dicho estadístico, es bueno comenzar entendiendo qué modelo de regresión aplicaría usted en una situación extrema. Imagine que desea obtener un modelo de regresión que explique el precio,

¡sin contar con ninguna variable explicativa! Ciertamente una situación complicada. La única información de la que dispone es el propio precio de las viviendas en su muestra, y con dicha información debe construir un modelo de valoración.

En la siguiente tabla tenemos de nuevo el ejemplo de 10 viviendas, de las que de momento sólo vamos a utilizar el precio (segunda columna). Si usted tuviera que estimar el precio de una undécima vivienda, muy probablemente el valor que daría en su pronóstico sería el promedio de estas 10. Es decir, 249.150€. Implícitamente lo que usted estaría haciendo es aplicar el modelo ingenuo o modelo *naive*. Dicho modelo asume que cuando sólo se conoce la información de la variable dependiente, y no existe ninguna independiente, entonces la mejor predicción posible para una nueva observación es la media muestral.

Vivienda	Precio	Modelo ingenuo		Modelo superficie		
		Precio estimado	Residuo	Superficie	Precio estimado	Residuo
1	234.500	249.150	-14.650	86	206.920,69	27.579,31
2	264.000	249.150	14.850	115	264.741,75	-741,75
3	204.000	249.150	-45.150	95	224.865,15	-20.865,15
4	333.000	249.150	83.850	101	236.828,13	96.171,87
5	244.500	249.150	-4.650	105,8	246.398,52	-1.898,52
6	330.000	249.150	80.850	127	288.667,70	41.332,30
7	124.000	249.150	-125.150	66	167.044,09	-43.044,09
8	225.000	249.150	-24.150	116	266.735,58	-41.735,58
9	246.000	249.150	-3.150	130	294.649,19	-48.649,19
10	286.500	249.150	37.350	130	294.649,19	-8.149,19
Media	249.150	249.150	0	107,18	249.150,00	0

En la columna "Modelo ingenuo" aparece el precio que estimaríamos para cada una de las 10 viviendas que componen la muestra: 249.150€. El residuo se tomaría como diferencia entre el precio observado y el precio estimado. La suma de los errores o residuos es cero, como ocurría en el caso del modelo de regresión simple. Si sumamos el cuadrado de los residuos obtenemos el siguiente valor:

$$-14.650\ ^2 + 14.850^2 + \cdots + 37.350^2 = 33.713.525.000$$

Las últimas columnas de la tabla recogen el resultado de aplicar el modelo de regresión simple entre el precio y la superficie. En este caso, la suma de los residuos al cuadrado es:

$$27.579{,}31^2 + \ -741{,}75\ ^2 + \cdots + \ -8.149{,}19\ ^2 = 18.185.320.306$$

Se comprueba como este valor es inferior al obtenido por el del modelo ingenuo. Por lo tanto, la desviación típica de los residuos también será inferior:

$$\sigma_{residuos\ mod\ ingenuo} = 58.063{,}35$$

$$\sigma_{residuos\ mod\ regresión\ simple} = 42.644{,}25$$

Al tener menor dispersión los residuos del modelo de regresión simple, podemos deducir que dicho modelo explica mejor el precio que el modelo ingenuo. Esto de por sí ya era evidente, ya que el modelo ingenuo realizaba la predicción de los precios a partir del propio precio, sin tener en cuenta ninguna otra variable exógena.

Pues bien, el estadístico R^2 se construye a partir de las dos sumas de cuadrados que hemos descrito:

$$R^2 = 1 - \frac{Suma\ cuadrados\ residuos\ mod\ regresión\ simple}{Suma\ cuadrados\ residuos\ mod\ ingenuo}$$

Para el caso de nuestro ejemplo, su valor se calcularía como:

$$R^2 = 1 - \frac{18.185.320.306}{33.713.525.000} = 0{,}4606 = 46{,}06\%$$

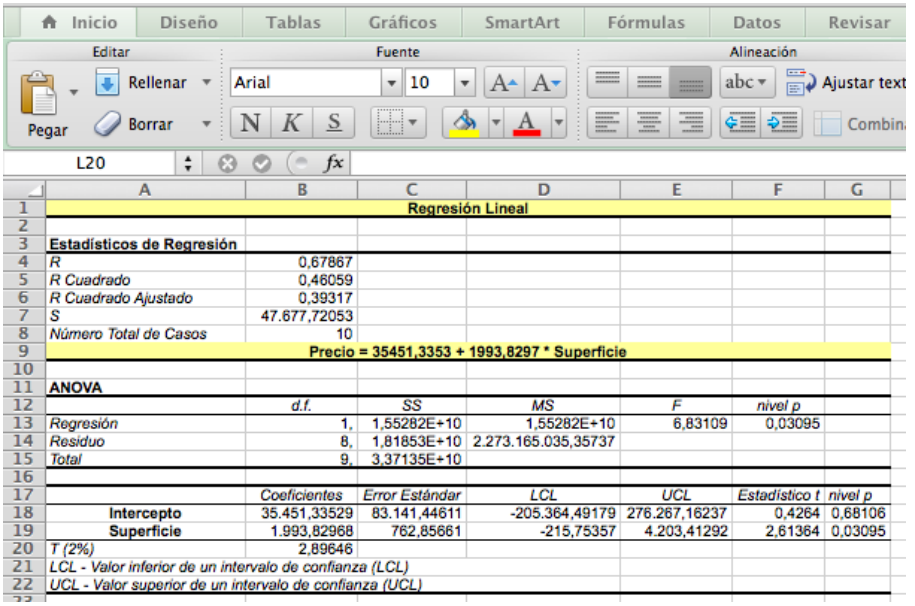
Obsérvese como cuanto más pequeña sea la suma de cuadrados de los residuos del modelo de regresión simple, mayor será el estadístico R^2. La situación ideal sería que esta suma de cuadrados fuera cero, lo que indicaría que

los precios estimados coinciden con los precios observados, y el estadístico tomaría valor 1.

Es fácil demostrar que el estadístico R^2 viene acotado entre los valores 0 y 1, de forma que cuánto más se aproxime a la unidad mayor capacidad explicativa tiene el modelo, mientras que un valor próximo a cero indicará que el modelo tiene escasa capacidad explicativa (y probablemente no será significativo en la tabla ANOVA):

$$0 \leq R^2 \leq 1$$

Una propiedad interesante de los modelos de regresión simple es que el estadístico R^2 coincide con el cuadrado del coeficiente de correlación entre la variable dependiente y la independiente. En el modelo entre el precio y la superficie, la correlación era de 67,87%. El cuadrado de este valor se corresponde con el valor del R^2: 46,06%. Por lo tanto, interesa regresar el precio frente a variables con las que esté altamente correlacionada. Cuanto mayor sea la correlación, positiva o negativa, mayor será el valor del estadístico R^2 y mejor ajustado nuestro modelo de valoración.



	A	B	C	D	E	F	G
1	Regresión Lineal						
2							
3	Estadísticos de Regresión						
4	*R*	0,67867					
5	*R Cuadrado*	0,46059					
6	*R Cuadrado Ajustado*	0,39317					
7	*S*	47.677,72053					
8	*Número Total de Casos*	10					
9	Precio = 35451,3353 + 1993,8297 * Superficie						
10							
11	ANOVA						
12		*d.f.*	*SS*	*MS*	*F*	*nivel p*	
13	*Regresión*	1,	1,55282E+10	1,55282E+10	6,83109	0,03095	
14	*Residuo*	8,	1,81853E+10	2.273.165.035,35737			
15	*Total*	9,	3,37135E+10				
16							
17		*Coeficientes*	*Error Estándar*	*LCL*	*UCL*	*Estadístico t*	*nivel p*
18	**Intercepto**	35.451,33529	83.141,44611	-205.364,49179	276.267,16237	0,4264	0,68106
19	**Superficie**	1.993,82968	762,85661	-215,75357	4.203,41292	2,61364	0,03095
20	*T (2%)*	2,89646					
21	*LCL - Valor inferior de un intervalo de confianza (LCL)*						
22	*UCL - Valor superior de un intervalo de confianza (UCL)*						

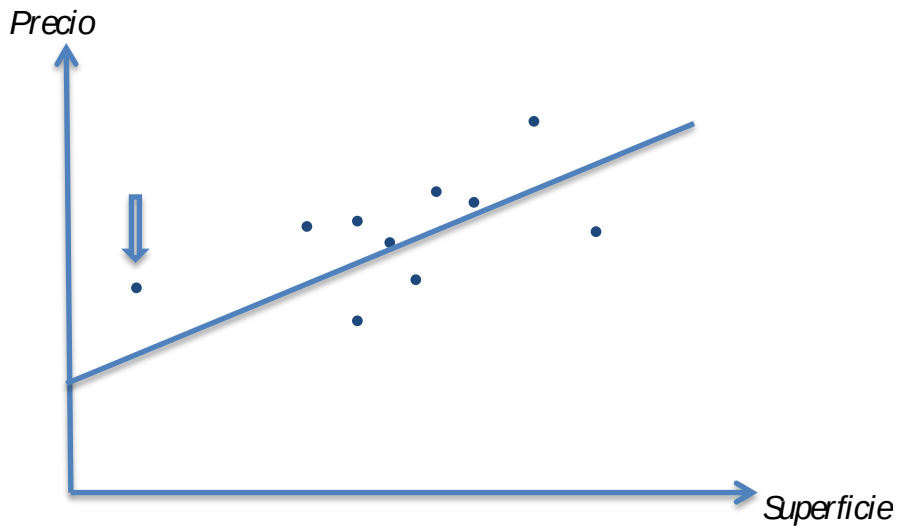
Figura 15. Modelo de regresión con la hoja de cálculo Excel de Microsoft. Detalle del estadístico R cuadrado, y el coeficiente de correlación (R)

3.8 Cómo influyen las observaciones atípicas o *outliers* en el análisis de regresión

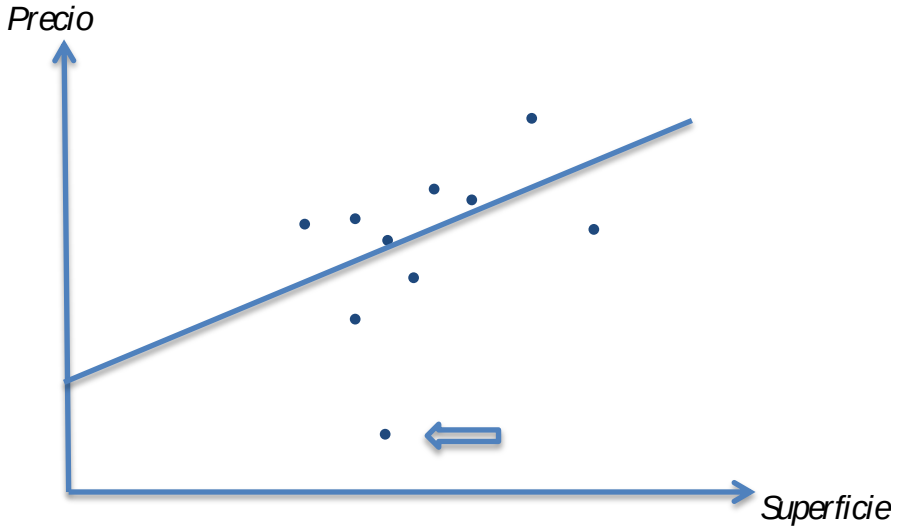
Los *outliers* u observaciones atípicas son aquellas que no representan ni son representadas por la tendencia central de los datos y que, por lo tanto, se separan mucho del comportamiento medio del resto de observaciones. Aparecen cuando se combinan observaciones muy heterogéneas entre sí, y para las que es difícil encontrar nexos en común. Se consideran observaciones atípicas las que reúnen una o más de las siguientes características: 1) tienen un valor de la variable dependiente muy alejado del promedio, 2) tienen un valor de la variable independiente muy alejado del promedio, 3) la relación entre la variable dependiente e independiente está muy alejada de la observada en el resto de la muestra.

Volviendo al ámbito de la valoración donde intentamos estimar el precio de las viviendas a partir de su superficie, una observación muestral se considerará atípica si tiene un precio excesivamente alto o bajo respecto del resto de viviendas, o si tiene una superficie muy superior o inferior al resto, o si el precio por metro cuadrado es muy diferente del resto de observaciones.

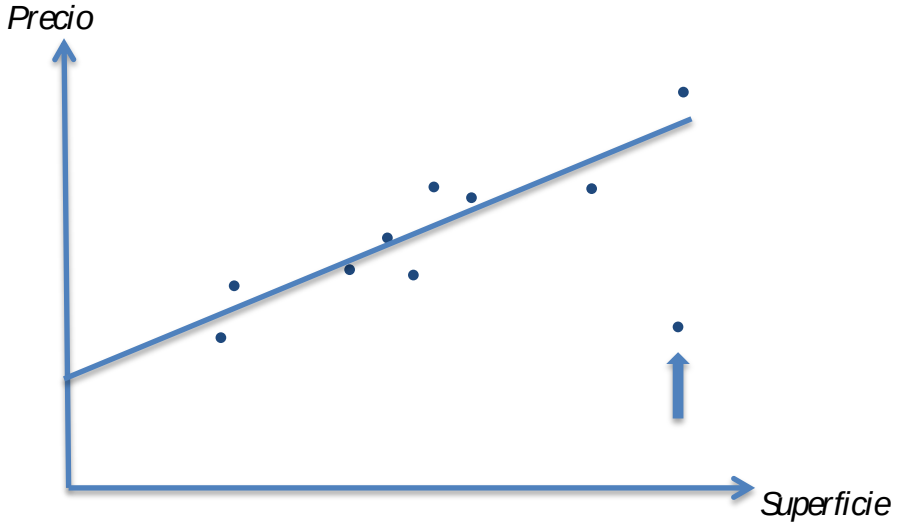
En las siguientes figuras aparecen las 3 circunstancias antes mencionadas, y por las que podríamos considerar atípica las observaciones destacadas sobre una flecha.



Panel A. Observación atípica en superficie, muy por debajo de la superficie en promedio del resto de viviendas



Panel B. Observación atípica en precio, muy por debajo del precio en promedio del resto de viviendas



Panel C. Observación atípica en su relación precio-superficie. El ratio es muy diferente respecto del observado en el resto de viviendas

Figura 16. Ejemplos de observaciones atípicas entre variable dependiente e independiente

La cuestión que debemos afrontar es qué hacer cuando nos encontramos en nuestra muestra observaciones como las señaladas en la anterior figura. Claramente su presencia va a distorsionar los resultados en muchos casos, puesto que afectará tanto a la pendiente como al intercepto de la recta de regresión. También pueden deteriorar el valor del estadístico R^2, con lo que un buen modelo teórico puede venirse al traste al observar un R^2 bajo como consecuencia de la presencia de observaciones atípicas. Caben dos posibilidades:

1) Incluir más variables explicativas. Si con la única variable independiente del modelo de regresión simple no somos capaces de explicar la presencia de esas observaciones atípicas, tendremos que recurrir a la regresión múltiple e incluir en el modelo de regresión más de una variable independiente. En nuestro ejemplo, observamos viviendas cuyo precio no se corresponde con la superficie. Pero es posible que tal disparidad sí se pueda explicar por la presencia de otra variable. Por ejemplo, si encontramos una vivienda con un precio excesivamente alto para su superficie, puede que se trate de una vivienda excepcionalmente bien localizada, o con calidades muy superiores a la media de la muestra. Estas características sí justificarían un precio superior para la superficie dada.
2) Eliminar las observaciones atípicas. Puesto que su inclusión afecta negativamente al modelo de regresión obtenido, distorsionando su R^2, la pendiente y/o el intercepto, la solución más sencilla pasa por eliminarlas de la muestra y repetir el análisis sin ellas. En primer lugar, veamos qué criterio seguir para su eliminación y después examinemos el efecto que tiene sobre el modelo resultante, para así poder discriminar cuándo es conveniente suprimirlas y cuándo no.

El criterio que más habitualmente se sigue para identificar las observaciones atípicas no es el visual, como hicimos anteriormente en una figura. Piénsese que su detección mediante un gráfico puede resultar sencilla en un plano, pero más difícil de llevar a cabo en 3 dimensiones o más.

Para la identificación de las observaciones anómalas se suele emplear el residuo estandarizado obtenido mediante el modelo de regresión. Básicamente, se procede de la siguiente forma: 1) se aplica el modelo de regresión sobre el conjunto de las observaciones de la muestra, 2) se calcula el residuo para cada observación, como diferencia entre el precio observado y el precio estimado por el modelo de regresión, 3) finalmente se estandarizan los residuos para que el conjunto tenga media cero y desviación típica unitaria.

La estandarización de una variable, en este caso los residuos, se obtiene restando de cada uno de sus valores la media, para después dividir por la desviación típica:

$$e_i^{estand.} = \frac{e_i - e}{\sigma_e}$$

donde $e_i^{estand.}$ representa el residuo estandarizado de la observación i-ésima, e_i es el residuo de la observación i-ésima, e es el promedio de los residuos y σ_e es la desviación típica de los residuos. Como vimos en un apartado anterior, los residuos obtenidos mediante un modelo de regresión tienen la propiedad de tener media cero, con lo que la anterior expresión se simplifica:

$$e_i^{estand.} = \frac{e_i}{\sigma_e}$$

En la siguiente tabla aparecen los residuos estandarizados para el ejemplo de regresión entre precio y superficie.

Vivienda	Precio	Superficie	Precio estimado mediante regresión	Residuo	Residuo estandarizado
1	234.500	86	206.920,69	27.579,31	0,6467
2	264.000	115	264.741,75	-741,75	-0,0174
3	204.000	95	224.865,15	-20.865,15	-0,4893
4	333.000	101	236.828,13	96.171,87	2,2552
5	244.500	105,8	246.398,52	-1.898,52	-0,0445
6	330.000	127	288.667,70	41.332,30	0,9692
7	124.000	66	167.044,09	-43.044,09	-1,0094
8	225.000	116	266.735,58	-41.735,58	-0,9787
9	246.000	130	294.649,19	-48.649,19	-1,1408
10	286.500	130	294.649,19	-8.149,19	-0,1911
Media	249.150	107,18	249.150,00	0,00	0,0000

Veamos, a modo de ejemplo, cómo se ha calculado el residuo estandarizado en la vivienda número 1. En primer lugar calculamos la desviación típica de los residuos:

$$\sigma_e = \sqrt{\frac{27.579{,}31\ ^2 + \ -741{,}75\ ^2 + \cdots + \ -8.149{,}19\ ^2}{10}} = 42.644{,}25$$

Con lo que el residuo de la primera observación se calcula como:

$$e_1^{estand.} = \frac{e_1}{\sigma_e} = \frac{27.579{,}31}{42.644{,}25} = 0{,}6467$$

El lector podrá comprobar cómo la media de los residuos estandarizados es cero, mientras que su desviación típica es uno:

$$e^{estand.} = \frac{0{,}6467 + \ -0{,}0174 \ + \cdots + \ -0{,}1911}{10} = 0$$

$$\sigma_e^{estand.} = \sqrt{\frac{0{,}6467\ ^2 + \ -0{,}0174\ ^2 + \cdots + \ -0{,}1911\ ^2}{10}} = 1$$

Pues bien, una vez ya sabemos cómo calcular el residuo estandarizado, procedemos ahora a indicar el criterio que habitualmente se emplea para eliminar observaciones anómalas, a partir del mencionado residuo estandarizado.

En un modelo de regresión en el que intervienen variables normales, como suponemos ocurre en nuestro ejemplo con el precio y la superficie, los residuos también seguirán una distribución normal. Al estandarizarlos, seguirán teniendo una distribución normal, pero ahora con media cero y desviación típica uno. Sabemos entonces que el 99% de los residuos estarán dentro del rango marcado por su media (cero) más/menos 3 veces aproximadamente su desviación típica (uno). Pues bien, lo que resulta muy habitual en la práctica es encontrar analistas que eliminan todas aquellas observaciones que tienen un residuo estandarizado por encima de +3, o por debajo de -3. Se justifican en que 1) dichas observaciones presentan un precio estimado muy diferente del observado, y 2) representan un porcentaje muy pequeño de la muestra como para afectar al tamaño muestral y a la validez de los resultados.

En el ejemplo de la tabla no encontramos ninguna vivienda que tenga un residuo con valor absoluto superior a 3. La vivienda con mayor valor absoluto en su residuo es la número 4, con un residuo estandarizado de 2,25. Únicamente el 5% de las observaciones tendrían un residuo estandarizado con valor absoluto superior a 2, en promedio.

Si analizamos la vivienda número 4, su precio observado es de 333.000€, mientras que el valor pronosticado para su superficie de 101 metros cuadrados es de 236.828,13€. Esto es, el error cometido en la predicción es de 96.171,87. Resulta ser la vivienda en la que mayor error absoluto de predicción se comete, ya que su precio es muy alto para la superficie que tiene. Ciertamente no será habitual encontrar viviendas con un error de predicción tan elevado. Sólo un 5% de las viviendas, una de cada 20, tendrán un residuo en promedio tan grande o más que el de la vivienda número 4.

De esta forma, un analista podría plantearse eliminar dicha observación del modelo de regresión, ya que su influencia en los parámetros a y b de la recta de regresión debe ser grande, y es evidente que no está adecuadamente representada por la relación general entre precio y superficie observada en el resto de casos.

Si repitiéramos el análisis de regresión considerando 9 en lugar de 10 observaciones (eliminando la vivienda número 4), la nueva recta de regresión tendría la siguiente forma:

$$Precio = 6.329{,}00 + 2.164{,}75 \times Superficie$$

Es decir, hemos pasado de valorar el metro cuadrado de superficie de 1.993,83€/m2 a 2.164,75€/m2; un cambio bastante notable. Pero la diferencia aún es mayor si comparamos las constantes. Antes era de 35.451,34€ y ahora toma un valor de 6.329,00€. Se demuestra que la influencia de la observación número 4 en la recta de regresión era muy grande[3]. De hecho, el estadístico R^2 también habría aumentado significativamente: de 46,06% a 69,90%.

En resumen, la bondad del ajuste ha aumentado de manera clara, los coeficientes a y b de la recta de regresión también se han visto modificados de manera significativa; y todo ello renunciando a solo una observación.

[3] De hecho, existe un estadístico que permite calcular la influencia de cada observación sobre la recta de regresión. No veremos su cómputo por alejarse del objetivo general de este manual, y ser un tema de interés específico para estadísticos.

Parecen todo ventajas. Pero también debemos considerar una serie de inconvenientes:

1) Hay una vivienda que ahora no se tiene en cuenta en el análisis. Y usted, como valorador, tendrá que valorar cualquier tipo de viviendas, no sólo las que se ajusten a determinado patrón estadístico. ¿Qué hará si tiene que valorar una vivienda que tiene un residuo estandarizado de valor 2 o 3? ¿Le dirá a su cliente que se busque otro valorador?
2) Si vuelve a calcular los residuos estandarizados de su nuevo modelo, es muy posible que se encuentre con que alguna observación ahora tiene un residuo estandarizado por encima del valor 2, o incluso del valor 3. ¿Qué hará? ¿Eliminarlo igualmente? Puede que llegue el momento en que su muestra se vea mermada considerablemente.

Ciertamente no soy muy partidario de eliminar datos de una muestra. Puede que la vivienda número 4, que parece tener un precio excesivo para sus metros cuadrados, pueda justificarse por alguna característica que no hemos tenido en cuenta. Es que, de hecho, ¡sólo hemos considerado la superficie! Veremos en el siguiente tema que al considerar más de una variable explicativa algunas viviendas que nos parecían *outliers*, ahora tienen residuos estandarizados de lo más "normales".

En la práctica, sólo elimino aquellas observaciones que tienen un residuo estandarizado por encima de 4 o 5. Encontrarse con viviendas de esas características es ciertamente muy poco probable, con lo que parece plausible que simplemente se trate de datos mal recopilados o introducidos en el ordenador por error. Por lo general el porcentaje de viviendas que presentan residuos estandarizados tan extremos no suele superar el 0,5% de la muestra, y eso en el peor de los casos. Así que mi muestra no se ve diezmada de manera significativa por eliminar este tipo de casos extremos.

Piense además en lo siguiente. Las viviendas con residuos estandarizado de valor 3 o más, tienen que representar en promedio el 1% de la muestra, por definición de distribución normal. Así pues, si su muestra está compuesta por 1.000 viviendas, en promedio tendrá 10 viviendas con un residuo estandarizado que cumpla esa condición. Eso será lo normal. Luego no podemos definir como observaciones anómalas aquellas que entran dentro de lo "normal". ¿Por qué suprimirlas entonces?

3.9 El problema de la heterocedasticidad (y cómo relajarlo)

La heterocedasticidad se relaciona con aquellas situaciones en las que la variabilidad de una variable no es constante con respecto a otra, sino que fluctúa según los valores que toma la segunda variable. En primer lugar veremos cómo detectarla gráficamente, para luego examinar los problemas que se

pueden derivar de la misma, y cómo intentar mitigarlos (que no eliminarlos totalmente).

En la siguiente figura aparece representada una instancia del problema de heterocedasticidad. Tenemos el precio y superficie de un conjunto de viviendas y, englobando a las mismas, hemos delimitado un par de rectas imaginarias. Vemos cómo la distancia entre las rectas aumenta conforme lo hace la superficie. Para viviendas de hasta 80 metros cuadrados, la dispersión en los precios es baja. Sin embargo, para viviendas de más de 140 metros cuadrados nos encontramos viviendas con precios muy diferentes entre sí: podemos ver viviendas entorno a los 200.000€, pero también de más de 1.000.000€. Así pues, parece evidente que la variabilidad del precio no es constante con la superficie, sino que depende de ésta: acabamos de constatar visualmente la presencia de heterocedasticidad.

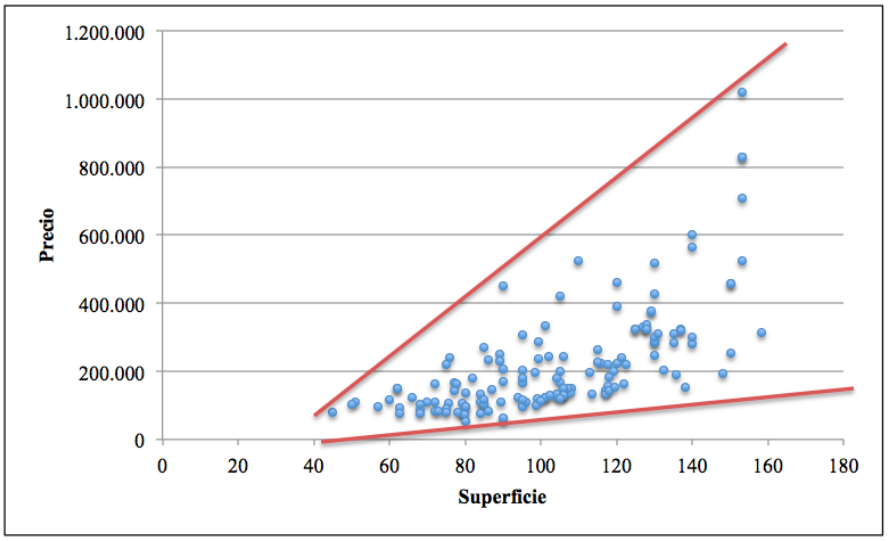


Figura 17. Ejemplo de heterocedasticidad

Para que no tuviéramos heterocedasticidad los precios tendrían que distribuirse de forma constante sobre la superficie, de manera que si trazáramos dos rectas que los envolviera, éstas fueran prácticamente paralelas.

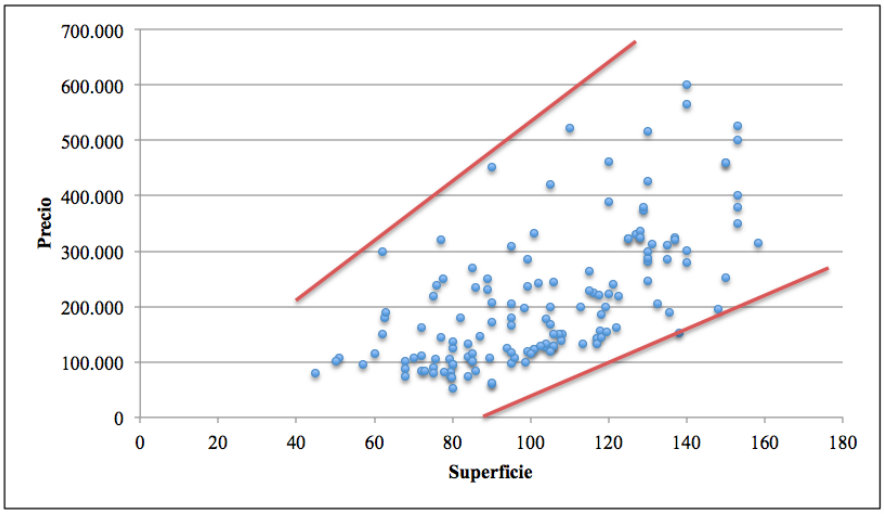


Figura 18. Ejemplo de muestra sin presencia aparente de heterocedasticidad

¿Por qué la heterocedasticidad puede ser un problema en nuestro análisis? Una propiedad interesante de los modelos de regresión es la *influencia* de las observaciones en la pendiente de la recta de regresión. No todos los puntos afectan o influyen por igual en el valor de *b*. Esta influencia es mayor conforme se alejan del punto central de los datos. Esto es, si la media de la superficie es de 90 metros cuadrados, una vivienda con 150 metros cuadrados tendrá mucho más peso en el cálculo de la pendiente *b* que una vivienda de 90 metros cuadrados (¡en realidad la de 90 metros cuadrados no influirá!). Así pues, la pendiente de la recta de regresión será mucho más pronunciada por la presencia de una vivienda de 150 metros cuadrados y más de 1.000.000€, que hace que la relación precio/superficie sea muy superior a la del resto de observaciones.

Dicho de otra forma, es como si la recta de regresión se acomodara de tal forma que la estimación de las viviendas con una superficie muy superior a la media sea mucho más ajustada, a costa de aquellas que tienen una superficie más acorde con el promedio.

Otro problema sobrevenido es que el error cometido en la estimación de las viviendas con una superficie anormalmente grande será mayor que para las viviendas con pequeña superficie. Imagine una recta de regresión sobre los datos de la figura 17, que se encuentre a mitad camino entre las dos rectas imaginarias. Si mide la distancia de los puntos a dicha recta, es evidente que para superficies pequeñas los residuos también serán pequeños, pero para las viviendas grandes los errores pueden ser exageradamente elevados.

Una manera de paliar el efecto negativo de la heterocedasticidad, que no eliminarlo totalmente, es suprimir aquellas observaciones que pueden originarlo. Pero esto puede hacer que la muestra se vea seriamente mermada, como comentamos en un apartado anterior.

Lo más habitual es tomar logaritmos en las variables del modelo de regresión. En la siguiente figura aparecen representadas las mismas viviendas que en la figura 17, pero ahora no sobre las variables originales de precio y superficie, sino sobre sus transformadas logarítmicas. Puede apreciarse como el problema se ha minimizado, y prácticamente las dos rectas que envuelven los datos pueden considerarse paralelas.

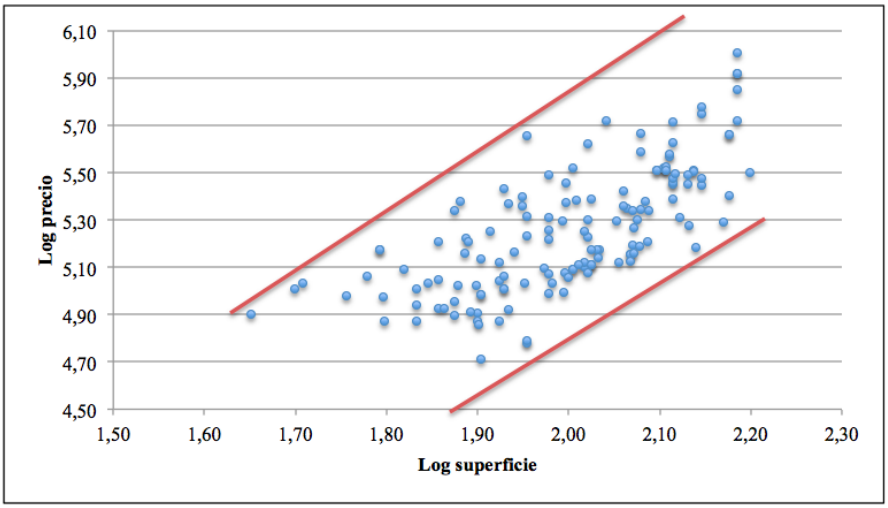


Figura 19. Toma de logaritmos en las variables de la regresión

En la figura 20 aparece representado el histograma de los residuos en el modelo que contempla las variables en su versión original, sin considerar los logaritmos. Podemos apreciar cómo los mayores errores aparecen a la derecha de la distribución. Es decir, aparecen viviendas cuyo precio observado es muy superior al estimado por la recta de regresión, un problema derivado de la presencia de heterocedasticidad. Este problema hace que el error cometido en la estimación de las viviendas de mayor precio sea mucho mayor que en las viviendas de menor precio.

En la figura 21 se representa el histograma de los residuos en el modelo transformado, donde se han considerado los logaritmos del precio y la superficie. Aquí los residuos sí se distribuyen de forma más armoniosa, con colas simila-

res a izquierda y derecha de la distribución, con valores extremos en torno a -0,4 y 0,4.

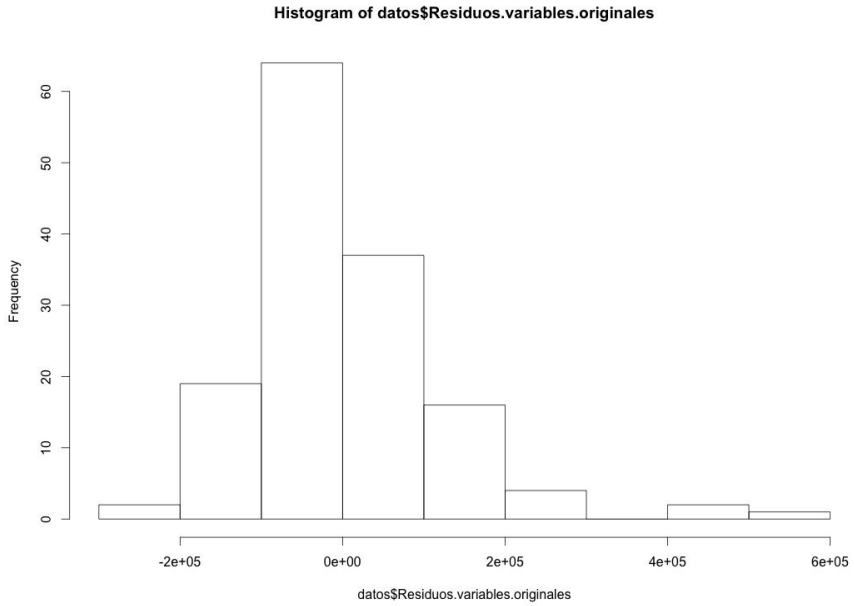


Figura 20. Histograma de residuos con variables originales

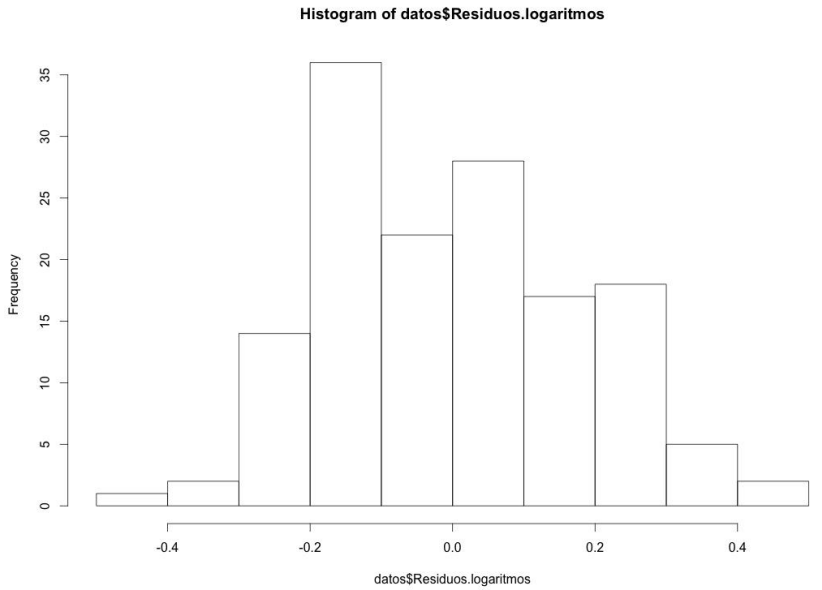


Figura 21. Histograma de residuos con transformación logarítmica de las variables

Al aplicar logaritmos sobre las variables originales tenemos que pensar que nuestro modelo ha cambiado sus unidades, con lo que tomaremos algunas precauciones a la hora de hacer predicción. En el modelo logarítmico, la recta de regresión obtenida ha sido:

$$Log\ Precio = 2{,}0408 + 1{,}6084 \times Log\ Superficie$$

con lo que si, por ejemplo, queremos valorar una vivienda de 120 metros cuadrados:

$$\begin{aligned} Log\ Precio &= 2{,}0408 + 1{,}6084 \times Log\ Superficie \\ &= 2{,}0408 + 1{,}6084 \times Log\ \ 120 \\ &= 2{,}0408 + 1{,}6084 \times 2{,}0792 = 5{,}3850 \end{aligned}$$

Y deshaciendo el logaritmo sobre el precio obtendremos el precio estimado que buscamos:

$$Precio = exp\ 5{,}3850\ = 242.661{,}01€$$

3.10 Limitaciones en la predicción mediante modelos de regresión

Los modelos de regresión no siempre son aplicables en la práctica valorativa. En primer lugar, debemos tener presente que para aplicarlos nuestra muestra debe tener un tamaño mínimo. Cuando se trabaja con modelos de regresión simple, con una única variable explicativa, se aconseja tener al menos 100 observaciones o viviendas para llevar a cabo el análisis. No obstante, en la práctica se puede trabajar con menos datos si resulta extremadamente complicada su obtención, o si la población es tan reducida que conseguir una muestra de 100 observaciones sea algo inabordable en la práctica. En cualquier caso, nunca deberíamos aplicar estos modelos si nuestro número de observaciones en la muestra no llega a 30.

En el siguiente capítulo abordaremos la regresión múltiple, mucho más interesante desde el punto de vista práctico que la regresión simple. En ese caso necesitaremos trabajar con 30 observaciones como mínimo por cada variable explicativa considerada. Esta necesidad de mayor información se verá recompensada con una mejora significativa en la capacidad explicativa de nuestros modelos de valoración.

También es importante que la muestra sea heterogénea en las variables consideradas, tanto la dependiente como la independiente. Esto significa que en la muestra tenemos que tener viviendas con un amplio rango de precios y superficies, si la superficie fuera la variable independiente. Imagine que ha seleccionado su muestra y que todas las viviendas tienen una superficie dentro del rango 90-100 metros cuadrados, y que los precios son muy diferentes entre sí: pongamos dentro del rango 70.000-250.000 euros. Difícilmente la superficie va a poder explicar dichas variaciones en los precios. Si todas tienen una superficie similar, es evidente que dicha variable no explica la diferencia en precios, y por lo tanto tendremos que recurrir a otra variable explicativa para poder justificar las diferencias.

De igual modo, si los precios de las viviendas en nuestra muestra son muy similares, entonces no tendrá sentido intentar explicar las diferencias a partir de ninguna variable independiente. ¡Simplemente porque no hay diferencias que explicar!

Un error muy habitual, y que debemos intentar evitar a toda costa, es intentar estimar el precio de una vivienda cuando el valor de la variable independiente está fuera del rango considerado en la muestra. Imagine que la superficie de

las viviendas de su muestra está entre un mínimo de 68 metros cuadrados, y un máximo de 150 metros cuadrados. En ese caso, no intente valorar una vivienda que tenga menos de 60 metros cuadrados, o una por encima de los 150. Dicho de otra forma, la función de valoración obtenida mediante análisis de regresión no es extrapolable a valores fuera del rango considerado en la variable explicativa.

Los estadísticos suelen incluir más hipótesis para considerar la aplicación del análisis de regresión, como que los datos se ajusten a una distribución normal, y que los residuos obtenidos mediante el análisis también cumplan con la hipótesis de normalidad.

Capítulo 4. El modelo de regresión múltiple

4.1 Introducción

La regresión múltiple es la extensión natural del modelo de regresión simple. En lugar de explicar la variable dependiente a través de una única variable explicativa, se amplía el número de variables independientes con el objetivo de mejorar los resultados. De esta forma se suele aumentar la capacidad explicativa de nuestros modelos, con un mayor valor en el estadístico R^2.

La forma funcional de los modelos de regresión múltiple es la siguiente:

$$Y = f\ X\ = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

donde Y es la variable dependiente (el precio en nuestro caso), β_0 es la constante o intercepto, X_i es la i-ésima variable explicativa, y β_i es el coeficiente asociado a la i-ésima variable explicativa. Suponemos que tenemos n variables explicativas. En el caso $n = 1$ estaríamos ante el modelo de regresión simple, en el que simplemente hemos denominado como β_0 al coeficiente a y como β_1 al coeficiente b.

La interpretación de los coeficientes varía con respecto al modelo de regresión simple, en concreto los asociados a las variables explicativas.

El coeficiente β_0 se interpreta como el valor del regresando Y cuando todas las variables explicativas toman valor cero. El coeficiente β_i se interpreta como el incremento experimentado por la variable Y cuando la variable X_i se incrementa en una unidad, manteniendo constantes el resto de variables, *ceteris paribus*.

Pongamos como ejemplo la siguiente función, donde explicamos el precio de las viviendas a partir de su superficie y el número de dormitorios:

$$Precio = 10.000 + 2.500 \times Superficie + 4.000 \times NumDormitorios$$

El valor que tendría una hipotética vivienda de cero metros cuadrados y ningún dormitorio sería de 10.000€. Por cada metro adicional de la vivienda, el precio de la misma se incrementaría en 2.500€, suponiendo que se mantiene constante el número de dormitorios. Esto significa que si, por ejemplo, esta-

mos valorando una vivienda de 100 metros cuadrados y 2 dormitorios, el incremento de 1 metro cuadrado de superficie supondría incrementar el precio estimado en 2.500€. Pero, ¿cuál sería el precio estimado si además de incrementarse en un metro cuadrado la superficie, también añadiéramos 1 dormitorio adicional? El incremento de dormitorios se valora en 4.000€. Luego en ese caso, a los 2.500€ por el incremento de un metro cuadrado de la superficie, habría que sumar 4.000€ por la adición de un dormitorio. En total, 6.500€.

4.2 Significación estadística del modelo y de los coeficientes

Como en el caso de la regresión simple, también debemos evaluar la significación estadística del modelo en su conjunto así como de cada uno de los coeficientes estimados. En este caso, tendremos un coeficiente para la constante y otro coeficiente por cada una de las variables explicativas.

En el siguiente ejemplo consideraremos una función de regresión que relaciona el precio de las viviendas con su superficie, número de dormitorios y antigüedad del edificio. El modelo obtenido ha sido el siguiente:

$$Precio = 39.711{,}15 + 2.516{,}35 \times Superficie - 21.996{,}55 \times NumDormitorios - 1.519{,}99 \times Antigüedad$$

Puesto que ahora trabajamos con más de una variable no hablaremos de una recta de regresión, sino de un hiperplano de regresión. Ciertamente más complicado de representar, pero el fondo de la cuestión no varía.

Se han incluido en la muestra un total de 136 observaciones (viviendas). Puesto que son 3 las variables explicativas consideradas, estamos claramente por encima de las 30 observaciones por variable que como mínimo requeriríamos para realizar el análisis de regresión múltiple. No obstante, los más puristas dirían que se necesitan al menos 120 observaciones puesto que son 4, y no 3, los coeficientes a estimar: la constante más los 3 coeficientes asociados a las variables explicativas. En cualquier caso, estamos también por encima de esas 120 observaciones.

Regresión Lineal							
Estadísticos de Regresión							
R	0,7829						
R Cuadrado	0,61293						
R Cuadrado Ajustado	0,60414						
S	56.820,07022						
Número Total de Casos	136						
Precio = 39771,1551 + 2516,3479 * Superficie - 21996,5527 * NumDormitorios - 1519,9954 * Antig?edad							
ANOVA							
	d.f.	*SS*	*MS*	*F*	*nivel p*		
Regresión	3,	6,74848E+11	2,24949E+11	69,67563	0,E+0		
Residuo	132,	4,26165E+11	3.228.520.379,86326				
Total	135,	1,10101E+12					
	Coeficientes	*Error Estándar*	*LCL*	*UCL*	*Estadístico t*	*nivel p*	*H0 (2%) rechazado?*
Intercepto	39.771,15515	24.341,72368	-17.551,89155	97.094,20184	1,63387	0,10467	*No*
Superficie	2.516,34786	223,83335	1.989,23608	3.043,45965	11,24206	0,E+0	*Si*
NumDormitorios	-21.996,55269	7.992,34046	-40.817,95215	-3.175,15322	-2,7522	0,00675	*Si*
Antig?edad	-1.519,99537	379,48454	-2.413,65476	-626,33597	-4,00542	0,0001	*Si*
T (2%)	2,35493						
LCL - Valor inferior de un intervalo de confianza (LCL)							
UCL - Valor superior de un intervalo de confianza (UCL)							

Figura 22. Ejemplo de regresión múltiple con la hoja de cálculo Microsoft Excel

El modelo ha obtenido un estadístico R^2 de 61,29%. Claramente es un porcentaje muy inferior al mínimo requerido en un trabajo de valoración. Por lo general, no deberíamos admitir modelos de valoración con un coeficiente de determinación inferior al 85%-90%. Esto no significa que debamos desechar el modelo, sino que tendremos que buscar mejoras que permitan conseguir ese coeficiente de determinación que nos hemos marcado como mínimo.

Es importante señalar que si queremos comparar diferentes modelos de regresión múltiple no lo haremos a través del estadístico R^2, sino a través de una variante del mismo: el estadístico R^2 corregido o ajustado, que tiene en cuenta tanto el número de observaciones como el número de variables explicativas del modelo. Es decir, un modelo será mejor que otro si obtiene un R^2 ajustado mayor, con independencia del valor obtenido en el R^2. Por lo tanto, a partir de ahora nos fijaremos en el valor del R^2 ajustado, en este caso el 60,41%, para analizar la bondad en el ajuste de los modelos de regresión múltiple.

Respecto de la significación del modelo en su conjunto, podemos ver que sí se obtiene una significación aceptable, puesto que el nivel p de la tabla ANOVA es inferior al 5% (nivel de confianza del 95%).

También son significativos los coeficientes asociados a las variables explicativas Superficie, Número de dormitorios y Antigüedad, con niveles p todos ellos inferiores al 5%. El coeficiente que más se acerca a ese límite es el asociado al Número de dormitorios, con un nivel p del 0,675%. El coeficiente que no ha resultado ser estadísticamente distinto de cero es el de la constante, con un nivel p del 10,47%. Por lo tanto, podríamos eliminarlo del modelo y repetir el análisis de regresión. Sin embargo, el cambio en el resto de coefi-

cientes no sería significativo, y tendríamos el hándicap de que el valor del estadístico R^2 ajustado ya no sería válido, pues en los modelos sin constante no se pueden comparar los valores del R^2 ajustado con los modelos que sí tienen constante.

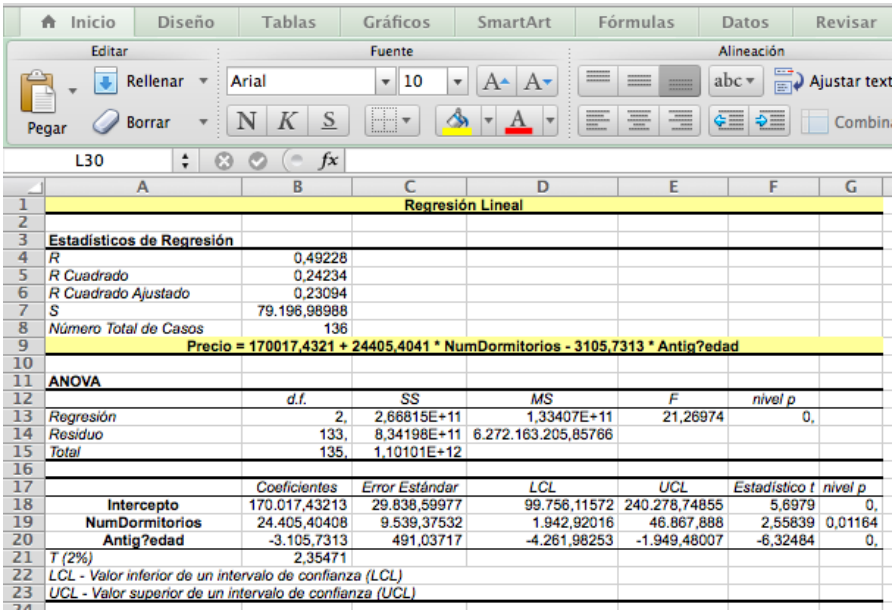
Por lo tanto, y atendiendo al mismo criterio defendido en el capítulo anterior, podríamos mantener la constante en nuestro modelo de valoración.

Un aspecto importante que siempre debemos vigilar es el signo de los coeficientes. Estos deben mantener cierta coherencia económica para que el modelo pueda ser considerado válido, con independencia del nivel alcanzado por el R^2 ajustado. Por ejemplo, si nuestro modelo presentara un coeficiente negativo en la superficie, sería un indicio de que algo extraño ocurre en nuestra muestra o en nuestro modelo de valoración, ya que difícilmente vamos a poder presentar un modelo a nuestro cliente que valore negativamente la superficie. Esta situación suele ser indicadora de un problema que veremos más adelante, el de la multicolinealidad.

Vemos que nuestro modelo valora positivamente la superficie, y negativamente el número de dormitorios y la antigüedad. Parecen razonables lo signos de la superficie (a mayor superficie, mayor precio) y el de la antigüedad (a mayor antigüedad, menor precio). Sin embargo puede parecer contradictorio que el número de dormitorios se valore negativamente: a mayor número de dormitorios, ¿menor precio de la vivienda? Esto podría hacernos pensar que algo ha fallado en nuestro modelo, que quizá las observaciones no se hayan escogido adecuadamente, o simplemente nos hayamos equivocado introduciendo algunos valores en la base de datos.

Pues bien, en este caso el signo del coeficiente asociado al número de dormitorios sí se encuentra justificado desde un punto de vista económico. En primer lugar, vamos a realizar una regresión entre el precio y dos de las variables explicativas: número de dormitorios y antigüedad. Al compararlo con el anterior podremos extraer conclusiones reveladoras.

Al regresar el precio frente a estas dos variables obtenemos el siguiente resultado:



	A	B	C	D	E	F	G
1	Regresión Lineal						
2							
3	**Estadísticos de Regresión**						
4	*R*	0,49228					
5	*R Cuadrado*	0,24234					
6	*R Cuadrado Ajustado*	0,23094					
7	*S*	79.196,98988					
8	*Número Total de Casos*	136					
9	Precio = 170017,4321 + 24405,4041 * NumDormitorios - 3105,7313 * Antig?edad						
10							
11	**ANOVA**						
12		*d.f.*	*SS*	*MS*	*F*	*nivel p*	
13	*Regresión*	2,	2,66815E+11	1,33407E+11	21,26974	0,	
14	*Residuo*	133,	8,34198E+11	6.272.163.205,85766			
15	*Total*	135,	1,10101E+12				
16							
17		*Coeficientes*	*Error Estándar*	*LCL*	*UCL*	*Estadístico t*	*nivel p*
18	**Intercepto**	170.017,43213	29.838,59977	99.756,11572	240.278,74855	5,6979	0,
19	**NumDormitorios**	24.405,40408	9.539,37532	1.942,92016	46.867,888	2,55839	0,01164
20	**Antig?edad**	-3.105,7313	491,03717	-4.261,98253	-1.949,48007	-6,32484	0,
21	*T (2%)*	2,35471					
22	*LCL - Valor inferior de un intervalo de confianza (LCL)*						
23	*UCL - Valor superior de un intervalo de confianza (UCL)*						

Figura 23. Regresión múltiple entre el precio y las variables explicativas número de dormitorios y antigüedad

He aquí la primera sorpresa: ¡el coeficiente del número de dormitorios es positivo! Parece entonces que dicho coeficiente no sólo cambia en valor de modelo en modelo –algo lógico por otra parte-, sino también en signo. El signo positivo desde luego nos puede parecer mucho más adecuado, sobre todo teniendo en cuenta que para esta muestra la correlación entre precio y número de dormitorios es también positiva: 12%.

Entonces, ¿por qué en el primer modelo, donde aparece como variable explicativa la superficie, el signo del número de dormitorios es negativo? La respuesta está en la aparición de la superficie como variable explicativa, y la relación que dicha variable guarda con el precio y el número de dormitorios. Imagine que tiene que valorar dos viviendas con exactamente la misma superficie y antigüedad, pero con diferente número de dormitorios. Supongamos que una de ellas tiene 2 dormitorios y la otra 4. Es evidente que la vivienda con 2 dormitorios tendrá, en promedio, habitaciones más espaciosas y amplias que la vivienda con 4 dormitorios. Por lo tanto, a igualdad de condiciones en el resto de variables, usted atribuiría un precio mayor a la vivienda que tiene un espacio más amplio, armonioso y equilibrado entre sus habitaciones. He ahí la razón de por qué al combinar superficie con número de dormitorios observará en la mayoría de los casos que el coeficiente de esta última variable es negativo.

No se trata, entonces, de un error de transcripción de los datos, o de mala elección de las observaciones en nuestra muestra. La lógica económica explica el signo de ese coeficiente.

4.3 Datos nominales y ordinales en los modelos de regresión

Aunque de momento no habíamos recabado en ello, la incorporación de las variables tipo nominal u ordinal en los modelos de regresión no es tan directa como en el caso de las variables numéricas.

Si usted incluye como variable explicativa la superficie en un modelo de regresión, sabe cómo interpretar el coeficiente obtenido: incremento en el precio por un incremento de un metro cuadrado en la superficie, *ceteris paribus*. Si ahora la variable fuera de tipo ordinal, como el número de dormitorios, la interpretación sería análoga: incremento en el precio por un incremento de una habitación, *ceteris paribus*.

Para la superficie es lógico pensar que pasar de 80 a 81 metros cuadrados se pueda valorar, aproximadamente, en la misma cantidad que pasar de 81 a 82 metros cuadrados. Pero en el caso del número de dormitorios esto ya no es tan evidente. ¿Pagaría usted lo mismo por pasar de una habitación a dos? ¿Y de dos a tres? Parece claro que los incrementos marginales en este caso no tienen por qué ser constantes. De no considerar esta posibilidad estamos restringiendo nuestro modelo, pues le imponemos que valore esos incrementos en la variable independiente de la misma forma, con exactamente la misma cantidad. Y restringir un modelo, como ya examinamos anteriormente, significa empeorar la solución en la mayoría de los casos. Demos entonces la posibilidad de que sea nuestro modelo el que valore si el precio a pagar por cada dormitorio adicional sea el mismo, con independencia del número de dormitorios.

Para incluir esta opción en nuestros modelos debemos transformar las variables nominales u ordinales en variables binarias. La transformación se debe hacer de la siguiente forma: por cada variable nominal u ordinal con n niveles diferentes, se deben construir $n - 1$ variables binarias.

Las variables binarias sólo pueden tomar dos valores, por lo general 0 y 1, para indicar la ausencia o la presencia de una propiedad.

Veamos un ejemplo con la variable número de dormitorios. En nuestra muestra esta variable fluctuaba entre los valores 1 y 4, según se muestra en el siguiente gráfico:

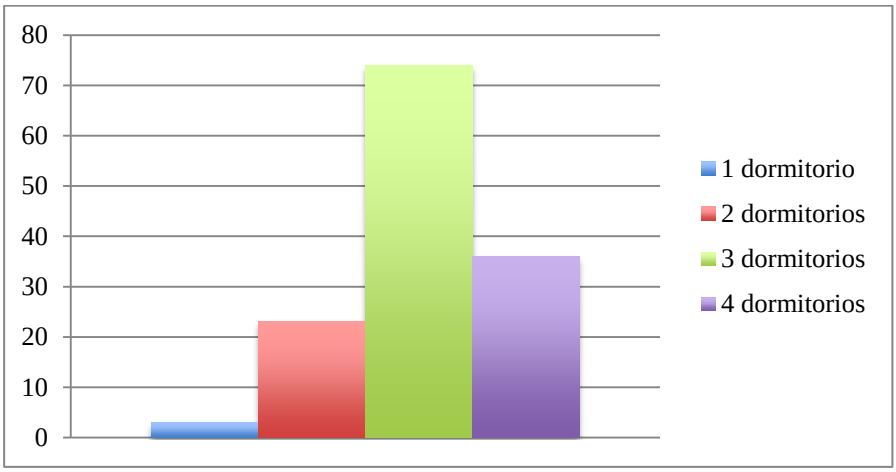


Figura 24. Gráfico de barras del número de dormitorios

Puesto que son 4 los niveles de esta variable, tendremos que crear 3 variables binarias. Por ejemplo:

- NumDormitorios2: Tomará valor 1 (se activará) cuando el número de dormitorios sea 2, y cero (se desactivará) en caso contrario. Es decir, que si el número de dormitorios es 1, 3 o 4, entonces la variable tomará valor cero.
- NumDormitorios3: Tomará valor 1 cuando el número de dormitorios sea 3, y cero en caso contrario.
- NumDormitorios4: Tomará el valor 1 cuando el número de dormitorios sea 4, y cero en caso contrario.

De esta forma, sólo necesitamos 3 variables para poder representar el número de dormitorios de cualquier vivienda en nuestra muestra, dejando como nivel de referencia el de 1 dormitorio. Por ejemplo, si una vivienda tiene 3 habitaciones, las variables binarias tomarán los siguientes valores:

$$NumDormitorios2 = 0, NumDormitorios3 = 1, NumDormitorios4 = 0$$

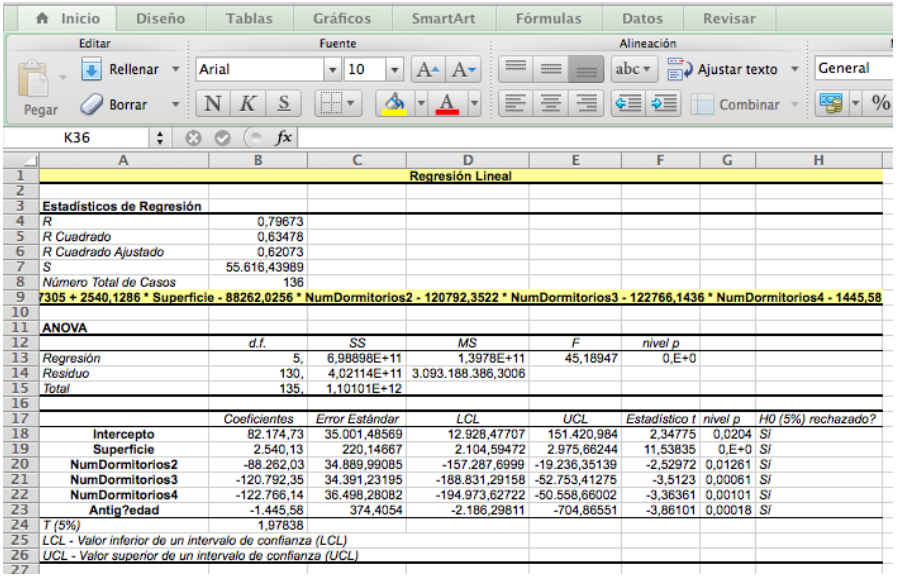
Se estará preguntando, ¿qué ocurre cuando una vivienda tenga 1 dormitorio? Este caso vendrá representado por la combinación:

$$NumDormitorios2 = 0, NumDormitorios3 = 0, NumDormitorios4 = 0$$

En la siguiente tabla aparecen algunos ejemplos de representación del número de dormitorios con estas 3 nuevas variables binarias:

NumDormitorios	NumDormitorios2	NumDormitorios3	NumDormitorios4
4	0	0	1
4	0	0	1
3	0	1	0
2	1	0	0
1	0	0	0
3	0	1	0
3	0	1	0
2	1	0	0
1	0	0	0
3	0	1	0

De esta forma, en la nueva regresión sustituiríamos la variable original NumDormitorios por las 3 variables binarias recién definidas. El resultado de la regresión múltiple sería el siguiente:



Regresión Lineal

Estadísticos de Regresión	
R	0,79673
R Cuadrado	0,63478
R Cuadrado Ajustado	0,62073
S	55.616,43989
Número Total de Casos	136

7305 + 2540,1286 * Superficie - 88262,0256 * NumDormitorios2 - 120792,3522 * NumDormitorios3 - 122766,1436 * NumDormitorios4 - 1445,58

ANOVA

	d.f.	SS	MS	F	nivel p
Regresión	5,	6,98898E+11	1,3978E+11	45,18947	0,E+0
Residuo	130,	4,02114E+11	3.093.188.386,3006		
Total	135,	1,10101E+12			

	Coeficientes	Error Estándar	LCL	UCL	Estadístico t	nivel p	H0 (5%) rechazado?
Intercepto	82.174,73	35.001,48569	12.928,47707	151.420,984	2,34775	0,0204	Sí
Superficie	2.540,13	220,14667	2.104,59472	2.975,66244	11,53835	0,E+0	Sí
NumDormitorios2	-88.262,03	34.889,99085	-157.287,6999	-19.236,35139	-2,52972	0,01261	Sí
NumDormitorios3	-120.792,35	34.391,23195	-188.831,29158	-52.753,41275	-3,5123	0,00061	Sí
NumDormitorios4	-122.766,14	36.498,28082	-194.973,62722	-50.558,66002	-3,36361	0,00101	Sí
Antig?edad	-1.445,58	374,4054	-2.186,29811	-704,86551	-3,86101	0,00018	Sí
T (5%)	1,97838						

LCL - Valor inferior de un intervalo de confianza (LCL)

UCL - Valor superior de un intervalo de confianza (UCL)

Figura 25. Análisis de regresión múltiple con variables binarias

En primer lugar, vemos que hemos mejorado ligeramente el ajuste, pasando de un R^2 ajustado de 60,41% a un valor de 62,07%. Tanto el modelo en su conjunto, como cada uno de los 6 coeficientes estimados han resultado ser estadísticamente significativos para un nivel de confianza del 95%. Los signos de los coeficientes asociados a la superficie y a la antigüedad tienen el signo esperado: positivo. Antes de analizar los signos de los coeficientes de las variables binarias es necesario conocer su interpretación:

- El coeficiente de NumDormitorios2 se interpreta como la diferencia de precio entre una vivienda con dos dormitorios respecto del caso base o nivel de referencia de una vivienda con un dormitorio. En este caso, la habitación adicional supone un menor precio de 88.262,03€.
- El coeficiente de NumDormitorios3 se interpreta como la diferencia de precio entre una vivienda con tres dormitorios respecto del caso base o nivel de referencia de una vivienda con un dormitorio. En este caso, tener 3 habitaciones supone un menor precio de 120.792,35€ respecto de tener una sola habitación. Vemos cómo el descenso en el precio continúa, pero así como tener dos dormitorios suponía un menor valor de 88.262,03€, pasar a un tercer dormitorio no significa minorar el precio en otros 88.262,03€. La diferencia respecto de dos dormitorios es de 120.792,35-88.262,03 = 32.530,32€. Si hubiéramos utilizado la variable original del número de dormitorios, cada uno de ellos hubiera disminuido el precio en la misma cantidad. Y vemos que, para las viviendas de nuestra muestra, el descenso en el precio es diferente según el número de dormitorios.
- El coeficiente de NumDormitorios4 se interpreta como la diferencia de precio entre una vivienda con cuatro dormitorios respecto del caso base o nivel de referencia de una vivienda con un dormitorio. En este caso, tener 4 habitaciones supone un menor precio de 122.766,14€ respecto de tener una sola habitación. ¿Qué diferencia tenemos respecto del caso de 3 habitaciones? Restando los coeficientes 122.766,14-120.792,35 vemos como la habitación adicional reduce el precio en sólo 1.973,79€.

Aunque estos coeficientes puedan parecer llamativos, debemos tener presentes que las interpretaciones son suponiendo constantes el resto de variables. Es decir, la diferencia de precios se entiende para viviendas con la misma superficie y antigüedad.

Pongamos un ejemplo para entender esta afirmación. Puede pensar que pasar de 3 a 4 dormitorios hace que el precio baje en 1.973,79€. ¿Cómo es posible que una vivienda con más habitaciones tenga un menor precio? La respuesta es que no necesariamente el precio será más bajo. Si una habitación adicional supone, por ejemplo, una superficie de 10 metros cuadrados más, entonces el

precio se verá incrementado por esa partida en 10x2.540,13 = 25.401,30€. Es decir, que el precio habría subido en 25.401,30-1.973,79 = 23.427,51€.

El mismo tratamiento debemos dar a las variables de tipo nominal, como por ejemplo la orientación de las viviendas. Supongamos que tenemos definida esta variable con 4 posibles niveles: Este, Oeste, Norte y Sur. Podríamos considerar la orientación Este como el nivel de referencia, y crear entonces tres variables binarias: BinariaOeste, BinariaNorte y BinariaSur. Por ejemplo.

4.4 Mejorando la capacidad explicativa del modelo

Está bien. Hemos construido nuestro modelo de regresión múltiple, pero el paupérrimo R^2 ajustado de 60,41% no nos permite aplicar nuestro modelo en la práctica profesional. ¿Qué podemos hacer? Tenemos varias posibilidades por explorar:

- Incluir más variables explicativas. Desde luego es la mejor forma de intentar aumentar la capacidad explicativa de cualquier modelo, buscar alternativas a las variables explicativas que estamos utilizando. Siempre es posible encontrar nueva información que nos ayude a explicar las diferencias en precio de nuestra muestra. Pero también para esto tenemos un límite, puesto que, por una parte, tendremos un número limitado de viviendas que no permita incrementar de manera continuada el número de variables explicativas y, por otro, toda nueva información supone un mayor coste de tiempo y dinero para el tasador.
- Comprobar que nuestro modelo no tiene problemas de heterocedasticidad. Puede que tomando logaritmos en todas o algunas de nuestras variables consigamos elevar el estadístico R^2 ajustado. En la mayoría de los casos los incrementos no serán muy significativos, pero algo es algo. Además, el coste de esta posibilidad es cero. Se trata simplemente de transformar una o más variables para trabajar con el logaritmo de las mismas.
- Detectar y eliminar *outliers*. Por lo general, siempre va a ser posible identificar, después de haber considerado todas las variables explicativas, alguna o varias observaciones que puedan ser consideradas anómalas. Su exclusión, si no supone una merma significativa de la muestra, puede mejorar sensiblemente el estadístico R^2 ajustado.

Vamos a considerar, de momento, las dos primeras posibilidades. Tomar el logaritmo del precio es sencillo, lo podemos hacer con cualquier programa que estemos utilizando para estimar nuestro modelo. Respecto de la inclusión de nuevas variables explicativas, junto con las que ya teníamos hemos consi-

derado varias adicionales. En conjunto, las variables explicativas consideradas son las siguientes:

- Superficie: superficie útil expresada en metros cuadrados.
- Antigüedad: expresada en años, se refiere a la antigüedad del edificio en que se ubica la vivienda.
- Número de viviendas en el edificio: esta variable está relacionada con las dimensiones del edificio.
- Número de plantas: mide la altura del edificio.
- Número de planta de la vivienda: referencia de la altura en la que se encuentra la vivienda dentro del edificio.
- Calidad de la construcción: medida cualitativa sobre la construcción y sus acabados. Se han considerado tres niveles, de más a menos calidad: Calidad constructiva alta, calidad constructiva media y calidad constructiva de viviendas sociales. Al tener 3 niveles, se han creado dos variables binarias: CalidadConstrAlta y CalidadConstrViviendaSocial. De esta forma, el nivel que se queda como referencia es el de las viviendas con calidad constructiva media.
- Calidad urbanística: variable cualitativa. Se relaciona con la calidad del entorno urbanístico del edificio. Se han definido tres niveles: alto, medio y bajo. Dejado como nivel de referencia el medio, las dos variables binarias que se han creado han sido CalidadUrbAlta y CalidadUrbBaja.
- Entorno comercial: define el número y calidad de los establecimientos comerciales de la zona de influencia del edificio. De esta forma, se diferencia entre las zonas más comerciales de las más deprimidas. Como en el caso de las dos variables anteriores, también se han diferenciado tres niveles: Excelente, Bueno y Básico. Tomando como nivel de referencia el entorno bueno, se han creado las variables binarias EntComExcel y EntComBas.
- Renta: de nuevo una variable cualitativa, en este caso se mide el nivel de renta disponible del entorno donde se sitúa el edificio. De esta manera se distinguen las zonas de mayor nivel adquisitivo de las que pueden ser más depresivas. Se han utilizado tres niveles: Alta, Media y Baja. Tomando como referencia una renta media, se han creado las variables binarias RentaAlta y RentaBaja.
- Número de dormitorios: en la muestra el número de dormitorios está en el rango 1-5. Ya se comentó con anterioridad la necesidad de transformar esta variable ordinal en n-1 variables binarias. Para ello hemos tomado como base las viviendas con un dormitorio, creando dos variables binarias: NumDorm2 y NumDorm3, correspondientes a las viviendas con 2, y 3 o más dormitorios, respectivamente. Esto es, si una vivienda tiene 3 dormitorios se activa la variable binaria

NumDorm3. Y también se activa la misma variable si la vivienda tiene 4 o 5 habitaciones.

- Número de baños: al igual que con el número de dormitorios, también se ha transformado esta variable ordinal. Tomando como base las viviendas de 1 sólo cuarto de baño, se han creado las variables binarias NumBaños2 y NumBaños3Mas. Esta última recoge las viviendas con 3 o más baños.
- Número de ascensores: variable definida a cuatro niveles: sin ascensor, con un ascensor, con dos ascensores, y con 3 o más ascensones. Se ha tomado como nivel base los edificios sin ascensor, con lo que se han creado las variables binarias NumAsc1, NumAsc2 y NumAsc3Mas.

Tomando este conjunto de variables explicativas, hemos construido un modelo que explique el logaritmo del precio de las viviendas, para una muestra compuesta por 136 apartamentos.

En las siguientes tablas presentamos algunos de los resultados. No hemos incluido los coeficientes, porque como veremos a continuación el modelo no es el definitivo.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,874[a]	,764	,749	,09922

a. Variables predictoras: (Constante), Antigüedad, NumAsc1, EntComBasico, Superficie, RentBaja, CalidadUrbAlta, NumPlantas, NumAsc2

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	4,047	8	,506	51,394	,000[b]
	Residual	1,250	127	,010		
	Total	5,298	135			

a. Variable dependiente: LogPrecio

b. Variables predictoras: (Constante), Antigüedad, NumAsc1, EntComBasico, Superficie, RentBaja, CalidadUrbAlta, NumPlantas, NumAsc2

Figura 26. Modelo de regresión múltiple

Podemos ver como la variable dependiente ya no es el precio, sino el logaritmo del precio (LogPrecio). Como variables explicativas hemos considerado la antig de la vivienda, si el nivel de renta de la zona era bajo (RentBaja), si la calidad urbanística del edificio era alta (CalidadUrbAlta) y el número de

plantas del edificio. Aunque no aparecen sus coeficientes ni su nivel de significación estadística, todos los coeficientes han resultado ser significativos con un nivel de confianza del 95% (valor p inferior al 5%).

La calidad del modelo ha mejorado sensiblemente, pasando a tener un R^2 ajustado de 74,9%. Ciertamente aún está lejos del 85% que consideraríamos como bueno para la práctica profesional, pero la mejora respecto del primero modelo ha sido más que notable.

Sin embargo, esta bondad en el ajuste se ha obtenido aumentando considerablemente el número de variables explicativas. En el modelo aparecen un total de 8 variables explicativas, cuando el número de viviendas en la muestra es de sólo 136.

Para mejorar un poco más el modelo aún nos queda examinar la posible presencia de observaciones atípicas. En los modelos de regresión múltiple su identificación no es tan sencilla como en los modelos de regresión simple, donde podíamos representar gráficamente las variables dependiente e independiente y, de un sólo vistazo, localizar aquellas observaciones cuyo comportamiento se salía de la tendencia central de los datos.

En el caso de la regresión múltiple la identificación de los outliers puede llevarse a cabo a través de los residuos estandarizados. Precisamente, en el siguiente histograma aparecen representados dichos residuos para el caso del último modelo de regresión. Podemos ver cómo aparece una barra próxima al valor 4, que claramente identificaríamos como un dato anómalo. Dicha barra se corresponde con una vivienda que, tras se identificada en la muestra, se excluye del modelo.

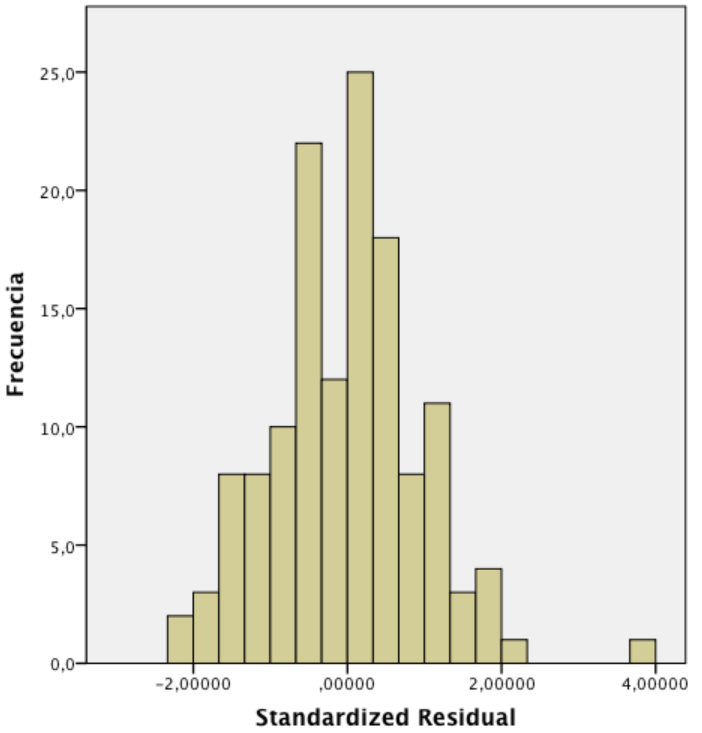


Figura 27. Histograma de residuos estandarizados

Al repetir el modelo de regresión sin la observación atípica, algunas variables explicativas han perdido significación, mientras que otras la han ganado. Más adelante veremos cómo llevar a cabo un modelo de regresión múltiple sin tener que introducir y retirar manualmente variables, según la significación que consigan sus coeficientes. Hasta entonces, lo que nos interesa es conocer cuál es la configuración del nuevo modelo de regresión, que podemos ver en las siguientes tablas.

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,890[a]	,793	,780	,09272

a. Variables predictoras: (Constante), NumAsc2, EntComBasico, NumPlantas, Superficie, CalidadUrbAlta, Antigüedad, NumAsc1, RentBaja

b. Variable dependiente: LogPrecio

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	4,149	8	,519	60,328	,000[b]
	Residual	1,083	126	,009		
	Total	5,232	134			

a. Variable dependiente: LogPrecio

b. Variables predictoras: (Constante), NumAsc2, EntComBasico, NumPlantas, Superficie, CalidadUrbAlta, Antigüedad, NumAsc1, RentBaja

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	4,785	,046		104,749	,000
	Superficie	,004	,000	,509	11,067	,000
	NumPlantas	,009	,002	,267	5,241	,000
	Antigüedad	-,002	,001	-,147	-3,035	,003
	RentBaja	-,044	,020	-,111	-2,186	,031
	NumAsc1	,086	,020	,219	4,363	,000
	CalidadUrbAlta	,088	,029	,149	3,071	,003
	EntComBasico	-,074	,029	-,107	-2,517	,013
	NumAsc2	,061	,028	,119	2,170	,032

a. Variable dependiente: LogPrecio

Figura 28. Modelo de regresión múltiple

El modelo ha mejorado aún más el estadístico R^2 ajustado, que ahora se sitúa en el 79% de capacidad explicativa. Ciertamente no alcanzamos el 85% que nos marcamos como objetivo general, pero la mejora respecto del modelo inicial es importante.

El modelo es significativo en su conjunto (tabla ANOVA) y también lo son cada uno de los coeficientes estimados, incluida la constante. Las variables explicativas del logaritmo del precio han sido la superficie, el número de

plantas del edificio, la antigüedad, y las variables binarias RentBaja, NumAsc1, CalidadUrbAlta, EntComBasico, y NumAsc2.

El signo de los coeficientes también es lógico desde un punto de vista económico. Se valora positivamente la superficie y el número de plantas, y de forma negativa la antigüedad. Respecto de las variables binarias, también se valora positivamente el número de ascensores y la calidad urbanística alta, mientras que se valoran negativamente la renta baja y el entorno comercial básico.

Conviene representar los nuevos residuos estandarizados, de forma que comprobemos que no han aparecido nuevos *outliers* que puedan distorsionar nuestro modelo. En la siguiente figura aparece el histograma de los mismos, y vemos como se encuentran en todos los casos dentro del rango [-3,+3].

A modo de conclusión, podemos apuntar dos aspectos negativos del modelo que acabamos de construir. En primer lugar el estadístico R2 ajustado, que aunque bastante bueno no alcanza el nivel del 85%. En segundo lugar, el excesivo número de variables explicativas para el limitado número de viviendas que tenemos en nuestra muestra. En realidad no es que tengamos pocas observaciones, sino que el ratio número de variables / observaciones es excesivo. Tendríamos que eliminar algunas de las variables, aún reduciendo a su costa la bondad del ajuste.

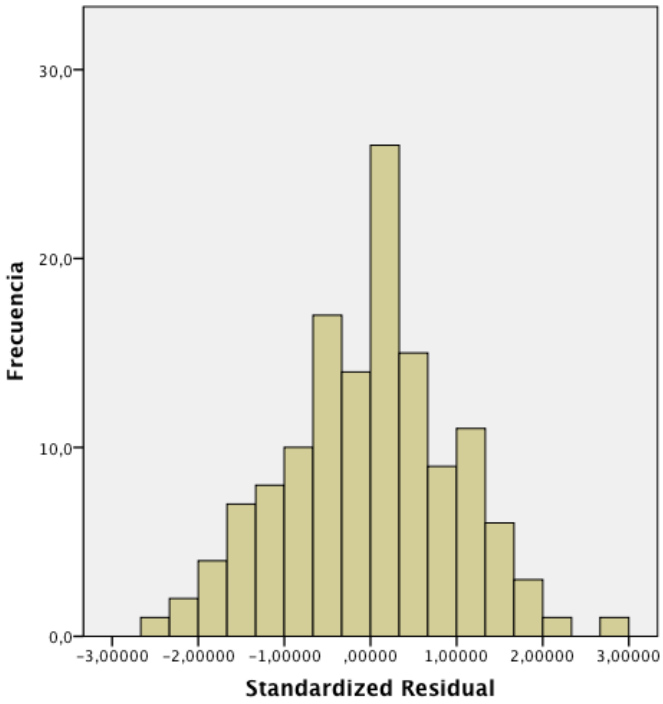


Figura 29. Histograma de los residuos estandarizados

Suponga que, para salvaguardar el ratio número de variables / número de viviendas tuviera que sacrificar algunas de las variables explicativas. ¿Cuáles eliminaría? El proceso de prueba y error, incluyendo y excluyendo variables una a una según su efecto en el R^2 ajustado, es ciertamente costoso en tiempo. Veamos cómo medir la importancia de las variables en nuestro modelo, para de esta manera poder identificar con mayor facilidad las variables candidatas a quedar fuera del modelo de valoración final.

Podríamos pensar que son más importantes aquellas variables que han obtenido un coeficiente, en valor absoluto, mayor que el resto. Por ejemplo, de nuestro último modelo podríamos destacar el coeficiente asociado a la calidad urbanística alta (0,088) o el asociado a los edificios con un único ascensor (0,086). Por el contrario, los coeficientes con valores absolutos más pequeños serían los asociados a los coeficientes de la antigüedad (-0,002) y superficie (0,004); por lo tanto, pensaríamos que estas variables son las menos relevantes en el precio.

Como ya supondrá el lector, la interpretación de estos coeficientes no es la adecuada. En ningún caso podemos asociar mayor importancia a los coefi-

cientes según el valor de los mismos. Y esto es así porque la unidad de medida de cada uno de esos coeficientes es distinta. Por ejemplo, la unidad de medida del coeficiente asociado a la superficie sería log(€)/metros cuadrados, mientras que la unidad de medida del coeficiente antigüedad sería log(€)/años. Como su unidad de medida es distinta, los coeficientes no son comparables entre sí.

Para determinar la relevancia de cada variable en la explicación del precio (en este caso en la explicación del logaritmo del precio) debemos hacer uso de los coeficientes tipificados (coeficientes Beta). A diferencia de los coeficientes del modelo de regresión, los coeficientes tipificados no tienen una unidad de medida determinada, y al estar tipificados sí que permiten la comparación unos con otros.

En el caso de nuestro modelo de regresión múltiple, el coeficiente Beta con mayor valor (absoluto) es el asociado a la superficie (0,509). Por lo tanto, entre las variables consideradas es la superficie la que tiene mayor relevancia en la explicación del logaritmo del precio. Siguiendo el orden de relevancia encontramos al número de planta (0,267), la variable NumAsc1 (0,219), la variable CalidadUrbAlta (0,149), o la antigüedad (-0,147). Obsérvese como la importancia o relevancia de las variables se mide por el valor absoluto de sus coeficientes Beta, sin tener en cuenta el signo de los mismos. De esta forma, la variable menos relevante sería EntComBasico, con un coeficiente Beta de -0,107.

Es posible que el lector considere que algunas variables importantes se han quedado fuera del modleo, como por ejemplo el número de dormitorios. Tenga en cuenta que en muchas ocasiones la información que puedan aportar estas variables ya viene recogida en el modelo por alguna/s otra/s. En el caso del número de dormitorios, la superficie se puede considerar un buen *proxy*, con lo que el número de dormitorios no aportaría gran cosa de cara a la valoración. Al menos para esta muestra.

Una vez determinada la distinta importancia de las variables en el modelo de regresión múltiple, suponga que desea limitar el número de las mismas a 3. Eso supone que tendremos que prescindir de 5 de las 8 variables que nuestro modelo considera.

Para llevar a cabo este "adelgazamiento" de nuestro modelo tenemos dos opciones:

1) Eliminar manualmente las 4 variables con coeficientes Beta menos relevantes: EntComBasico, RentBaja, NumAsc2, Antigüedad y CalidadUrbAlta. De esta forma únicamente seleccionaríamos las 3 más relevantes: Superficie, número de plantas y NumAsc1
2) Aplicar la regresión *paso a paso* (*stepwise regression*). Es un procedimiento algorítmico en el que las variables entran o salen del mode-

lo de regresión atendiendo únicamente al efecto que dicha entrada o salida tiene en la explicación de la variable dependiente.

Supongamos que optamos por la primera opción, e intentamos explicar el logaritmo del precio únicamente a partir de las variables superficie, número de plantas y NumAsc1. El resultado aparece en la siguiente figura:

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,825[a]	,680	,673	,11298

a. Variables predictoras: (Constante), NumAsc1, Superficie, NumPlantas

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	3,560	3	1,187	92,966	,000[b]
	Residual	1,672	131	,013		
	Total	5,232	134			

a. Variable dependiente: LogPrecio

b. Variables predictoras: (Constante), NumAsc1, Superficie, NumPlantas

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	4,631	,039		119,227	,000
	Superficie	,005	,000	,606	11,732	,000
	NumPlantas	,015	,002	,435	8,094	,000
	NumAsc1	,058	,021	,148	2,802	,006

a. Variable dependiente: LogPrecio

Figura 30. Modelo de regresión múltiple. El logaritmo del precio es explicado por la superficie, el número de plantas del edificio, y NumAsc1

Tanto el modelo en su conjunto como los 4 coeficientes estimados son estadísticamente significativos. Si embargo el estadístico R^2 ajustado ha descendido hasta el valor 67,3%. La pregunta que nos hacemos es, ¿realmente es el mejor modelo que podemos construir con únicamente 3 variables explicativas? La respuesta es no.

Los modelo de regresión paso a paso, en inglés *stepwise regresssion*, permiten la obtención del modelo óptimo, en lo que a capacidad explicativa se refiere. Su obtención se basa en un algoritmo informático, que permite encon-

trar la combinación ideal de variables para maximizar la capacidad explicativa de nuestros modelos. No obstante, veremos más adelante que no todo son ventajas, ni mucho menos.

Existen diferentes variantes de este tipo de regresión. En la versión *forward* las variables se van incorporando una a una al modelo de regresión múltiple, mientras que la variante *backward* parte de un modelo con todas las variables explicativas, y las va eliminando de manera individual. En figura 21 aparecen los resultados de la variante *forward*.

En la tabla Resumen del modelo se recoge el estadístico R^2 ajustado para cada uno de los modelos. El más sencillo de todos ellos, con la etiqueta de *Modelo 1*, obtiene un R^2 ajustado de 51,7%, considerando como única variable explicativa la más relevante: la superficie. El siguiente modelo pasa a tener dos variables explicativas, añadiendo a la superficie el número de plantas, y consiguiendo aumentar el R^2 ajustado hasta el 65,6%.

Vemos como para cada uno de estos modelos también obtenemos los coeficientes de cada una de las variables explicativas, justamente en la última tabla de la figura 21. Para todas las tablas, la última fila recoge el modelo definitivo. Aquél en el que se obtiene el mayor estadístico R^2 ajustado y todos los coeficientes han resultado ser estadísticamente significativos. Pero la información del resto de modelos también es valiosa. Si buscáramos el mejor modelo posible con 3 variables explicativas, seleccionaríamos el que aparece como *Modelo 3*. Su R^2 ajustado es del 70,2%, un valor sensiblemente mayor que el obtenido con el modelo a partir de las variables con mayor relevancia según el coeficiente Beta: 67,3%. En lugar de seleccionar la superficie, el número de plantas del edificio y la variable NumAsc1, la regresión paso a paso ha mejorado ese modelo sustituyendo NumAsc1 por la antigüedad del edificio.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,721[a]	,521	,517	,13734
2	,813[b]	,661	,656	,11588
3	,842[c]	,709	,702	,10786
4	,857[d]	,734	,726	,10342
5	,870[e]	,757	,747	,09934
6	,879[f]	,773	,763	,09624
7	,886[g]	,785	,773	,09406
8	,890[h]	,793	,780	,09272

a. Variables predictoras: (Constante), Superficie

b. Variables predictoras: (Constante), Superficie, NumPlantas

c. Variables predictoras: (Constante), Superficie, NumPlantas, Antigüedad

d. Variables predictoras: (Constante), Superficie, NumPlantas, Antigüedad, RentBaja

e. Variables predictoras: (Constante), Superficie, NumPlantas, Antigüedad, RentBaja, NumAsc1

f. Variables predictoras: (Constante), Superficie, NumPlantas, Antigüedad, RentBaja, NumAsc1, CalidadUrbAlta

g. Variables predictoras: (Constante), Superficie, NumPlantas, Antigüedad, RentBaja, NumAsc1, CalidadUrbAlta, EntComBasico

h. Variables predictoras: (Constante), Superficie, NumPlantas, Antigüedad, RentBaja, NumAsc1, CalidadUrbAlta, EntComBasico, NumAsc2

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	2,724	1	2,724	144,388	,000[b]
	Residual	2,509	133	,019		
	Total	5,232	134			
2	Regresión	3,460	2	1,730	128,838	,000[c]
	Residual	1,772	132	,013		
	Total	5,232	134			
3	Regresión	3,708	3	1,236	106,251	,000[d]
	Residual	1,524	131	,012		
	Total	5,232	134			
4	Regresión	3,842	4	,960	89,798	,000[e]
	Residual	1,390	130	,011		
	Total	5,232	134			
5	Regresión	3,959	5	,792	80,247	,000[f]
	Residual	1,273	129	,010		
	Total	5,232	134			
6	Regresión	4,047	6	,674	72,825	,000[g]
	Residual	1,185	128	,009		
	Total	5,232	134			
7	Regresión	4,109	7	,587	66,336	,000[h]
	Residual	1,124	127	,009		
	Total	5,232	134			
8	Regresión	4,149	8	,519	60,328	,000[i]
	Residual	1,083	126	,009		
	Total	5,232	134			

Modelo		Coeficientes no estandarizados		Coeficientes tipificados		
		B	Error típ.	Beta	t	Sig.
1	(Constante)	4,704	,045		103,670	,000
	Superficie	,005	,000	,721	12,016	,000
2	(Constante)	4,655	,039		119,809	,000
	Superficie	,005	,000	,633	12,173	,000
	NumPlantas	,013	,002	,385	7,405	,000
3	(Constante)	4,768	,044		109,276	,000
	Superficie	,004	,000	,594	12,072	,000
	NumPlantas	,010	,002	,303	5,861	,000
	Antigüedad	-,003	,001	-,240	-4,621	,000
4	(Constante)	4,850	,048		101,326	,000
	Superficie	,004	,000	,571	11,985	,000
	NumPlantas	,008	,002	,225	4,156	,000
	Antigüedad	-,003	,001	-,237	-4,762	,000
	RentBaja	-,073	,021	-,182	-3,534	,001
5	(Constante)	4,830	,046		104,208	,000
	Superficie	,004	,000	,539	11,570	,000
	NumPlantas	,009	,002	,273	5,074	,000
	Antigüedad	-,003	,001	-,231	-4,842	,000
	RentBaja	-,080	,020	-,200	-4,029	,000
	NumAsc1	,063	,018	,161	3,451	,001
6	(Constante)	4,790	,047		102,452	,000
	Superficie	,004	,000	,550	12,143	,000
	NumPlantas	,009	,002	,259	4,953	,000
	Antigüedad	-,003	,001	-,179	-3,641	,000
	RentBaja	-,065	,020	-,162	-3,265	,001
	NumAsc1	,069	,018	,176	3,873	,000
	CalidadUrbAlta	,090	,029	,152	3,073	,003
7	(Constante)	4,799	,046		104,701	,000
	Superficie	,004	,000	,539	12,109	,000
	NumPlantas	,009	,002	,253	4,944	,000
	Antigüedad	-,002	,001	-,164	-3,383	,001
	RentBaja	-,057	,020	-,142	-2,880	,005
	NumAsc1	,065	,018	,166	3,731	,000
	CalidadUrbAlta	,097	,029	,163	3,349	,001
	EntComBasico	-,079	,030	-,114	-2,642	,009
8	(Constante)	4,785	,046		104,749	,000
	Superficie	,004	,000	,509	11,067	,000
	NumPlantas	,009	,002	,267	5,241	,000
	Antigüedad	-,002	,001	-,147	-3,035	,003
	RentBaja	-,044	,020	-,111	-2,186	,031
	NumAsc1	,086	,020	,219	4,363	,000
	CalidadUrbAlta	,088	,029	,149	3,071	,003
	EntComBasico	-,074	,029	-,107	-2,517	,013
	NumAsc2	,061	,028	,119	2,170	,032

Figura 31. Modelo de regresión múltiple paso a paso

4.5 El problema de la multicolinealidad

Una de las críticas que podemos realizar a los modelos obtenidos mediante regresión paso a paso está justamente relacionada con lo que denominamos multicolinealidad.

La multicolinealidad se presenta cuando en un mismo modelo combinamos variables que están fuertemente correlacionadas entre sí. Las diferentes relaciones de dependencia que puedan existir entre ellas acaban afectando y sesgando al resultado final, de forma que el modelo de regresión puede invalidarse por la concurrencia de variables correlacionadas, en lugar de mejorarse por la incorporación de nuevas variables explicativas.

Entre los efectos negativos de la multicolinealidad están:

- La obtención de elevados y artificiales valores del estadístico R^2, dando la falsa sensación de que un modelo con alta capacidad explicativa también será un modelo excelente para la predicción. Es precisamente lo que ocurre al incorporar un número excesivo de variables explicativas, en relación con el número de observaciones. Cuando además estas variables están altamente correlacionadas entre sí, se habla de modelos de regresión sobreajustados (*overfitted*). Es entonces cuando se obtienen R^2 próximos al 100%, y creemos que el modelo tendrá también una alta capacidad predictiva. Sin embargo, cuando el modelo es aplicado sobre nuevas observaciones se detectan errores en las estimaciones muy superiores a las registradas en las observaciones que conforman la muestra.
- La aparición de coeficientes con un signo que no se justifica desde un punto de vista económico. Es el caso de variables que pueden tener el mismo grado y signo de relación con el precio, y que al combinarse en el mismo modelo obtienen coeficientes con signo contrario al obtenido por el coeficiente de correlación. Pongamos un ejemplo. Supongamos que pretendemos estimar el precio de las viviendas y, entre las variables explicativas, incluimos la superficie total y la superficie útil de la vivienda. Lógicamente, ambas variables van a estar muy relacionadas entre sí, y también con el precio. La relación parece, a priori, que debe ser claramente positiva. Pues bien, podemos comprobar cómo en la mayoría de los modelos que estimemos, una de las variables aparecerá con un coeficiente negativo. ¿A qué se debe esto? La razón la encontramos en que una variable tiende a "compensar" la presencia de la otra, de forma que el modelo resultante se encuentra sobreajustado.
- La significación estadística de variables que, a priori, no guardan aparente relación con la variable dependiente. Las variables que presentan esta característica se dice que ponen de manifiesto una rela-

ción espuria. Esto es, podemos encontrar una variable que no tenga nada que ver con el precio, de forma que el coeficiente de correlación entre ambas variables no sea significativo. Sin embargo, al combinar la variable con otras que sí están relacionadas con el precio, puede que en el modelo de regresión resultante sí obtenga un coeficiente estadísticamente significativo. Es como querer vincular el precio del queso a los kilómetros de carretera construidos en el país. Es posible que ambas variables pudieran formar parte del mismo modelo de regresión, pero ello no debe hacernos pensar que existe una relación entre ellas.

Para evitar estos efectos sobre nuestro modelos de regresión tendremos que llevar a cabo algunas de las siguientes acciones:

- Eliminar variables explicativas. Tendremos que identificar cuál es el foco que origina la multicolinealidad. Para ello deberemos examinar la relación existente entre las variables explicativas, de forma que podamos señalar qué variables están fuertemente correlacionadas entre sí y pueden, por lo tanto, ser el origen del problema.
- Aumentar el número de observaciones. En muchas ocasiones la multicolinealidad se presenta porque el número de observaciones es relativamente bajo, de forma que la fuerte relación entre variables se origina por la escasa representatividad de la muestra. Si ampliamos la misma con nuevos individuos, veremos como en muchos casos el problema de la multicolinealidad desparece o disminuye de manera relevante.

Empleemos un ejemplo financiero para ilustrar el problema generado por la multicolinealidad. Para ello hemos recabado información sobre la evolución en bolsa del índice de referencia bursátil español Ibex-35, compuesto por las 35 compañías cotizadas más importantes del país.

Junto con la evolución de su rentabilidad, también hemos incluido en nuestra base de datos otros dos títulos españoles de gran peso. Se trata de los bancos Santander y BBVA, con sede en España pero amplia presencia internacional. En la siguiente figura hemos representado la evolución en la rentabilidad para los 3 activos, durante el periodo 2008-2013. Resalta en el gráfico la crisis financiera de 2008, con origen en las hipotecas subprime y que afectó seriamente a los principales mercados financieros del mundo. En el caso español, no sólo se vio afectado el índice Ibex-35, sino muy especialmente el conjunto de las entidades financieras. A la crisis internacional se unió, en el caso español, el pinchazo de la burbuja inmobiliaria. El efecto sobre los balances y cuentas de resultados de los bancos no sólo lastró el valor bursátil de estas entidades, sino al conjunto de la economía nacional. De ahí que veamos un claro paralelismo entre la evolución del índice y la evolución de los dos ban-

cos, que a mediados de 2013 continuaban con unas pérdidas en su valoración respecto del comienzo del 2008 del 50% de su capitalización.

Figura 32. Evolución de la rentabilidad del índice Ibex-35 en color verde, y las entidades financieras Banco Santander Central Hispano (Santander – SAN.MC) en color rojo y Banco Bilbao Vizcaya Argentaria (BBVA.MC) en color azul

El objetivo es intentar construir un modelo predictivo de la rentabilidad diaria del Ibex. Para ello, obtenemos la rentabilidad diaria de BBVA y Santander para cada día, y en cada una de ellas colocamos la rentabilidad del Ibex en el día siguiente.

Aunque por la figura parezca que la rentabilidad de los 3 activos debe estar muy relacionada entre sí, al calcular el coeficiente de correlación entre las rentabilidades diarias vemos que el grado de relación entre el índice y los dos bancos es prácticamente insignificante. Si entramos en detalle, vemos como la rentabilidad diaria del BBVA sólo obtiene un coeficiente de correlación del 9,9% con la rentabilidad diaria del Ibex, mientras que para el Santander la correlación es aún menor: 4,2%. Difícilmente entonces podemos generar un modelo de regresión que explique la rentabilidad del índice a partir de la rentabilidad de los dos títulos.

Sin embargo, sí podemos ver cómo las rentabilidades diarias de BBVA y Santander están altamente correlacionadas entre sí: 93,9%.

Correlaciones

		IBEX	BBVA	SAN
IBEX	Correlación de Pearson	1	,099**	,042
	Sig. (bilateral)		,000	,134
	N	1288	1288	1288
BBVA	Correlación de Pearson	,099**	1	,939**
	Sig. (bilateral)	,000		,000
	N	1288	1288	1288
SAN	Correlación de Pearson	,042	,939**	1
	Sig. (bilateral)	,134	,000	
	N	1288	1288	1288

**. La correlación es significativa al nivel 0,01 (bilateral).

Figura 33. Evolución de la rentabilidad del índice Ibex-35 en color verde, y las entidades financieras Banco Santander Central Hispano (Santander – SAN.MC) en color rojo y Banco Bilbao Vizcaya Argentaria (BBVA.MC) en color azul

4.6 Cómo detectar el grado de multicolinealidad de nuestro modelos

En este epígrafe terminaremos de desarrollar nuestro modelo sobre el índice Ibex-35, explicando su rentabilidad a partir de dos únicos títulos: BBVA y Santander.

En la siguiente figura tenemos el resultado de regresar la rentabilidad diaria del Ibex frente a las rentabilidades diarias de BBVA y Santander. Ciertamente la capacidad explicativa del modelo es muy pobre, con un estadístico R2 corregido del 6,1%. El modelo en su conjunto no es significativo según la tabla ANOVA. Sin embargo, los coeficientes asociados a BBVA y Santander sí han resultado ser estadísticamente significativos para un nivel de confianza del 95%.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,306[a]	,093	,061	,013990128

a. Variables predictoras: (Constante), SAN, BBVA

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	,001	2	,001	2,887	,064[b]
	Residual	,011	56	,000		
	Total	,012	58			

a. Variable dependiente: IBEX

b. Variables predictoras: (Constante), SAN, BBVA

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.	Estadísticos de colinealidad	
		B	Error típ.	Beta			Tolerancia	FIV
1	(Constante)	,000	,002		-,063	,950		
	BBVA	,677	,287	,967	2,355	,022	,096	10,412
	SAN	-,676	,283	-,980	-2,387	,020	,096	10,412

a. Variable dependiente: IBEX

Figura 34. Modelo de regresión entre rentabilidad diaria del Ibex, y rentabilidades diarias de BBVA y Santander

Debe llamarnos la atención que ambos coeficientes hayan resultado significativos, cuando los coeficientes de correlación nos indicaban nula o escasa relación con la rentabilidad diaria del índice. Éste es un primer aviso de que podemos tener un problema de multicolinealidad. Pero es que además vemos cómo el coeficiente asociado al Banco Santander es negativo (-0,676). ¿Tiene sentido que las rentabilidad del índice y del Banco Santander se muevan en sentido contrario? La lógica, y el signo del coeficiente de correlación entre ambas variables, nos indican lo contrario.

Parece claro que estamos ante una situación donde las variables independientes están altamente relacionadas entre sí, el modelo no es significativo en su conjunto pero los coeficientes de las variables explicativas sí lo son, y el signo de uno de estos coeficientes no tiene el más mínimo sentido económico. El modelo está gritando: "¡Multicolinealidad!".

Para poder medir este efecto, y no quedarnos únicamente con la sospecha, podemos emplear el Factor de Inflado de la Varianza (FIV). Cuando este estadístico está por debajo del valor 10, podemos afirmar que no tenemos un problema de multicolinealidad. Cuando el valor está entre 10 y 30, el problema puede ser considerado leve, mientras que valores superiores a 30 informarían de un grave problema de multicolinealidad. En nuestro caso, ambos coeficientes tienen un FIV de 10,412, con lo que no podemos descartar un leve problema de multicolinealidad. Ello justificaría las incongruencias que hemos resaltado anteriormente.

Para superar esta situación podemos eliminar una de las variables. Si optáramos por esta situación, lógicamente el problema habría desaparecido totalmente, ya que ahora sólo tendríamos una variable explicativa.

La segunda posibilidad es ampliar el tamaño de la muestra, incluyendo un periodo de rentabilidades mayor al considerado inicialmente de 59 días. En la siguiente figura aparece el resultado de considerar un total de 1.288 rentabilidades diarias. Vemos cómo ahora el modelo en su conjunto sí es significativo, así como los coeficientes de BBVA y Santander. Los FIV también han bajado de 10. El único problema que no hemos conseguido eliminar es el singo nega-

tivo del coeficiente asociado al Banco Santander. Muy probablemente la solución que deberíamos acometer, finalmente, sería eliminar uno de los activos –el de menor correlación con el Ibex-.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,180[a]	,033	,031	,018585920

a. Variables predictoras: (Constante), SAN, BBVA

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	,015	2	,007	21,630	,000[b]
	Residual	,444	1285	,000		
	Total	,459	1287			

a. Variable dependiente: IBEX

b. Variables predictoras: (Constante), SAN, BBVA

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.	Estadísticos de colinealidad	
		B	Error típ.	Beta			Tolerancia	FIV
1	(Constante)	,000	,001		-,712	,477		
	BBVA	,353	,055	,512	6,399	,000	,118	8,505
	SAN	-,300	,055	-,439	-5,489	,000	,118	8,505

a. Variable dependiente: IBEX

Figura 35. Modelo de regresión entre rentabilidad diaria del Ibex, y rentabilidades diarias de BBVA y Santander. Ampliación de la muestra

Pongamos ahora un ejemplo más próximo al ámbito de la valoración. En la siguiente figura se encuentra el modelo de regresión múltiple obtenido al intentar explicar el precio (el logaritmo del precio en realidad) a partir de la superficie y el número de dormitorios. Como esta última variable es ordinal, se han creado 3 variables binarias para representar a las viviendas con 2, 3 y 4 dormitorios.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,748[a]	,559	,545	,13356

a. Variables predictoras: (Constante), NumDormit4, NumDormit2, Superficie, NumDormit3

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	2,961	4	,740	41,499	,000[b]
	Residual	2,337	131	,018		
	Total	5,298	135			

a. Variable dependiente: LogPrecio

b. Variables predictoras: (Constante), NumDormit4, NumDormit2, Superficie, NumDormit3

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados			Estadísticos de colinealidad	
		B	Error típ.	Beta	t	Sig.	Tolerancia	FIV
1	(Constante)	4,894	,081		60,384	,000		
	Superficie	,006	,000	,804	12,237	,000	,780	1,282
	NumDormit2	-,158	,084	-,300	-1,885	,062	,133	7,504
	NumDormit3	-,268	,082	-,676	-3,287	,001	,080	12,572
	NumDormit4	-,279	,086	-,624	-3,244	,001	,091	11,001

a. Variable dependiente: LogPrecio

Figura 35. Modelo de regresión múltiple entre el precio (variable dependiente), la superficie y el número de dormitorios (variables independientes). Detalle de los coeficientes FIV

En primer lugar, el modelo es significativo en su conjunto, tal y como indica la tabla ANOVA. Otra cuestión son los coeficientes de algunas variables explicativas, como el asociado a NumDormit2. La no significación de esta variable nos estaría informando de que el modelo no encuentra diferencias significativas en precio para viviendas de 1 o 2 dormitorios, considerando la misma superficie. Mientras que la significación de las variables NumDormit3 y NumDormit4 informarían de diferencias significativas en precio entre el nivel de referencia de 1 dormitorio, y las viviendas con 3 o 4 dormitorios.

Más allá de estas consideraciones, vemos cómo tenemos dos coeficientes con un valor FIV superior a 10. Se trata de las variables NumDormit3 (FIV=12,572) y NumDormit4 (FIV=11,001). Pueden indicar un posible problema de multicolinealidad, con lo que debemos plantear la eliminación de alguna de las variables de nuestro modelo. Para ello podemos, por ejemplo, llevar a cabo una regresión paso a paso.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,698[a]	,488	,484	,14231
2	,722[b]	,522	,515	,13800

a. Variables predictoras: (Constante), Superficie

b. Variables predictoras: (Constante), Superficie, NumDormit2

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	2,584	1	2,584	127,585	,000[b]
	Residual	2,714	134	,020		
	Total	5,298	135			
2	Regresión	2,765	2	1,382	72,590	,000[c]
	Residual	2,533	133	,019		
	Total	5,298	135			

a. Variable dependiente: LogPrecio

b. Variables predictoras: (Constante), Superficie

c. Variables predictoras: (Constante), Superficie, NumDormit2

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.	Estadísticos de colinealidad	
		B	Error típ.	Beta			Tolerancia	FIV
1	(Constante)	4,725	,047		101,471	,000		
	Superficie	,005	,000	,698	11,295	,000	1,000	1,000
2	(Constante)	4,680	,047		98,568	,000		
	Superficie	,006	,000	,737	12,034	,000	,957	1,045
	NumDormit2	,099	,032	,189	3,082	,002	,957	1,045

a. Variable dependiente: LogPrecio

Figura 36. Modelo de regresión múltiple paso a paso entre el precio (variable dependiente), la superficie y el número de dormitorios (variables independientes). Detalle de los coeficientes FIV

Vemos cómo sacrificando dos variables explicativas eliminamos cualquier indicio de multicolinealidad, y sin apenas sacrificar el valor del estadístico R^2 ajustado. Incluso el coeficiente asociado a la variable NumDormit2 parece tener ahora mejor interpretación económica.

Aunque la regresión paso a paso pueda parecer una posible solución para la construcción de modelos libres de multicolinealidad, veremos en el próximo capítulo cómo el análisis factorial facilita esta tarea. Además, esta técnica también permite reducir la dimensión de nuestros modelos de valoración, dejando que el menor número de variables explicativas posible explique el mayor porcentaje de variabilidad de la variable respuesta (el precio).

Capítulo 5. El análisis factorial (I)

5.1 Introducción

El análisis factorial es una de las técnicas que conforman el grupo de técnicas de reducción de la dimensión. Aunque en valoración también se aplica para obtener modelos libres de multicolinealidad, es una técnica que fue ideada originariamente para intentar minimizar el número de dimensiones de los modelos estadísticos.

Planteemos el problema de la siguiente forma. Supongamos que usted tiene una base de datos en la que ha recopilado información sobre un número significativo de viviendas (muestra). En la mayor parte de los casos la información sobre las viviendas se reduce a un pequeño número de variables, pues sabemos que obtener dicha información tiene un coste en tiempo y dinero elevado. Sin embargo, y sobre todo cuando hablamos de sociedades de tasación y no de tasadores particulares, la muestra de viviendas puede llegar a incluir un número relativamente importante de variables. Ciertamente es posible encontrar tasadoras que reúnen más de 50 variables por cada una de las viviendas que tienen en sus registros.

Cuando el número de variables es muy elevado es importante distinguir el polvo de la paja; esto es, discernir qué variables son realmente importantes en el proceso de configuración de los precios, y cuáles actúan como meras comparsas sin aportar nada relevante sobre el resto.

Mucho esfuerzo, dinero y tiempo se habrían ahorrado la mayoría de las tasadoras si comprendieran que el precio está, finalmente, configurado a partir de un pequeño número de variables relevantes y que, por lo tanto, no es necesario recopilar 50 datos de cada una de las viviendas para poder emitir un informe de valoración con mínimas garantías. En este capítulo mostraremos cómo los modelos no son necesariamente mejores cuanto más variables consideran, sino que llegado un límite la adición de nuevas variables no aporta nada al resultado.

Imaginemos el caso de una valoración inmobiliaria en el que el tasador trabaja sobre un total de 50 variables. Las más habituales serían la superficie, el número de dormitorios, el número de cuartos de baño, la calidad de la construcción, el número de viviendas en el edificio, la planta en la que se sitúa la vivienda, el número de ascensores, el entorno comercial, el nivel económico del entorno, etc.

Si nuestra pretensión es trabajar con modelos parsimoniosos, pero sin sacrifica a cambio capacidad explicativa y predictiva de nuestros modelos, tenemos la posibilidad de emplear el análisis factorial. La idea es la siguiente: del conjunto de variables mencionado, podemos encontrar claras relaciones entre algunas de ellas. Por ejemplo, es lógico que la superficie y el número de dormitorios puedan estar directamente relacionados; o que la altura del edificio también guarde relación con la presencia o ausencia de ascensor, o con el número de ascensores.

El análisis factorial tiene por objetivo encontrar asociaciones entre variables observables, que nos informen de dimensiones que no podemos observar directamente. Por ejemplo, las variables superficie, número de dormitorios y número de cuartos de baño, nos están informando sobre la "dimensión" de la vivienda. Esto es, las 3 variables formarían parte de una variable de mayor entidad, que las aglutinara a todas ellas, pero que no fuera directamente observable: la dimensión. Por lo tanto, y a partir de unas variables que sí podemos observar y cuantificar, el análisis factorial pretende inferir y cuantificar nuevas variables inobservables, pero que sabemos resumen el comportamiento del conjunto de variables. Dichas variables inobservables reciben el nombre de factores latentes.

Los factores latentes son, finalmente, esas variables que el tasador no puede incluir directamente en sus bases de datos, pero de las que sí que tiene certeza de su existencia, pues entre las variables que sí están recopiladas en la base de datos tenemos variables que se relacionan directamente con esos factores latentes.

El análisis factorial permite trabajar con factores latentes, en lugar de con las variables originales, pero ¿dónde está la ventaja de su uso? Pues que en la mayoría de los casos el número de factores latentes será muy inferior al número de variables, lo que permite reducir enormemente la dimensión del problema. Seremos capaces de trabajar con modelos de valoración de no más de 5 o 6 variables, en lugar de tener que incluir 50 variables, y ello sin renunciar a un modelo con un buen ajuste.

5.2 Las bases del análisis factorial

El análisis factorial fue propuesto inicialmente por un psicólogo ingles, Charles Spearman, que decidió plantear su tesis doctoral sobre la medición de la inteligencia en los individuos. Para ello seleccionó a un grupo de alumnos de los que recopiló las notas en diferentes asignaturas. Spearman supuso que las calificaciones en cada asignatura dependían de dos elementos independientes: la inteligencia general del alumno, y la aptitud específica del alumno para esa asignatura en concreto.

Por lo tanto, Spearman contaba con variables que pudo observar y medir para cada uno de sus alumnos: las calificaciones en las distintas asignaturas; y a partir de estas mediciones quiso cuantificar la inteligencia, que no era observable directamente: factor latente.

Para comprender la forma en que llevó a cabo su análisis, pongamos un ejemplo ilustrativo. Supongamos que, para un grupo de alumnos, se ha recopilado información de las calificaciones obtenidas en 6 asignaturas: Matemáticas (M), Física (F), Química (Q), Historia (H), Lengua (L) e Inglés (E). Supongamos también que la calificación en cada una de estas asignaturas se puede obtener a partir de dos elementos independientes: la inteligencia general (I) y la aptitud específica del alumno a dicha asignatura. En ese caso, podríamos llegar a obtener el siguiente sistema de ecuaciones:

$$M = 0.90I + A_M$$
$$F = 0.85I + A_F$$
$$Q = 0.75I + A_Q$$
$$H = 0.60I + A_H$$
$$L = 0.70I + A_L$$
$$E = 0.65I + A_E$$

La primera ecuación nos diría que la calificación en Matemáticas se obtiene como 0,9 veces el factor inteligencia, más la aptitud específica de cada alumno hacia esa materia; la calificación en Física se obtiene multiplicando el factor de inteligencia por 0,85, más la aptitud específica del alumno hacia la Física; etc.

Suponiendo que las variables estuvieran normalizadas, algo que ahora no va a influir en la explicación de nuestro caso, el coeficiente de cada una de estas ecuaciones medirá el grado de relación entre el factor de inteligencia general y la calificación de la asignatura correspondiente. Es decir, la calificación en Matemáticas y la inteligencia general tendrán un coeficiente de correlación del 90%. Dicho coeficiente recibe el nombre de carga factorial.

Por lo tanto, y a la vista de estas 6 ecuaciones, la asignatura que mayor correlación tiene con la inteligencia de los alumnos es la de Matemáticas, mientras que la menos vinculada a la inteligencia es la Historia.

Puesto que ya conocemos de capítulos anteriores el análisis de regresión, podemos plantear la siguiente pregunta. Dado que la correlación entre la calificación de Matemáticas y la inteligencia general es del 90%, ¿qué R^2 obten-

dríamos si explicáramos mediante regresión simple la nota en Matemáticas a partir de la inteligencia? Como sabemos que el estadístico R^2 es el cuadrado del coeficiente de correlación, la respuesta sería un R^2 del 81%. Por lo tanto, hay un 19% de variabilidad en las notas de Matemáticas de los alumnos que no puede explicarse por su inteligencia general, sino que depende únicamente de la aptitud que cada uno de los alumnos tiene hacia esa asignatura. En Historia ocurre todo lo contrario: un 36% de la variabilidad en las calificaciones se explica por la inteligencia general de los alumnos (el cuadrado de 0,6), mientras que el 64% restante se debe a la aptitud específica de los alumnos hacia esa asignatura.

Al cuadrado de la carga factorial, que acabamos de ver coincide con el R^2 de la regresión entre las dos variables, se le conoce como comunalidad. Informa sobre qué porcentaje de variabilidad de la correspondiente variable viene explicada por el factor latente; mientras que al valor $1-R^2$ se le denomina varianza única o específica.

En la siguiente tabla hemos representado el coeficiente de correlación entre la inteligencia general y cada una de las calificaciones, así como la comunalidad y la varianza específica.

	Carga factorial (correlación)	Comunalidad	Varianza específica
Matemáticas (M)	0,90	0,81	0,19
Física (F)	0,85	0,72	0,28
Química (Q)	0,75	0,56	0,44
Historia (H)	0,60	0,36	0,64
Lengua (L)	0,70	0,49	0,51
Inglés €	0,65	0,42	0,58
Total		3,36	2,64

Tabla 1. Cargas factoriales, comunalidades y varianzas específicas

En la última fila aparece la suma de ambas componentes. La suma de estos parciales tiene que coincidir con el número de variables: 6. De esta forma, podemos afirmar que el factor inteligencia explicaría en promedio el 56% de la variabilidad de las calificaciones. El 44% restante de variabilidad no podría ser explicado por la inteligencia general de los alumnos, sino que dependería exclusivamente de su aptitud hacia cada una de las materias.

$$Comunalidad\ promedio = \frac{3{,}36}{6} = 0{,}56$$

Supongamos ahora que decidimos añadir un nuevo factor, además del de inteligencia general. De momento no sabemos la interpretación de este factor, pero imaginemos que las ecuaciones obtenidas con la presencia de dos factores latentes son las siguientes:

$$M = 0.85I + 0{,}3J + A_M$$
$$F = 0.8I + 0{,}3J + A_F$$
$$Q = 0.7I + 0{,}35J + A_Q$$
$$H = 0.15I + 0{,}85J + A_H$$
$$L = 0.35I + 0{,}8J + A_L$$
$$E = 0.4I + 0{,}7J + A_E$$

Con esta nueva configuración de cargas factoriales y número de factores latentes, las comunalides y varianzas específicas deben actualizarse. En la siguiente tabla aparecen los valores.

Tabla 2. Cargas factoriales, comunalidades y varianzas específicas para un modelo con dos factores latentes

	Carga factorial		Comunalidad		Varianza específica	
Factor	I	J	I	J	I	J
Matemáticas (M)	0,85	0,3	0,7225	0,09	0,2775	0,91
Física (F)	0,8	0,3	0,64	0,09	0,36	0,91
Química (Q)	0,7	0,35	0,49	0,1225	0,51	0,8775
Historia (H)	0,15	0,85	0,0225	0,7225	0,9775	0,2775
Lengua (L)	0,35	0,8	0,1225	0,64	0,8775	0,36
Inglés €	0,4	0,7	0,16	0,49	0,84	0,51
Total			2,1575	2,155	3,8425	3,845

Vemos cómo la inclusión de un nuevo factor ha modificado los pesos de las cargas factoriales en el factor de inteligencia general. Por ejemplo, la asignatura de Matemáticas cargaba un 0,9 en el primer modelo unifactorial, mientras que ahora su peso es de 0,85.

A simple vista también podemos comprobar que algunas asignaturas han obtenido mayores cargas factoriales en el primer factor latente (I) que en el segundo (J). Es el caso de Matemáticas, Física y Química. Sin embargo, las asignaturas de Historia, Lengua e Inglés tienen una mayor carga factorial, y por tanto correlación, con el segundo factor latente (J). Esto nos ayuda a interpretar el significado de ambos factores. El primero I está vinculado a asignaturas de ciencias, con un alto grado de entendimiento cuantitativo. El segundo factor J se vincula a las asignaturas de letras, más vinculadas a aspectos verbales.

Así, podríamos concluir que:

$$I = Inteligencia\ Cuantitativa$$

$$J = Inteligencia\ Verbal$$

Es muy importante destacar el hecho de que asumimos que ambos factores son independientes. Esto significa que un alumno con una elevada inteligencia cuantitativa no tiene por qué distinguirse por una inteligencia verbal mayor que el resto, ni tampoco inferior. El modelo asume que ambas inteligencias no están conectadas (correlacionadas) entre sí.

En la misma tabla 2 también hemos recogido la comunalidad de cada variable con los dos factores latentes. Al suponer que los factores son independientes entre sí, podemos sumar comunalidades de una misma variable para medir la forma en que los dos factores explican cada una de las calificaciones. Por ejemplo, para el caso de Matemáticas la comunalidad con el factor de inteligencia cuantitativa es del 72,25%, y con el factor de inteligencia verbal es del 9%. La comunalidad conjunta será del 81,25%, lo que indica que entre los dos factores se explica el 81,25% de la variabilidad encontrada en las calificaciones de Matemáticas. El resto de variabilidad vendrá explicado por la aptitud de los alumnos en concreto a esa asignatura, o por otras causas que no vienen recogidas en estos dos factores.

¿Ha mejorado el segundo modelo la explicación de la variabilidad en las calificaciones respecto del primero? Para ello sólo tenemos que sumar la comunalidad conjunta de ambos factores: 2,1575 + 2,155 = 4,3125. A partir de este valor, podemos calcular la comunalidad conjunta del modelo:

$$Comunalidad\ promedio = \frac{4{,}3125}{6} = 0{,}7188$$

Por lo tanto, hemos pasado de un modelo unifactorial que explicaba el 56% de la variabilidad en las calificaciones, a otro bifactorial que explica el 71,88%. La mejora es evidente.

Pues bien, ya sólo queda saber cómo obtener las cargas factoriales para entender un poco mejor el funcionamiento del análisis factorial. Entre otras cosas.

La obtención de estas cargas necesitaría algún conocimiento sobre álgebra, matrices de varianzas-covarianzas, y cálculo de vectores y valores propios. Afortunadamente, el paquete estadístico SPSS nos ofrece el resultado obviando estos cálculos intermedios.

5.3 Un ejemplo: la valoración de viviendas en la ciudad de Valencia (España)

El software SPSS incorpora el análisis factorial entre su baterías de análisis estadísticos. Utilizaremos esta opción para intentar obtener un conjunto de factores latentes que agrupe a las variables explicativas del precio. Debemos resaltar que no utilizamos el análisis factorial para obtener de manera directa un modelo de valoración. Hacemos uso de él para identificar una serie de factores latentes que resuman un número importante de variables explicativas del precio. Veremos más adelante como el número de factores limitará el número de variables explicativas en nuestro modelo, siendo ésta la forma en que el análisis factorial participa en la construcción del modelo de valoración. Eso sí, el modelo final será obtenido como hasta ahora, a través del análisis de regresión múltiple.

En el modelo que vamos a presentar a continuación hemos utilizado una muestra compuesta por 126 viviendas. Todas ellas pertenecen a un mismo código postal de la ciudad de Valencia (España). Dicho código es el 46001, localizado en el centro de la Ciudad, con importantes dotaciones de medios de transporte, comercio y zonas de ocio.

Antes de explicar qué variables empleamos en el análisis factorial, señalemos qué variable no va a ser incluida. Se trata del precio, la variable que luego queremos explicar mediante un modelo de regresión. El motivo es que tratamos de reducir la dimensión de las variables explicativas, intentando sacar a la luz los factores latentes que se encuentran detrás de ellas. Si incluyéramos el precio, éste se agruparía dentro de algún factor con variables que se encontraran en su misma dimensión, pero sin aportar mayor información útil para el

valorador. El proceso de extracción de factores latentes debe omitir al precio, pues a priori todas las variables consideradas en el proceso están relacionadas con él, y podrían configurar un único factor latente. Piense en el ejemplo de las calificaciones en las asignaturas. Si se añadiera un examen fin de curso que incluyera preguntas de todas las asignaturas, el resultado acabaría promediando las calificaciones individuales de cada asignatura, y podríamos extraer un factor en el que la componente más destacada fuera el examen final, por delante de las asignaturas particulares.

Las variables que vamos a considerar para este análisis son las siguientes:

- Antigüedad: se refiere a la antigüedad del edificio, medida en años.
- numVivEdif: número de viviendas en el edificio.
- numAscEdif: número de ascensores en el edificio. En la muestra aparecen edificios sin ascensor, con un ascensor y con dos ascensores. A partir de esta información, y como pensamos que es posible que el precio de pasar de 0 ascensores a 1, pueda ser distinto de pasar de 1 a 2, creamos dos variables binarias: ascensor1 (edificios con un ascensor) y ascensor2 (edificios con dos ascensores). El nivel base o referencia serán los edificios sin ascensor.
- numPlantas: número de plantas o niveles del edificio.
- plantaVivienda: número de planta en que se sitúa la vivienda.
- numDormitorios: número de dormitorios. Como en el caso del número de ascensores, tenemos varios niveles que debemos transformar en variables binarias. Tomamos como nivel de referencia las viviendas con un sólo dormitorio, creando las variables: numDormit2, numDormit3, numDormit4 y numDormit5omas. La última variable se activa cuando la vivienda tiene 5 o más dormitorios.
- numBanyos: número de cuartos de baño / aseos. Se ha tomado como nivel de referencia las viviendas con un cuarto de baños, por lo que se han creado las variables binarias: numBaños2 y numBaños3omas.
- superficie: superficie útil de la vivienda expresada en metros cuadrados.
- supTerrazaDesc: superficie de terraza descubierta.
- calidadConstructiva: indica la calidad de la construcción y los acabados. Se han identificado 3 niveles, dejando como nivel base el de las viviendas con una calidad constructiva media. Las dos variables binarias que recogen los otros dos casos son calidadConstrAlta (calidad constructiva alta) y calidadConstrVivSocial (calidad constructiva de viviendas sociales).
- entornoComercial: el entorno comercial viene determinado por los establecimientos comerciales de la zona. Se distinguen 3 niveles, tomando como referencia el entorno comercial bueno. Se han creado dos variables binarias para los otros dos niveles: entComExcel (en-

torno comercial excelente) y entComBasico (entorno comercial básico).

- nivelRenta: como nivel de renta se han considerado 3 posibilidades: nivel de renta alto (variable binaria rentaAlta), nivel de renta intermedio (nivel de referencia) y nivel de renta bajo (variable binaria rentaBaja).
- densidadPoblacion: la densidad de población se ha resumido en los niveles alto (variable binaria densidadAlta), nivel intermedio (nivel de referencia) y nivel bajo (variable binaria densidadBaja).

Al llevar a cabo el análisis factorial sobre la muestra compuesta por 126 viviendas, el programa SPSS lanza la siguiente advertencia:

Advertencia

Hay menos de dos casos, al menos una de las variables tiene varianza cero, hay sólo una variable en el análisis o no se han podido calcular los coeficientes de correlación para todos los pares de variables. No se calculará ningún estadístico más.

Figura 37. Advertencia de SPSS al encontrar en el análisis factorial una variable sin varianza

Hemos seleccionado este ejemplo con el objetivo de alertar sobre estas situaciones. Para poder llevar a cabo el análisis factorial, es necesario que la matriz de varianzas-covarianzas cumpla una serie de condiciones. Sin entrar en demasiados detalles matemáticos, deberemos comprobar que no concurren ninguna de estas condiciones:

- Una o más variables tienen varianza cero. Esto significa que todos sus valores son iguales, en este caso para todas las viviendas.
- Una variable tiene dependencia perfecta respecto de otra u otras variables. Si al llevar a cabo una regresión entre una variable (dependiente) y otra u otras variables de la base de datos (independiente/s) obtenemos un R^2 del 100%, entonces la dependencia es perfecta e impedirá que podamos obtener ningún resultado con el análisis factorial.

En nuestro ejemplo estamos dentro del primer caso. Al obtener la tabla de frecuencias de la variable densidadBaja vemos que ninguna vivienda se encuentra en una zona con densidad baja de población; de hehco, todas las viviendas están en una zona similar por compartir código postal. De ahí que SPSS nos haya alertado de esa situación.

densidadBaja

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	,00	126	100,0	100,0	100,0

Figura 38. Tabla de frecuencias para la variable densidadBaja

Para poder realizar el análisis factorial tendremos que excluir esta variable.

El segundo caso, el de dependencia perfecta, suele ser menos habitual. No obstante, sí podremos encontrar ejemplos cuando el tamaño de la muestra sea muy reducido. En esas situaciones tendremos que desprendernos de una de las variables que originan esa dependencia perfecta para que el análisis factorial pueda ser llevado a cabo.

Una vez identificado el problema de nuestra muestra y eliminada la variable densidadBaja, los resultados del análisis factorial son los que se muestran en la siguiente figura.

Matriz de componentes rotados[a]

	Componente							
	1	2	3	4	5	6	7	8
numPlantas	,815	,222	,063	,069	,132	-,053	-,068	,041
plantaVivienda	,757	-,060	,033	,038	,043	,092	-,058	,130
numVivEdif	,653	,045	-,056	-,039	-,027	-,136	-,002	-,027
calidadConstrAlta	,097	,687	,115	-,088	-,075	-,093	,062	,230
calidadConstrVivSocial	-,005	-,617	,019	,056	,002	,006	,306	,028
numBaños2	,106	,587	-,055	,284	,419	,190	-,014	-,071
numBaños3omas	,035	-,024	,815	-,021	-,011	-,065	-,022	,036
numDormit5omas	-,043	,043	,801	-,023	-,051	,044	-,023	,043
superficie	,113	,331	,585	,300	,449	,143	-,020	,016
numDormit2	-,036	,040	-,101	-,905	-,168	-,009	-,032	-,044
numDormit3	,016	-,043	-,156	,799	-,545	,019	-,022	-,088
numDormit4	,044	,003	,020	-,022	,924	-,006	,064	,098
supTerrazaDesc	,013	-,012	-,011	-,004	,098	,019	-,028	-,054
ascensor1	,024	,135	,052	,082	,173	,855	-,115	-,032
ascensor2	,450	,287	,071	,080	,080	-,685	-,039	-,073
rentaBaja	,002	-,122	-,024	-,006	-,047	-,013	,774	,002
entComBasico	-,115	,000	-,030	,020	,012	-,060	,702	-,066
entComExcel	,078	,038	,011	,002	-,040	,018	-,114	,841
rentaAlta	,172	,454	,178	-,055	-,088	,048	,110	,572
densidadAlta	,171	,332	,044	-,081	-,141	,286	,299	-,364

Método de extracción: Análisis de componentes principales.
Método de rotación: Normalización Varimax con Kaiser.

a. La rotación ha convergido en 8 iteraciones.

Varianza total explicada

	Suma de las saturaciones al cuadrado de la rotación		
Componente	Total	% de la varianza	% acumulado
1	1,983	9,917	9,917
2	1,802	9,010	18,928
3	1,749	8,747	27,675
4	1,670	8,349	36,024
5	1,663	8,313	44,337
6	1,389	6,946	51,283
7	1,334	6,671	57,954
8	1,282	6,409	64,363

Método de extracción: Análisis de Componentes principales.

Figura 39. Resultados del análisis factorial

Aunque SPSS ofrece más información en otras tablas de resultados, en la anterior figura sólo hemos incluido las dos que nos parecen más importantes.

En la primera de ellas se recogen las cargas factoriales del conjunto de variables consideradas sobre los factores extraídos por SPSS. En la segunda tabla se recogen algunas características de los 8 factores.

Entrando en detalle sobre la primera tabla, debemos destacar que hemos seleccionado la rotación VARIMAX. La rotación de los factores originales se hace para resaltar las diferencias entre ellos, de forma que se puedan identificar más fácilmente las variables con sus factores latentes correspondientes. SPSS ha extraído 8 factores, por lo que tenemos 8 columnas de cargas factoriales.

Examinando las cargas factoriales, vemos cómo las variables numPlantas, plantaVivienda y numVivEdif obtienen sus mayores coeficientes, en valor absoluto, en el primer factor. Sus cargas factoriales son, respectivamente, 0,815, 0,757 y 0,653. Esto indica que dichas variables están altamente relacionadas con el primer factor latente. Vemos como la correlación con el segundo factor es mucho menor, con cargas factoriales de 0,222, -0,060 y 0,045, respectivamente. Y lo mismo ocurre con el resto de factores. Todo esto nos indica que las 3 variables están claramente vinculadas al factor latente.

De igual forma procedemos con el resto de variables. Por ejemplo, las variables calidadConstrAlta, calidadConstrVivSocial y numBaños2 obtienen las mayores cargas factoriales, en valor absoluto, con el segundo factor latente. Mientras que la correlación con el segundo factor de calidadConstrAlta (0,687) y numBaños2 (0,587) es positiva, la variable calidadConstrVivSocial (-0,617) es negativa. Esto indicaría que las dos primeras variables están relacionadas positivamente con el segundo factor latente, mientras que la variable calidadConstrVivSocial tiene una relación inversa. Es lógico pensar que sea así, ya que si se tiene una calidad constructiva alta, entonces se está en el extremo opuesto a tener una calidad constructiva propia de las viviendas sociales; y viceversa. Examinando el signo de la variable numBaños2 podemos afirmar que las viviendas con calidad constructiva alta suelen tener dos baños, justamente lo contrario de las viviendas con calidad constructiva de viviendas sociales.

Continuando con este mismo procedimiento, la agrupación de variables en factores latentes sería la siguiente:

- Factor 1: agrupa a las variables numPlantas (0,815), plantaVivienda (0,757) y numVivEdif (0,653). Todas estas variables están relacionadas positivamente con el factor. Claramente es una dimensión que nos informa sobre el tamaño del edificio, permitiendo diferenciar edificios con muchas viviendas y un número elevado de plantas respecto de aquellos otros más pequeños.

- Factor 2: calidadConstrAlta (0,687), calidadConstrVivSocial (-0,617) y numbaños2 (0,587). El factor se relaciona con la dimensión de la calidad constructiva del edificio. Llama la atención la inclusión en el factor de la variable numbaños2, que vincularemos a las viviendas con calidad constructiva alta. No obstante, la inclusión de variables que a priori no guardan relación con la dimensión es algo habitual, sobre todo ocupando las últimas posiciones en el factor según su carga factorial. Esto significa que aunque asociamos las viviendas con dos baños a este factor, dicha variable es la menos correlacionada con el factor latente, viéndose superada por las otras dos variables sobre la calidad constructiva.
- Factor 3: numBaños3omas (0,815), numDormit5omas (0,801) y superficie (0,585). Claramente el significado del factor latente está vinculado a la dimensión de la vivienda. Tenemos entonces dos factores latentes que hablan de la dimensión: el factor 1 sobre la dimensión del edificio, y el factor 3 sobre la dimensión de la vivienda en particular.
- Factor 4: numdormit2 (-0,905) y numdormit3 (0,799). Al contraponer dos variables relacionadas con el número de dormitorios, se trata de un factor latente que distingue entre las viviendas de tamaño medio. A diferencia del factor 3 que captura las viviendas grandes, el factor 4 se centra en las viviendas de tamaño intermedio.
- Factor 5: numDormit4 (0,924) y supTerrazaDesc (0,098). Aunque tengamos dos variables, la carga factorial nos indica que la variable más relevante en este factor es numDormit4. De esta forma, volvemos a tener un factor que informa sobre la dimensión de las viviendas, a mitad camino entre los factores 3 y 4.
- Factor 6: ascensor1 (0,855) y ascensor2 (-0,685). Factor latente que recoge la diferencia entre los edificios con uno y dos ascensores. Si el edificio tiene un ascensor, obtendrá una alta correlación con el factor; si tienes dos ascensores, la correlación sería negativa. Es interesante comprobar la carga factorial de la variable ascensor2 con el factor 1: 0,450. Esto indica que los edificios grandes normalmente tendrán dos ascensores.
- Factor 7: rentaBaja (0,774) y entComBasico (0,702). Factor que informa sobre un entorno de la vivienda con renta baja y poco comercio.
- Factor 8: entornoComExcel (0,841), rentaAlta (0,572) y densidadAlta (-0,364). Factor que indica la pertenencia a una zona con gran desarrollo comercial, renta alta y poca densidad poblacional. También podemos observar que la variable rentaAlta tiene una correlación con el factor 2 nada desdeñable: 0,454. Indicaría que el factor relacionado con la calidad constructiva está muy vinculado a la zona definida por el factor 8.

La construcción de los factores se lleva a cabo de forma que se busca independencia entre los mismos. Esto es, aunque podamos ver variables que están relacionados con varios factores, la formación de los mismos se lleva a cabo para que capturen dimensiones lo más independientes entre sí.

Justamente es esta cualidad la que hace del análisis factorial una técnica muy interesante para el valorador. Al capturar dimensiones que son independientes entre sí, podemos construir modelos de valoración que sólo incluyan una variable por dimensión. De esta manera estaremos minimizando el posible problema de multicolinealidad, que se da en aquellos casos en que combinamos variables altamente correlacionadas entre sí.

A modo de ejemplo, supongamos que queremos incluir en nuestro modelo una variable que recoja la dimensión del tamaño de la vivienda. En ese caso tendremos que escoger alguna variable del factor 3, pero sólo una, ya que todas las variables del factor informan sobre la misma dimensión: tamaño de la vivienda. Incluir más de una variable por factor implicaría solapar variables que informan sobre lo mismo, sin añadir nada relevante y creando problemas de multicolinealidad.

Una vez identificados los factores latentes que resumen toda la información sobre las viviendas, podemos construir un modelo seleccionando una variable como representante de cada factor. Lo más habitual es seleccionar como representante aquella variable que mayor carga factorial, en valor absoluto, tiene dentro de su factor. En la siguiente figura aparecen recuadradas las cargas factoriales de cada variable en su correspondiente factor, de las que escogeremos las que ocupan las primeras posiciones en su respectivo factor.

Así, las variables representantes de cada factor sería las siguientes:

- Factor 1: numPlantas
- Factor 2: calidadConstrAlta
- Factor 3: numBaños3omas
- Factor 4: numDormit2
- Factor 5: numDormit4
- Factor 6: ascensor1
- Factor 7: rentaBaja
- Factor 8: entComExcel

Matriz de componentes rotados[a]

	Componente							
	1	2	3	4	5	6	7	8
numPlantas	,815	,222	,063	,069	,132	-,053	-,068	,041
plantaVivienda	,757	-,060	,033	,038	,043	,092	-,058	,130
numVivEdif	,653	,045	-,056	-,039	-,027	-,136	-,002	-,027
calidadConstrAlta	,097	,687	,115	-,088	-,075	-,093	,062	,230
calidadConstrVivSocial	-,005	-,617	,019	,056	,002	,006	,306	,028
numBaños2	,106	,587	-,055	,284	,419	,190	-,014	-,071
numBaños3omas	,035	-,024	,815	-,021	-,011	-,065	-,022	,036
numDormit5omas	-,043	,043	,801	-,023	-,051	,044	-,023	,043
superficie	,113	,331	,585	,300	,449	,143	-,020	,016
numDormit2	-,036	,040	-,101	-,905	-,168	-,009	-,032	-,044
numDormit3	,016	-,043	-,156	,799	-,545	,019	-,022	-,088
numDormit4	,044	,003	,020	-,022	,924	-,006	,064	,098
supTerrazaDesc	,013	-,012	-,011	-,004	,098	,019	-,028	-,054
ascensor1	,024	,135	,052	,082	,173	,855	-,115	-,032
ascensor2	,450	,287	,071	,080	,080	-,685	-,039	-,073
rentaBaja	,002	-,122	-,024	-,006	-,047	-,013	,774	,002
entComBasico	-,115	,000	-,030	,020	,012	-,060	,702	-,066
entComExcel	,078	,038	,011	,002	-,040	,018	-,114	,841
rentaAlta	,172	,454	,178	-,055	-,088	,048	,110	,572
densidadAlta	,171	,332	,044	-,081	-,141	,286	,299	-,364

Método de extracción: Análisis de componentes principales.
Método de rotación: Normalización Varimax con Kaiser.

a. La rotación ha convergido en 8 iteraciones.

Figura 40. Agrupación de variables por factores

Al regresar el logaritmo del precio respecto de este grupo de 8 variables obtenemos los resultados de la figura 41.

Lo primero que llama la atención es que, pese a haber combinado un número elevado de variables (8), el estadístico R^2 corregido es alarmantemente bajo: 33,9%. Claramente insuficiente como para considerar la función obtenida en la práctica profesional.

Además, no todas las variables representativas de los factores han resultado ser estadísticamente significativas. Tengamos en cuenta que un modelo construido a partir de sólo 126 viviendas no puede incluir muchas variables explicativas, sino 3 o 4 a lo sumo. Por lo tanto, tendremos que reducir el número de variables atendiendo a su significación estadística.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,618[a]	,381	,339	,17823

a. Variables predictoras: (Constante), entComExcel, numBaños3omas, ascensor1, rentaBaja, calidadConstrAlta, numPlantas, numDormit2, numDormit4

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	2,292	8	,287	9,020	,000[b]
	Residual	3,716	117	,032		
	Total	6,009	125			

a. Variable dependiente: logPrecio

b. Variables predictoras: (Constante), entComExcel, numBaños3omas, ascensor1, rentaBaja, calidadConstrAlta, numPlantas, numDormit2, numDormit4

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	5,283	,060		88,341	,000
	numPlantas	,010	,008	,094	1,245	,216
	calidadConstrAlta	,026	,040	,048	,657	,513
	numBaños3omas	,312	,084	,278	3,724	,000
	numDormit2	-,104	,036	-,229	-2,924	,004
	numDormit4	,078	,048	,128	1,620	,108
	ascensor1	,156	,035	,338	4,505	,000
	rentaBaja	-,050	,181	-,020	-,274	,785
	entComExcel	,031	,033	,070	,942	,348

a. Variable dependiente: logPrecio

Figura 41. Análisis de regresión entre el logaritmo del precio y los representantes de cada factor

Tras eliminar recursivamente las variables con peor significación estadística, el modelo resultante mejora incluso el R^2 ajustado, y con un número razonable de variables dado el tamaño de la muestra.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,604[a]	,365	,344	,17752

a. Variables predictoras: (Constante), ascensor1, numBaños3omas, numDormit4, numDormit2

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	2,196	4	,549	17,419	,000[b]
	Residual	3,813	121	,032		
	Total	6,009	125			

a. Variable dependiente: logPrecio

b. Variables predictoras: (Constante), ascensor1, numBaños3omas, numDormit4, numDormit2

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	5,360	,033		161,243	,000
	numBaños3omas	,327	,083	,292	3,959	,000
	numDormit2	-,101	,035	-,222	-2,865	,005
	numDormit4	,094	,047	,155	2,003	,047
	ascensor1	,157	,034	,340	4,581	,000

a. Variable dependiente: logPrecio

Figura 42. Análisis de regresión entre el logaritmo del precio y variables estadísticamente significativas

No obstante, sigue siendo un modelo con una capacidad explicativa muy pobre y aunque eliminásemos datos anómalos no mejoraríamos lo suficiente como para alcanzar el mínimo que debemos proponernos en cualquier modelo de valoración.

Otra posibilidad que podemos plantearnos es cambiar las variables representantes de cada factor. En el anterior ejemplo escogíamos como representantes aquellas que tenían la mayor carga factorial, en valor absoluto, dentro de su factor. Haciendo variaciones entre las variables seleccionadas podemos mejorar de forma considerable el R^2 ajustado. Eso sí, teniendo siempre la precaución de no incluir en el mismo modelo dos variables que estén agrupadas dentro de un mismo factor.

En la siguiente figura aparecen dos pequeños cambios que van a propiciar una mejora sustancial en el estadístico R^2 corregido. Se han reemplazado las va-

riables representantes de los factores 3 y 8, escogiendo ahora a la superficie y renta alta como variables seleccionadas para el modelo de valoración.

Matriz de componentes rotados[a]

	Componente							
	1	2	3	4	5	6	7	8
numPlantas	,815	,222	,063	,069	,132	-,053	-,068	,041
plantaVivienda	,757	-,060	,033	,038	,043	,092	-,058	,130
numVivEdif	,653	,045	-,056	-,039	-,027	-,136	-,002	-,027
calidadConstrAlta	,097	,687	,115	-,088	-,075	-,093	,062	,230
calidadConstrVivSocial	-,005	-,617	,019	,056	,002	,006	,306	,028
numBaños2	,106	,587	-,055	,284	,419	,190	-,014	-,071
numBaños3omas	,035	-,024	,815	-,021	-,011	-,065	-,022	,036
numDormit5omas	-,043	,043	,801	-,023	-,051	,044	-,023	,043
superficie	,113	,331	,585	,300	,449	,143	-,020	,016
numDormit2	-,036	,040	-,101	-,905	-,168	-,009	-,032	-,044
numDormit3	,016	-,043	-,156	,799	-,545	,019	-,022	-,088
numDormit4	,044	,003	,020	-,022	,924	-,006	,064	,098
supTerrazaDesc	,013	-,012	-,011	-,004	,098	,019	-,028	-,054
ascensor1	,024	,135	,052	,082	,173	,855	-,115	-,032
ascensor2	,450	,287	,071	,080	,080	-,685	-,039	-,073
rentaBaja	,002	-,122	-,024	-,006	-,047	-,013	,774	,002
entComBasico	-,115	,000	-,030	,020	,012	-,060	,702	-,066
entComExcel	,078	,038	,011	,002	-,040	,018	-,114	,841
rentaAlta	,172	,454	,178	-,055	-,088	,048	,110	,572
densidadAlta	,171	,332	,044	-,081	-,141	,286	,299	-,364

Método de extracción: Análisis de componentes principales.
Método de rotación: Normalización Varimax con Kaiser.

a. La rotación ha convergido en 8 iteraciones.

Figura 43. Cambio de variables representantes en los factores 3 y 8

Con este pequeño cambio, volvemos a lanzar el análisis de regresión. Como en el caso anterior, gran número de variables no resultan ser estadísticamente significativas, y tras descartarlas obtenemos el modelo de la figura 44.

Podemos observar una clara mejoría en el estadístico R^2 ajustado, que ahora alcanza un valor del 75,3%. Las tres variables estadísticamente significativas son la superficie, ascensor1 y rentaAlta. El signo de todos los coeficientes son positivos, como parece razonable. Es lógico pensar que a mayor superficie, mayor precio (o logaritmo del precio) de la vivienda. El signo positivo del coeficiente asociado a la variable ascensor1 debemos interpretarlo de la siguiente forma: se está dispuesto a pagar más por una vivienda situada en un edificio con un ascensor, que por esa misma vivienda (con esas mismas características) en un edificio sin ascensor, que es el nivel o categoría tomado como base. De igual forma, también se valora positivamente las viviendas

ubicadas en zonas de renta alta, frente a la zona tomada como base: renta media.

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,871[a]	,759	,753	,10903

a. Variables predictoras: (Constante), rentaAlta, ascensor1, superficie

b. Variable dependiente: logPrecio

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	4,558	3	1,519	127,829	,000[b]
	Residual	1,450	122	,012		
	Total	6,009	125			

a. Variable dependiente: logPrecio

b. Variables predictoras: (Constante), rentaAlta, ascensor1, superficie

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	5,014	,025		200,725	,000
	superficie	,003	,000	,773	16,817	,000
	ascensor1	,094	,021	,203	4,440	,000
	rentaAlta	,066	,020	,146	3,264	,001

a. Variable dependiente: logPrecio

Figura 44. Modelo de regresión múltiple

Capítulo 6. El análisis factorial (II)

6.1 Introducción

En el anterior capítulo hemos visto cómo a través del análisis factorial podemos hacer una selección de variables explicativas del precio que elimine o reduzca el impacto negativo de la multicolinealidad en nuestros modelos de valoración. Al escoger un único representante por factor, y saber que estos factores son independientes entre sí, estamos evitando la coexistencia dentro de nuestro modelo de variables altamente correlacionadas entre sí. Ello permitirá que los modelos tengan mayor poder predictivo, frente a los modelos con multicolinealidad de alta capacidad explicativa pero escasa capacidad predictiva.

Sin embargo, también hemos podido comprobar la dificultad que entraña seleccionar el representante adecuado para cada factor. Podemos optar por escoger aquél que tenga un mayor coeficiente, en valor absoluto, dentro de su factor. El problema de esta elección es que no por ello dicho representante será necesariamente el de mayor capacidad explicativa del precio. Esto significa que, en muchas ocasiones, incluiremos en el modelo variables con escasa o nula significación estadística, descartando otras que sí tendrían cabida en el modelo de valoración y lo mejorarían sensiblemente. Justamente esta situación se daba en el ejemplo final del anterior capítulo, donde quedaba excluida inicialmente la superficie por no ser seleccionada como representante de su factor.

La exploración de diferentes combinaciones de variables, hasta encontrar una que satisfaga los estándares del grado de explicación del precio (o logaritmo del precio), puede resultar una tarea harto tediosa. Es por ello que debemos buscar una alternativa que, a priori, nos facilite la selección óptima de los representantes de los factores, entendiendo por óptima la que nos asegura un mayor valor del estadístico R^2 corregido en la correspondiente función de valoración.

6.2 El modelo factorial sobre los residuos

La elección de representantes poco adecuados en los factores afecta negativamente a la bondad de nuestro modelos de valoración.

El problema radica en que el precio no ha intervenido en la formación de los factores, ni directa ni indirectamente. Incorporarlo como una variable más no

tendría sentido, ya que las variables incluidas en factores distinto al suyo estarían, a priori, poco correlacionadas con el precio y, por tanto, no entrarían en los modelos de valoración. Únicamente tendría sentido incluir variables que pertenecieran al mismo factor que el precio, pero para evitar problemas de multicolinealidad sólo podríamos considerar una de estas variables. En definitiva, trabajaríamos con modelos de regresión simple.

En la siguiente figura aparecen los 9 factores extraídos al incluir el precio junto con las variables explicativas. El precio (logaritmo del precio) aparece en el primer factor, junto con la superficie y las variables indicadoras de 5 o más dormitorios y 3 o más baños. El resto de variables, por encontrarse en otros factores, se entiende se encuentran poco correlacionados con las variables del primer factor, con lo que a priori no tendría sentido incluirlas en el modelo de valoración. Únicamente se seleccionaría una variable de primer factor, la superficie. Obviamente, este modelo univariante tendría escasa capacidad explicativa.

Matriz de componentes rotados[a]

	Componente								
	1	2	3	4	5	6	7	8	9
superficie	,869	,142	,163	,011	,180	,038	,217	-,008	-,042
numDormit5omas	,818	,101	,078	,031	-,099	-,039	-,328	,132	-,029
logPrecio	,777	,166	,280	,228	,139	-,006	,206	-,140	-,023
numBaños3omas	,627	,072	-,370	-,072	-,092	,006	,156	,247	,071
numPlantas	,097	,886	,064	-,018	-,028	-,013	,066	-,012	,005
numVivEdif	,068	,882	,040	,078	-,029	-,003	,101	-,011	-,043
ascensor2	,261	,632	-,492	,028	,058	,105	-,079	-,139	-,031
plantaVivienda	,154	,562	,047	-,159	-,074	-,053	,069	,522	,047
ascensor1	,182	-,025	,822	,059	-,045	-,044	,115	-,006	,102
numBaños2	,266	,174	,509	-,057	,352	,236	,075	-,193	-,222
rentaAlta	,130	-,019	,007	,809	,033	-,160	-,019	-,007	-,116
densidadAlta	-,017	,066	,130	,680	,014	,185	-,222	-,116	,246
entComExcel	-,126	-,048	-,165	,472	-,110	-,432	,341	,170	-,294
numDormit3	-,133	-,101	-,048	,045	,899	-,053	-,200	-,043	,081
numDormit2	-,358	-,017	-,095	,036	-,716	,102	-,302	-,043	-,067
calidadConstrVivSocial	,092	-,156	-,341	-,011	-,022	-,664	,015	-,171	,229
entComBasico	-,005	-,128	-,261	-,136	-,156	,580	-,008	-,193	,048
calidadConstrAlta	,101	-,034	-,100	,442	-,058	,538	,253	,302	-,024
numDormit4	,143	,175	,167	-,113	-,005	,043	,866	-,078	,017
supTerrazaDesc	,049	-,063	-,037	,012	,003	,032	-,080	,843	-,016
rentaBaja	-,029	-,023	,020	,004	,089	-,072	,016	,009	,915

Método de extracción: Análisis de componentes principales.
Método de rotación: Normalización Varimax con Kaiser.

a. La rotación ha convergido en 27 iteraciones.

Figura 45. Composición y cargas factoriales, incluyendo el logaritmo del precio

Siendo evidente que el precio no debe formar parte directa de este proceso, sí parece razonable que sea tenido en cuenta, aunque sea de forma indirecta.

Sería adecuado que las variables se agruparan en factores no sólo por su "afinidad" o correlación, sino por el grado de explicación conjunto respecto de la variable precio. Esto es, si dos variables explican una parte similar de la variabilidad del precio, entonces deberían compartir factor. Y si dos variables explican partes independientes de la variabilidad del precio, entonces debería estar en factores distintos.

Pero, ¿cómo saber cuándo explican partes similares o diferentes de la variabilidad del precio? El siguiente algoritmo[4] permite la agrupación de variables en factores siguiendo esta máxima:

1. Realizar n regresiones simple entre el precio (variable dependiente) y cada una de las n variables independientes.
2. Guardar los residuos de las anteriores regresiones simples, generando por tanto n variables.
3. Realizar un análisis factorial sobre las n variables generadas.
4. Escoger como representante de cada factor el residuo con mayor carga factorial.
5. Obtener un modelo de regresión entre el precio y las variables escogidas como representantes de cada factor, eliminando aquellas variables cuyos coeficientes no resulten estadísticamente significativos.

En la siguiente figura aparece el resultado de regresar el logaritmo del precio frente a la superficie (paso 1). De esta regresión nos guardaríamos el residuo: diferencia el precio observado en cada vivienda respecto del estimado por la función (paso 2).

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,834[a]	,696	,694	,12130

a. Variables predictoras: (Constante), superficie

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	4,184	1	4,184	284,392	,000[b]
	Residual	1,824	124	,015		
	Total	6,009	125			

a. Variable dependiente: logPrecio

b. Variables predictoras: (Constante), superficie

[4] Pueden encontrarse detalles en García, F., Guijarro, F. y Moya, I. (2009) "An algorithm for variable selection in firm valuation models", *International Journal of Business Performance and Supply Chain Modelling*, 1(2), pp. 144-161.

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados		
		B	Error típ.	Beta	t	Sig.
1	(Constante)	5,073	,025		201,399	,000
	superficie	,004	,000	,834	16,864	,000

a. Variable dependiente: logPrecio

Figura 46. Regresión entre el logaritmo del precio y la superficie

El proceso se repite con el resto de variables explicativas del precio, guardando los correspondientes residuos. De esta forma, habremos generado tantas nuevas variables como variables explicativas tengamos en nuestra muestra.

A continuación llevaríamos a cabo el análisis factorial entre las variables que recogen los residuos (paso 3). La siguiente figura representa el resultado de dicho factorial.

Matriz de componentes rotados[a]

a. Sólo se ha extraído un componente. La solución no puede ser rotada.

Figura 47. Resultado del análisis factorial entre las variables que contienen los residuos de las regresiones simples

El paquete estadístico SPSS sólo ha podido extraer un factor en el análisis. Esto significa que todas las variables están correlacionadas entre sí, en mayor o menor grado, pero de manera significativa. Al haber analizado los residuos de las regresiones, esto es, la parte del precio que no explica cada variable por separado, el análisis nos está diciendo que esa parte sin explicar es bastante similar para todas las variables. Expresado de otra forma, que la parte del precio que no explica la superficie es bastante parecida a la parte del precio que no explica el número de habitaciones, o la parte del precio que no explica la calidad de la construcción, etc.

La extracción de un único factor implica que en sólo seleccionaríamos una variable que representaría a dicho factor (paso 4), con lo que el modelo de valoración se obtendría a partir de un análisis de regresión simple (paso 5).

Un modelo con una única variable explicativa, como podemos imaginar, no ofrecerá buenos resultados. En estos casos, podemos pedir al paquete estadístico SPSS qu extraiga más de un factor.

En la siguiente figura aparecen las cargas factoriales de cada variable para un modelo bifactorial. Atendiendo a dichas cargas podemos ver que el primer factor incluye a todas las variables excepto a una, la superficie, que formaría parte del segundo factor.

Matriz de componentes rotados[a]

	Componente	
	1	2
r_densidadAlta	,946	,288
r_entComExel	,945	,318
r_supTerrazaDesc	,945	,318
r_calidadConstrVivSocial	,944	,324
r_numDormit3	,943	,327
r_rentaBaja	,943	,327
r_entComBasico	,940	,322
r_calidadConstrAlta	,937	,331
r_rentaAlta	,935	,277
r_plantaVivienda	,931	,350
r_numVivEdif	,923	,303
r_numPlantas	,922	,344
r_ascensor2	,918	,333
r_ascensor1	,907	,209
r_numDormit4	,887	,348
r_numBaños3omas	,849	,425
r_numDormit2	,846	,408
r_numBaños2	,843	,329
r_numDormit5omas	,733	,537
r_superficie	,251	,958

Método de extracción: Análisis de componentes principales.
Método de rotación: Normalización Varimax con Kaiser.

a. La rotación ha convergido en 3 iteraciones.

Figura 48. Resultado del análisis factorial entre las variables que contienen los residuos de las regresiones simples: modelo con dos factores

Siguiendo con el mismo procedimiento en los pasos 4 y 5, escogeríamos como representantes de los factores a las variables densidadAlta (factor 1) y superficie (factor 2).

El modelo de regresión múltiple entre el logaritmo del precio y estas dos variables explicativas obtiene un R^2 corregido de 71,6%. Un valor aceptable para ser un modelo con dos únicas variables explicativas, pero insuficiente para la práctica profesional.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,849[a]	,720	,716	,11694

a. Variables predictoras: (Constante), densidadAlta, superficie

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	4,327	2	2,163	158,201	,000[b]
	Residual	1,682	123	,014		
	Total	6,009	125			

a. Variable dependiente: logPrecio

b. Variables predictoras: (Constante), densidadAlta, superficie

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	5,029	,028		180,587	,000
	superficie	,004	,000	,835	17,510	,000
	densidadAlta	,070	,022	,154	3,227	,002

a. Variable dependiente: logPrecio

Figura 49. Regresión entre el logaritmo del precio (variable dependiente), la superficie y densidadAlta (variables independientes)

Podríamos pensar en seguir ampliando el número de factores, hasta alcanzar una bondad del ajuste adecuada para nuestro fines profesionales. Considerando tres factores en el análisis factorial, los representantes que se escogerían serían las variables numBaños2, numDormit5omas y superficie.

Matriz de componentes rotados[a]

	Componente		
	1	2	3
r_numBaños2	,867	,217	,329
r_ascensor1	,844	,372	,176
r_numDormit4	,840	,354	,321
r_supTerrazaDesc	,809	,522	,258
r_calidadConstrVivSocial	,809	,520	,264
r_entComExel	,804	,532	,256
r_rentaBaja	,804	,528	,266
r_numDormit3	,804	,529	,266
r_entComBasico	,802	,526	,261
r_densidadAlta	,799	,538	,224
r_calidadConstrAlta	,795	,532	,269
r_plantaVivienda	,789	,534	,287
r_numPlantas	,787	,519	,284
r_numVivEdif	,781	,525	,241
r_rentaAlta	,779	,551	,210
r_numDormit2	,767	,407	,369
r_ascensor2	,751	,573	,261
r_numBaños3omas	,647	,626	,341
r_numDormit5omas	,477	,704	,434
r_superficie	,230	,215	,944

Método de extracción: Análisis de componentes principales.
Método de rotación: Normalización Varimax con Kaiser.

a. La rotación ha convergido en 6 iteraciones.

Figura 50. Resultado del análisis factorial entre las variables que contienen los residuos de las regresiones simples: modelo con tres factores

Sin embargo, el modelo de regresión múltiple obtenido a partir de estas variables obtiene peores resultados que el modelo anterior: un coeficiente de determinación ajustado del 69,8%, y un modelo en el que algunos coeficientes no son estadísticamente significativos.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,840[a]	,705	,698	,12052

a. Variables predictoras: (Constante), numDormit5omas, numBaños2, superficie

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	4,237	3	1,412	97,230	,000[b]
	Residual	1,772	122	,015		
	Total	6,009	125			

a. Variable dependiente: logPrecio

b. Variables predictoras: (Constante), numDormit5omas, numBaños2, superficie

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	5,073	,028		178,311	,000
	superficie	,004	,000	,797	11,689	,000
	numBaños2	,046	,025	,102	1,865	,065
	numDormit5omas	-,008	,047	-,010	-,165	,870

a. Variable dependiente: logPrecio

Figura 51. Regresión entre el logaritmo del precio (variable dependiente), la superficie, numBaños2 y numDormit5omas (variables independientes)

¿Dónde se encuentra entonces el problema? Capítulos atrás indicábamos que para poder llevar a cabo un modelo de regresión debemos comprobar una serie de hipótesis: normalidad de los datos, homocedasticidad, normalidad de los residuos, etc. Si representamos en un histograma los residuos de la regresión entre el logaritmo del precio y la superficie, podemos ver que el resultado no se corresponde exactamente con el de una distribución Normal. En la figura 52 se ha representado, junto con los residuos, la distribución Normal teórica. Podemos comprobar cómo existe una clara asimetría en los resultados de los residuos que no debería tenerse si asumiéramos que estos siguen una distribución normal.

Pues bien, ¿y si en lugar de emplear el logaritmo del precio volviéramos al precio? Comprobemos si el modelo resultante obtiene mejor coeficiente de determinación ajustado y no tenemos problemas de heterocedasticidad.

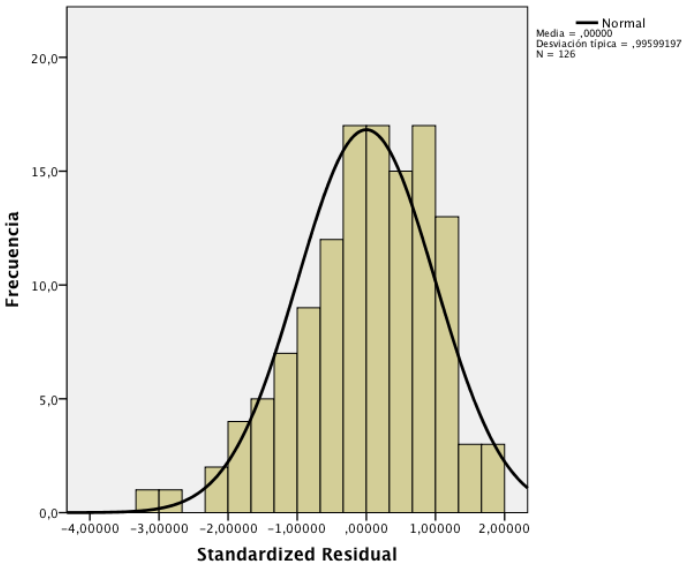


Figura 52. Histograma de residuos

Al considerar el precio, en lugar de su logaritmo, deberemos repetir todos los análisis de regresión simple, para obtener una nueva serie de residuos con los que llevar a cabo el análisis factorial.

Al repetir el análisis factorial, de nuevo se ha extraído un único factor. Pero ampliando el modelo a dos factores, se obtienen las cargas factoriales de la figura 53.

Podemos ver como en el primer factor las cargas factoriales de las primeras variables son muy similares entre sí, lo que indicaría que tendrán un efecto muy similar en el modelo de regresión. En el segundo factor sólo se encuentra la superficie.

De esta forma, llevamos a cabo un análisis de regresión entre el precio y las variables densidadAlta (factor 1) y superficie (factor 2). El resultado aparece en la figura 54.

Matriz de componentes rotados[a]

	Componente	
	1	2
res_densidadAlta	,959	,250
res_entComExcel	,959	,276
res_calidadConstrVivSocial	,956	,282
res_numDormit3	,956	,283
res_antiguedad	,956	,277
res_remtaBaja	,956	,288
res_supTerrazaDesc	,956	,288
res_entComBasico	,953	,287
res_calidadConstrAlta	,950	,291
res_rentaAlta	,949	,221
res_plantaVivienda	,942	,312
res_numVivEdif	,941	,293
res_numPlantas	,935	,310
res_ascensor2	,934	,279
res_ascensor1	,925	,277
res_numDormit4	,902	,325
res_numBaños2	,877	,364
res_numDormit2	,867	,387
res_numBaños3omas	,850	,351
res_numDormit5omas	,742	,489
res_superficie	,235	,964

Método de extracción: Análisis de componentes principales.
Método de rotación: Normalización Varimax con Kaiser.

a. La rotación ha convergido en 3 iteraciones.

Figura 53. Cargas factoriales del modelo con dos factores

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,876[a]	,767	,763	99288,5336

a. Variables predictoras: (Constante), densidadAlta, superficie

b. Variable dependiente: Precio

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	3,984E+12	2	1,992E+12	202,069	,000[b]
	Residual	1,213E+12	123	9,858E+9		
	Total	5,197E+12	125			

a. Variable dependiente: Precio

b. Variables predictoras: (Constante), densidadAlta, superficie

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	-78978,143	23643,723		-3,340	,001
	superficie	3535,002	178,049	,865	19,854	,000
	densidadAlta	59759,377	18290,638	,142	3,267	,001

a. Variable dependiente: Precio

Figura 54. Regresión entre el precio (variable dependiente), la superficie y densidadAlta (variables independientes)

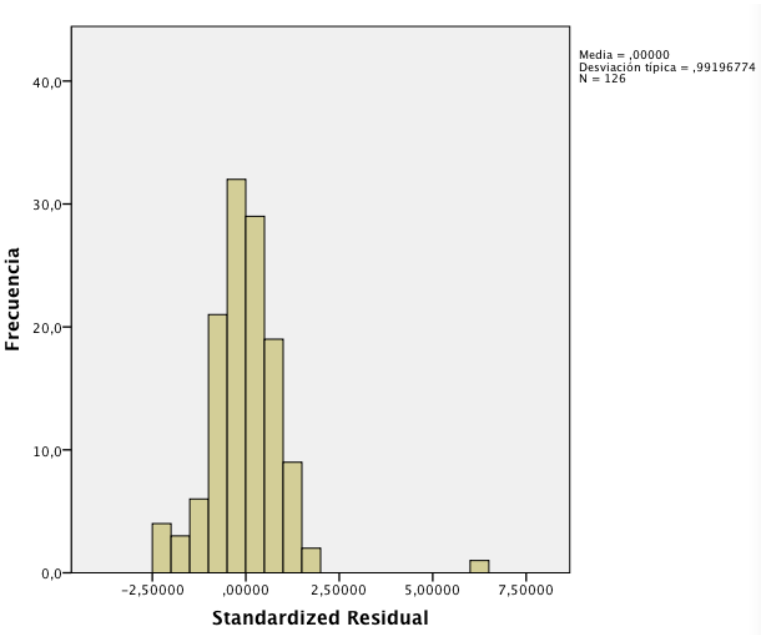


Figura 55. Histograma de residuos

Al representar los residuos estandarizados de la regresión, claramente detectamos una vivienda con un residuo extremo y aislado del resto. Lo eliminamos del análisis y repetimos la regresión. Finalmente, nuestro modelo alcanza un nivel de explicación del 79,4%.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,893[a]	,798	,794	79430,0943

a. Variables predictoras: (Constante), densidadAlta, superficie

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	3,037E+12	2	1,518E+12	240,679	,000[b]
	Residual	7,697E+11	122	6,309E+9		
	Total	3,807E+12	124			

a. Variable dependiente: Precio

b. Variables predictoras: (Constante), densidadAlta, superficie

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	-44861,320	19348,191		-2,319	,022
	superficie	3209,972	147,627	,885	21,744	,000
	densidadAlta	50784,041	14671,540	,141	3,461	,001

a. Variable dependiente: Precio

Figura 56. Regresión entre el precio (variable dependiente), la superficie y densidadAlta (variables independientes)

Aún no llegando al 85%, es cierto que se encuentra muy cerca de ese mínimo que nos marcamos para la aplicación de los modelos en la práctica profesional. Con lo que podríamos empezar a pensar en su utilización para el código postal analizado.

En cualquier caso, debemos insistir en la principal ventaja del modelo factorial sobre los residuos: se pueden escoger los representantes de cada factor e introducirlos directamente como variables explicativas en el modelo de regresión múltiple final. Frente al modelo de análisis factorial del capítulo anterior, en el que en ocasiones debemos sustituir alguna variable representante del factor por otra que tenga menor carga factorial, pero que sea más efectiva en los modelos de regresión. En el modelo sobre los residuos tenemos mayor seguridad en que las variables seleccionadas como representantes de los factores funcionarán mejor en los modelos de regresión múltiple. No tendremos que realizar sustituciones entre variables, lo que puede suponer mucho tiempo para el tasador.

BIBLIOGRAFÍA

Cuadras, C. M. (1996). *Métodos de Análisis Multivariante*. Ed. EUB.

Hair, J.F. Tatham, R.L. Black, W.C. (1999). *Análisis Multivariante*. Ed. Prentice Hall.

Peña, D. (2003). *Análisis de datos multivariantes*. Ed. McGraw Hill.

Sharma, S. (1996). *Applied Multivariate Techniques*. Ed. John Wiley & Sons.

Uriel, E. y Aldás, J. (2005): *Análisis multivariante aplicado*. Ed. Thomson.