Análisis bivariante

PID_00268329

Jordi Mas Elias

Tiempo mínimo de dedicación recomendado: 4 horas

Universitat Oberta de Catalunya

Jordi Mas Elias

El encargo y la creación de este recurso de aprendizaje UOC han sido coordinados por el profesor: Jordi Mas Elias (2019)

Primera edición: septiembre 2019

Avda. Tibidabo, 39-43, 08035 Barcelona
Realización editorial: FUOC

Índice

Introducción

El análisis bivariante se ocupa del análisis estadístico con dos variables. A veces las funciones del análisis bivariante se confunden con las del análisis univariante, puesto que este último también admite a menudo el análisis con dos variables. La gran distinción, sin embargo, es que el análisis univariante está orientado a hacer comparaciones entre subgrupos y describir las similitudes y las diferencias de esta variable entre subgrupos, mientras que el análisis bivariante busca principalmente explicar cómo dos variables se relacionan entre sí (Babbie, 2013). Por ejemplo, en análisis univariante podemos observar si los países que tienen acceso al mar son más ricos que los países que no lo tienen. En cambio, en análisis bivariante, buscaremos la asociación entre estas dos variables y diremos que tener acceso al mar tiene un efecto sobre la riqueza. Cuando hablemos, pues, de asociaciones o efectos entre variables, solo lo haremos mediante las técnicas del análisis bivariante, que nos permitirá sugerir desde un punto de vista estadístico si el comportamiento de una variable está parcialmente determinado por la otra (King y otros, 1994).

Establecer asociaciones entre dos variables es una práctica frecuente en el día a día. Acostumbramos a decir que la contaminación tiene un efecto sobre el calentamiento global, que los países democráticos suelen ser más ricos o que tener una ideología de derechas influye a la hora de votar un partido determinado. En otras palabras, creemos que si conocemos los valores de la primera variable (niveles de contaminación, democracia o ideología) podremos saber algo sobre los valores de la segunda variable (calentamiento global, riqueza o voto a partido) porque creemos que están relacionadas. Llamaremos a la primera variable, que es la que pensamos que causa el efecto, **variable independiente** y la ilustraremos con el símbolo x. La segunda variable, que es la que creemos que está afectada por la variable independiente, la denominaremos **variable dependiente** y la ilustraremos con el símbolo y.

Otras denominaciones de las variables

Hay otras maneras de referirse a las variables independiente y dependiente. A la variable independiente se la denomina variable explicativa, mientras que la variable dependiente es llamada variable explicada o de respuesta.

Para visualizar y cuantificar estadísticamente una asociación bivariante, utilizaremos una técnica diferente según si las variables independiente y dependiente son numéricas o categóricas. En la tabla 1 resumimos la técnica utilizada según el tipo de variable y si ocupa la posición de independiente o dependiente. Si tanto la variable independiente como la variable dependiente son categóricas, utilizaremos una tabla de contingencia. Si, en cambio, tanto la variable independiente como la variable dependiente son numéricas, utilizaremos la regresión y la correlación. Si la variable independiente es categórica y la dependiente es continua, usaremos la diferencia de medias. Finalmente, si la variable independiente es continua y la variable dependiente es categórica, haremos una regresión logística.

Tabla 1. Técnicas de análisis bivariante

		Variable independiente	
		Categórica	**Numérica**
Variable dependiente	**Categórica**	Tablas de contingencia	(Regresión logística)
	Numérica	Diferencia de medias	Regresión y correlación

En este módulo no explicaremos la regresión logística, de forma que las páginas siguientes describirán las otras tres técnicas de análisis bivariante:

- la tabla de contingencia,
- la diferencia de medias y
- la regresión y la correlación.

Cada apartado muestra primero cómo representar visualmente una relación bivariante según si las variables independiente y dependiente son numéricas o categóricas, y a continuación se explica la manera de cuantificar la relación.

Aunque las técnicas utilizadas cambien según los tipos de variable, todas las combinaciones tienen unos requisitos similares para indicar si dos variables tienen relación.

1) El primer requisito, que es el que trataremos con más profundidad en estas páginas, es que la asociación observada tenga **apoyo estadístico**. A veces podemos observar asociaciones poco consistentes entre variables. Por ejemplo, si echamos una moneda y sale dos veces cara, sería una inferencia débil decir que siempre saldrá cara. Desde un punto de vista estadístico, basado en las teorías de la probabilidad, para hacer inferencias a partir de una muestra necesitamos un número bastante elevado de casos y/o que la variación de la variable dependiente cuando varía la independiente sea suficientemente grande. Por lo tanto, la relación observada tiene que ser lo bastante consistente como para que no sea muy probable pensar que realmente no hay relación en el fenómeno observado.

2) El segundo requisito se refiere a **la temporalidad**. Es decir, por lógica, la variable independiente se tiene que haber producido antes o más o menos al mismo tiempo que la variable dependiente.

3) Y finalmente, para poder decir qué variable tiene un efecto sobre otra, no solamente hace falta que la estadística y la temporalidad nos den la razón, sino que la teoría también nos la tiene que dar. Este es el requisito más difícil de conseguir y por razones de espacio no trabajaremos de manera específica esta parte en este módulo. Sin embargo, hay que tener claro que para sugerir que

dos variables tienen relación hará falta que esta relación sea lógica y plausible y, por eso, es importante haber descartado las otras alternativas que, teóricamente, pueden explicar el fenómeno que queremos estudiar.

La necesidad de apoyo teórico en el análisis

La relación entre riqueza y democracia es un buen ejemplo para explicar la necesidad de teoría. Si los datos nos dicen que los países democráticos son más ricos que los países no democráticos, esto no quiere decir que haya un efecto entre ser una democracia y ser rico. Podría ser que el efecto fuera a la inversa: que ser rico te haga ser más democrático. O bien que haya otra variable que desconocemos que afecte a la riqueza y la democracia a la vez, pero que democracia y riqueza no se afecten entre ellas. Encontraréis esta idea ampliada en Halperin y Heath (2016, pág. 369-370).

1. Relación entre variables categóricas

La **tabla de contingencia** es la manera de representar una relación bivariante cuando las variables independiente y dependiente son categóricas. Normalmente representaremos los valores de la variable independiente en el eje horizontal y los valores de la dependiente en el eje vertical, de forma que cada celda de dentro de la tabla representa el recuento total de observaciones que caben en cada combinación entre niveles de las dos variables. Una vez tengamos en cada celda el número de frecuencias, calcularemos los porcentajes de cada frecuencia sobre el total de la columna.

Para ilustrar la tabla de contingencia, utilizaremos en esta sección el paquete *unvotes* (Voeten, 2017), que contiene información sobre las votaciones en la Asamblea General de Naciones Unidas (AGNU) separadas en tres paquetes. Lo primero que haremos es instalar y cargar el paquete y pediremos la estructura de los tres marcos de datos.

```
str(un_votes)
str(un_roll_calls)
str(un_roll_call_issues)
```

En este apartado, utilizaremos la tabla de contingencia para responder a una pregunta sobre la información contenida en el paquete *unvotes*: como sabéis, Sudán del Sur es el último país que ha ingresado en la Organización de las Naciones Unidas (ONU) después de conseguir la independencia en 2011. Partimos de la premisa de que Sudán del Sur se ha querido diferenciar respecto de Sudán en política internacional y queremos comprobar nuestra hipótesis mirando si estos dos países han votado de modo diferente en la AGNU a partir de la independencia de Sudán del Sur.

El paquete *unvotes*

El marco de datos `un_votes` tiene un registro de todas las votaciones en la AGNU según si cada estado votó a favor, en contra o se abstuvo. El marco de datos `un_roll_calls` tiene información más concreta sobre cada votación, como el número de resolución y una descripción del contenido de la votación. El marco de datos `un_roll_calls_issues` separa las votaciones según la temática.

1.1. Visualización con tabla

En la tabla de contingencia situaremos dos variables categóricas:

- La independiente es *country*, situada en las columnas, que puede adoptar los valores Sudán o Sudán del Sur.
- La dependiente es *vote*, situada en las filas, que puede tomar los valores Sí o No.

Creemos que cada país tiene un patrón de voto diferente, de forma que si conocemos los valores de la variable independiente *country* podremos saber algo sobre los valores de la dependiente *vote*. Por eso, en el código siguiente hemos filtrado los datos por los dos países que queremos mirar y por el número de

Sacar niveles de factores

Como la variable *vote* es un factor, cuando eliminamos todos los valores de una de las categorías (en este caso la categoría *abstain*), no nos elimina el nivel correspondiente. Para eliminar el nivel, tendremos que incluir la función `droplevels()`. Mirad cómo se visualiza la tabla de contingencia si sacáis esta función del código.

votación 5117 a partir de la cual participó Sudán del Sur en la AGNU. También hemos eliminado las abstenciones y hemos pedido solo las votaciones relacionadas con el desarrollo económico. Para visualizar la tabla de contingencia, solo hay que indicar los dos vectores categóricos como argumentos de la función `table()`.

```
sudan_ec_votes <- un_votes %>% inner_join(un_roll_call_issues) %>%
  filter(country == c("Sudan", "South Sudan"), rcid >= 5117,
         vote != "abstain", short_name == "ec") %>%
  select(country, vote) %>% droplevels()

> table(sudan_ec_votes$vote, sudan_ec_votes$country)
      South Sudan Sudan
  yes          11    16
  no            1     3
```

Cada celda de la tabla que hemos generado representa una combinación del recuento total de observaciones entre categorías de las dos variables. Parece que Sudán ha votado más veces y más a favor que Sudán del Sur, puesto que ha votado de manera favorable en 16 ocasiones y en contra 3 veces. Sudán del Sur ha votado de manera favorable en 11 ocasiones y en contra en 3 ocasiones.

Sumar filas o columnas

Si queremos ver la suma de las filas o las columnas de la tabla de contingencia, tendremos que introducir la función anterior adentro de la función `rowSums()` o `colSums()`. R nos devolverá un vector numérico con la suma de cada fila o columna.

Aunque desde un punto de vista descriptivo es bueno ver cuántas observaciones tenemos en cada combinación, el análisis de una tabla de contingencia lo haremos de manera más esmerada con proporciones. Si sabemos los porcentajes nos será más fácil comparar los valores entre columnas y distinguir las probabilidades de que la variable dependiente tome un valor determinado si conocemos los valores de la independiente. Para ver las proporciones utilizaremos la función `prop.table()` con el argumento `margin = 2`, que nos permitirá ver los porcentajes en columnas. También hemos multiplicado la tabla por 100 para ver el resultado en tantos por ciento y hemos utilizado `round()` para redondear los valores a un decimal.

Proporciones por columnas, filas o totales

La mejor manera de ver una tabla de contingencia es con las proporciones por columnas. Sin embargo, también podemos ver las proporciones por filas si introducimos el argumento `margin = 1`. Si no introducimos ningún argumento, veremos el porcentaje de cada celda sobre el total de celdas de la tabla.

```
> round(prop.table(table(sudan_ec_votes$vote, sudan_ec_votes$country), margin = 2) * 100, 2)
      South Sudan Sudan
  yes       91.67 84.21
  no         8.33 15.79
```

Vista con proporciones, la tabla de contingencia nos da otro ángulo de los datos en comparación con la tabla de contingencia vista en frecuencias. Desde 2011, Sudán del Sur tiende a votar las resoluciones sobre desarrollo económico más favorablemente que Sudán. Los resultados, pues, parece que de momento no contradicen nuestra hipótesis de partida.

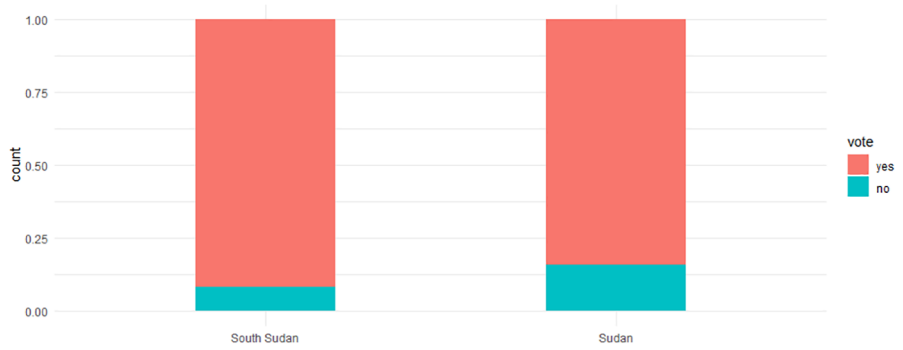
1.2. Visualización gráfica

El diagrama de barras es la manera más habitual de visualizar gráficamente las tablas de contingencia. Si utilizamos el paquete *ggplot2* usaremos la geometría `geom_bar()`. En primer lugar, visualizaremos el recuento de observaciones. En el código siguiente, intentad cambiar la posición por defecto de la geometría `position = "stack"` por `position = "dodge"`, que nos mostrará horizontalmente las diferentes categorías de cada columna.

```
sudan_ec_votes %>%
  ggplot(aes(x = country, fill = vote)) +
  geom_bar() + theme_minimal()
```

Si queremos visualizar la tabla de contingencia en proporciones, solo hará falta que cambiemos la posición de la geometría del código anterior. En este caso, situaremos `position= "fill"`.

Figura 1. Visualización gráfica de una tabla de contingencia



En el código de esta tabla también hemos disminuido la anchura de las barras con el argumento `width = 0.4` y hemos eliminado el nombre del eje horizontal con `xlab(NULL)`.

Diferencia de voto con las administraciones Carter y Reagan.

¿Cómo cambió el voto de los Estados Unidos en Naciones Unidas entre las administraciones Carter y Reagan? En este código tenéis la respuesta.

```
un_roll_calls %>% inner_join(un_votes) %>% select(-c(country, descr, short)) %>%
filter(country_code == "US", vote != "abstain", between(date, as.Date("1977-01-20"),
as.Date("1989-01-20"))) %>% mutate(president = if_else(date < as.Date("1980-01-20"),
"Carter", "Reagan")) %>% ggplot(aes(x = president, fill = vote)) + geom_bar(position = "fill")
```

Fijaos en cómo hemos filtrado las votaciones para poder separar la muestra en dos periodos, el que gobernó Carter y el que gobernó Reagan.

1.3. Cuantificar

Ya hemos explicado que cuantificar una relación va más allá de describir y comparar dos variables (de esto ya se encarga el análisis univariante). Es importante remarcar que uno de los objetivos cruciales de la cuantificación bivariante es saber si los datos apoyan o no a nuestra hipótesis inicial, que acostumbra a ser que las dos variables están relacionadas. En caso de que los resultados confirmen la hipótesis mediante los procedimientos que veremos a continuación, es muy importante no correr: esto no querrá decir que la hipótesis sea correcta. Si es correcta o no nunca lo podremos saber en realidad. Lo único que querrá decir es que los resultados no parecen indicar que las variables son independientes entre sí y, por lo tanto, no tienen relación. Estas dos frases anteriores parecen un lío, pero tienen mucho sentido. Por eso es importante que entendamos el significado de la **hipótesis nula**, que es lo que realmente comprobamos estadísticamente en un análisis bivariante. La hipótesis nula acostumbra a ser la hipótesis contraria a nuestra hipótesis inicial, también llamada **hipótesis alternativa**. Es decir, si decimos que hay relación entre *a y b*, la hipótesis nula que testaremos es que no hay relación entre *a y b*.

Hipótesis nula

Para entender la hipótesis nula, tenemos que partir de la base de que nunca podemos estar seguros de que lo que queremos probar sea cierto. Muchos hallazgos de grandes científicos que en su tiempo fueron un gran descubrimiento han sido falsificados por científicos más contemporáneos. Esto indica que tenemos que ir con cuidado con nuestra investigación y que, en lugar de probar que dos variables tienen relación, siempre es más prudente descartar lo contrario: que las dos variables no tienen relación. En el análisis de datos, pues, siempre comprobaremos la hipótesis nula, y los resultados nos permitirán descartar o no que no haya diferencia desde un punto de vista estadístico.

Cuando cuantificamos, pues, lo que hacemos es tratar de aportar argumentos estadísticos para desmontar la hipótesis nula de que no hay relación entre las variables. El primer argumento que podemos aportar es la diferencia entre los porcentajes de la tabla de contingencia: observamos un 7,46% de diferencia en el porcentaje de voto afirmativo respecto del voto negativo en temas de desarrollo entre Sudán del Sur y Sudán. Este argumento es más bien descriptivo y estadísticamente no es suficiente, aunque no deje de ser un número que nos pueda ayudar a explicar las diferencias entre categorías de la relación que analizamos.

Un segundo argumento que podemos aportar, más asociado con el análisis explicativo, es la fuerza de la relación entre las dos variables. Diremos que una relación entre variables categóricas es muy fuerte cuando, por medio de los valores de la variable independiente, podemos saber con mucha exactitud los valores que tomará la variable dependiente. Hay varios métodos para medir la fuerza de la relación, como el test de Cramer o el de Lambda. La mayoría de estos métodos, que se llaman formalmente **coeficientes de asociación**, devuelven un valor cercano a 1 si la relación tiene mucha fuerza y un valor cercano a 0 si la relación no tiene fuerza. Para utilizar estas herramientas, tendremos que instalar y cargar el paquete *DescTools*. En el código siguiente hemos pedido el test de Cramer con `CramerV()`, donde introducimos adentro de la función un argumento por cada variable.

```
library(DescTools)
CramerV(sudan_ec_votes$vote, sudan_ec_votes$country)
```

Ejemplos de la fuerza de la relación

Si todos los países pobres del mundo fueran autocracias y todos los países ricos fueran democracias, entonces diríamos que la relación entre nivel de desarrollo y régimen político es muy fuerte. Es decir, si sabemos los valores de la variable *desarrollo* podremos adivinar con toda seguridad los valores de la variable *democracia*. Esto, sin embargo, raramente pasa, sobre todo, en ciencias sociales. Siempre tendremos casos de democracias pobres y autocracias ricas, así como también tendremos casos de personas ideológicamente de derechas que voten a partidos de izquierdas o de años en que los niveles de contaminación han aumentado, pero en cambio la temperatura del planeta ha disminuido.

La operación del código anterior nos habrá devuelto un coeficiente de asociación de Cramer cercano a 0,11. Esto quiere decir que si sabemos los valores de *x* tenemos un 11% de probabilidades de acertar los valores de *y*. Ciertamente, no es una probabilidad de acierto muy alta.

¿Vota de modo diferente Israel cuando se tratan temas de Oriente Medio?.

Encontraréis un coeficiente de asociación de Cramer más alto que en el ejemplo anterior si miráis la orientación del voto de Israel en temas relacionados con Oriente Medio en comparación con el resto de temas de la AGNU. Primero, hemos creado el objeto `israel_un` y después hemos creado una tabla de contingencia y pedido el test de Cramer.

```
israel_un <- un_votes %>%
  inner_join(un_roll_call_issues) %>%
  filter(country == "Israel",
         vote != "abstain") %>%
  mutate(me = if_else(short_name == "me", "Middle East", "Other")) %>%
  select(me, vote) %>%droplevels()
prop.table(table(israel_un$vote, israel_un$me), 2)
CramerV(israel_un$vote, israel_un$me)
```

Fijaos en la tabla de contingencia y el coeficiente de asociación de Cramer. En este caso, la relación tiene mucha más fuerza porque, conociendo los valores de la variable independiente (si el tema es sobre Oriente Medio o no), tendremos cerca de un 33% de posibilidades de acertar la orientación del voto de Israel.

Aparte de la fuerza de la relación, la otra medida relevante para cuantificar una relación es saber hasta qué punto es significativo nuestro resultado. Volviendo al ejemplo anterior de la moneda, las conclusiones que podamos sacar sobre el comportamiento de una moneda si la lanzamos dos veces al aire serán poco significativas comparadas con las conclusiones que podamos sacar sobre la misma moneda si la lanzamos cien veces. Con el ejemplo anterior de los dos Sudanes nos pasa algo parecido. Supongamos que mañana hay una votación en la ONU y Sudán del Sur vota en contra de una resolución sobre desarrollo económico. Esto cambiaría los porcentajes de Sudán del Sur en la tabla de contingencia de una manera bastante significativa: los porcentajes pasarían a ser aproximadamente del 85% a favor y del 15% en contra, exactamente los mismos porcentajes que tiene Sudán. Si únicamente un solo caso adicional derroca nuestra hipótesis de que Sudán del Sur se ha querido diferenciar respecto de Sudán en política internacional, no podemos estar muy convencidos de que la diferencia observada en nuestro análisis sea generalizable de manera suficientemente significativa.

Para ayudarnos a decidir si el resultado obtenido es suficientemente significativo o no, la estadística tiene varias técnicas, que se denominan pruebas de significación, que nos ayudan a decidir si damos por buena la diferencia observada en nuestros datos. En otras palabras, un **test de significación** nos dice qué probabilidad hay de que la hipótesis nula sea cierta y que, consecuentemente, no haya realmente ninguna diferencia en nuestros resultados si los extrapoláramos a una población de casos más amplia.

Para saber más

Para una revisión a fondo de la significación y otros temas estadísticos en ciencias sociales explicados en este módulo, podéis consultar Agresti y Finlay (2009), Babbie (2013), Halperin y Heath (2016) y Johnson y otros (2007).

El test de significación más habitual que se utiliza para tablas de contingencia es la **prueba del Khi-cuadrado de Pearson**. Este test lo hacemos con la función `chisq.test()` y nos tendremos que fijar en el p-valor (*p-value*) de la última línea que devuelve la consola. Este valor nos dice la probabilidad de que la hipótesis nula sea cierta.

```
> chisq.test(sudan_ec_votes$vote, sudan_ec_votes$country)
X-squared = 0.0028326, df = 1, p-value = 0.9576
```

Según el test de significación que acabamos de aplicar, el 0,95 que hemos obtenido querrá decir que hay un 95% de posibilidades de que la hipótesis nula sea cierta. En ciencias sociales, existe la convención de que los test de significación necesitan un p-valor inferior a 0,05 para rechazar la hipótesis nula. Un p-valor más bajo que 0,05 querrá decir que las probabilidades de que realmente no haya ninguna diferencia son muy bajas, inferiores a un 5%, proporción que los científicos consideran suficiente para rechazar la hipótesis nula. En nuestro ejemplo, la hipótesis nula tiene más de un 95% de probabilidades de ser cierta. Es un número demasiado alto, ya que necesitamos menos de un 5% para descartarla y poder aceptar nuestra hipótesis alternativa. Diremos que no podemos asegurar con un 95% de confianza que nuestra hipótesis sea cierta. Por tanto, en este caso, tendremos que aceptar la hipótesis nula.

Test de Fisher como alternativa

Otro test de significación para tablas de contingencias es el test de Fisher. Podemos pedirlo con la función `fisher.test()` y en el retorno a la consola también podremos observar un p-valor.

¿Votan de modo diferente los Estados Unidos si en el título de la resolución aparece la palabra USSR?

Nos hemos preguntado si el comportamiento de voto de los Estados Unidos se escapa de lo habitual cuando aparece la palabra *USSR* en el título de la resolución. Por eso, hemos utilizado el paquete *stringr* para crear una nueva columna, que sea verdad cuando aparece USSR y falso cuando no aparece. Al acabar, hemos pedido la tabla de contingencia y el test de significación.

```
us_ussr_un <- un_roll_calls %>%
  mutate(ussr = str_detect(descr, "USSR")) %>%
  inner_join(un_votes) %>%
  filter(country_code == "US", vote != "abstain") %>%
  select(ussr, vote) %>%droplevels()

prop.table(table(us_ussr_un$vote, us_ussr_un$ussr), 2)
chisq.test(us_ussr_un$vote, us_ussr_un$ussr)
```

Podemos observar que la diferencia en el voto de Estados Unidos cuando en la resolución aparece USSR y cuando no aparece es significativa, con un 95% de confianza. En otras palabras, la probabilidad de que realmente no haya diferencia en una población más elevada de casos es inferior al 5%.

2. Relación entre variable categórica y numérica

Cuando queramos mirar la asociación entre una variable independiente categórica y una dependiente numérica utilizaremos la **diferencia de medias**. Para ilustrar esta técnica, en esta sección partimos de la hipótesis de que las democracias y las no democracias votan de modo diferente en la ONU. Lo que no tenemos claro es qué regímenes políticos votan más favorablemente las resoluciones en la AGNU, pero por ahora solo comprobaremos la hipótesis de que hay diferencia entre la media de voto favorable en las democracias en comparación con las no democracias.

Tenemos, pues, una variable independiente categórica, si un país es una democracia o no, y una dependiente numérica, la media de voto afirmativo de cada país por año. Para comprobar nuestra hipótesis, utilizaremos y juntaremos las bases de datos *unvotes* y *psData*. La segunda base de datos, acrónimo de Political Science Data, contiene la función `DDGet()` con la cual podemos extraer la clasificación dicotómica de democracia de Cheibub y otros (2010). Por eso instalaremos y cargaremos la librería *psData* e importaremos las variables *democracy* y *un_continent_name* con el código siguiente.

```
library(psData)
DDdata <- DDGet(vars = c("democracy", "un_continent_name"))
```

El objetivo es construir un solo marco de datos que separe los países por año, que diga si cada año el país era una democracia o no lo era y cuál es la proporción anual de votos favorable de cada país. Primero, hemos limpiado los marcos de datos que necesitamos de *unvotes*, transformando la variable *year* y convirtiendo en binaria la variable *vote*, de forma que adoptará el valor 1 cuando el voto sea favorable y el valor 0 en cualquier otro caso. Una vez convertida en binaria, hemos pedido la media agrupada por año y país para obtener la proporción de voto favorable sobre los votos totales.

```
un_year_mean <- un_roll_calls %>% separate(date, "year", extra = "drop") %>%
inner_join(un_votes) %>% mutate(year = as.numeric(year), vote = if_else(vote == "yes", 1, 0)) %>%
group_by(year, country, country_code) %>% summarize(mean = mean(vote))
```

Fijémonos en el marco de datos `un_year_mean` que hemos creado. Ahora ya tenemos la tabla limpia con los datos de *unvotes* que necesitamos. El segundo paso es unir el marco de datos `DDdata` por medio de las columnas que tienen en común: por código de país y por año. A continuación seleccionamos las columnas que nos interesan y transformamos la variable binaria *democracy* en las categorías “Democracy” y “No Democracy”.

Evitar unir los datos por nombre de país

Unir marcos de datos por el nombre del país es peligroso, porque no podemos estar seguros de que los nombres estén tipificados igual. Por ejemplo, hay marcos de datos que aluden a los Estados Unidos como United States y los que los llaman United States of America. Esto nos crearía dos categorías diferentes cuando, realmente, nos refiramos a la misma. Siempre será mejor, pues, unir marcos de datos por medio de un código estandarizado de país o variables estables como el año.

```
un_dd_year <- un_year_mean %>% inner_join(DDdata, by = c("country_code" = "iso2c",
"year" = "year")) %>% select(country = country.x, country_code, year, mean, democracy,
continent = un_continent_name) %>% mutate(democracy = if_else(democracy == 1,
"Democracy", "No democracy"))
```

El marco de datos `un_dd_year` está formado por las variables necesarias para visualizar y cuantificar una diferencia de medias. Tenemos una variable con el año, otra con la media de cada país por año y otra que indica si el país era o no una democracia.

2.1. Visualizar

Hay varias maneras de visualizar una diferencia de medias. Podemos, por ejemplo, utilizar un diagrama de barras en el cual la altura de las barras represente la media de cada categoría. Dos herramientas también comunes para representar visualmente la diferencia de medias son el diagrama de cajas y el diagrama de dispersión. En este apartado, utilizaremos el diagrama de dispersión porque creemos que es una herramienta con más posibilidades visuales y que nos permite representar más variables en el gráfico. Si el código siguiente no os funciona, intentad cargar el paquete *Hmisc*.

```
un_dd_year %>% filter(year == 1946) %>% ggplot(aes(x = democracy, y = mean, col = continent)) +
geom_jitter(alpha = 0.5) + stat_summary(geom = "point", fun.data = mean_se, col= "red", size = 1.2) +
stat_summary(geom = "errorbar", fun.data = mean_se, col = "red", width = 0.1) + theme_classic()
```

Con el código anterior hemos creado el diagrama de dispersión con *jittering* que observamos en la figura 2 y que contiene la variable categórica *democracy* en el eje de las *x* y la numérica *mean* en el eje de las *y*. Hemos querido saber si el primer año que empezaron las votaciones en la AGNU, en 1946, había diferencia entre la orientación del voto entre las democracias y no democracias. Por eso hemos pedido un sumario estadístico de la media que toma la variable dependiente para cada categoría de la variable independiente, representado con un punto rojo. También hemos pedido un estadístico denominado el **error típico de la media**, representado con líneas rojas, que nos da el intervalo donde podemos asegurar que se encuentra realmente la media con un 95% de confianza.

Evitar los diagramas de barras

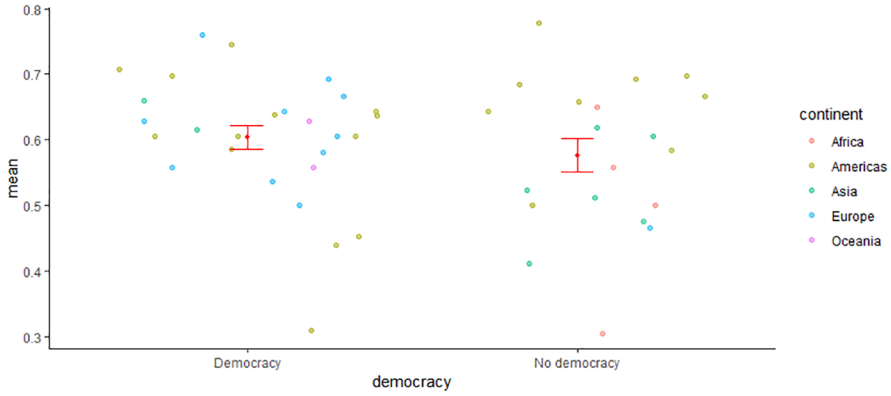
Si sustituimos la altura de la barra por un punto tendremos la misma información y habremos gastado menos tinta. Es por eso que los diagramas de barras no son una buena herramienta para visualizar la diferencia de medias. Utilizaremos siempre diagramas de cajas, que nos permiten ver más descriptivos en el gráfico, o diagramas de dispersión, que nos permiten observar la distribución de la variable dependiente y añadir más variables al gráfico si lo creemos oportuno. En algunos GitHub en internet encontraréis comparaciones entre tipos de representaciones visuales.

Error típico de la media

El error típico de la media nos ayuda a estimar en qué intervalos se puede encontrar realmente la media en la población según un nivel de confianza determinado. Cuantos menos casos tengamos y cuantos más dispersos estén estos casos alrededor de la media, más grande será el error típico y, por lo tanto, menos precisión tendremos para saber

dónde se encuentra la media. Para obtener el intervalo donde se encuentra la media con un 95% de confianza, tendremos que sumar dos errores típicos a la media y restarle dos errores típicos.

Figura 2. Diferencia de media de voto por régimen político en 1946



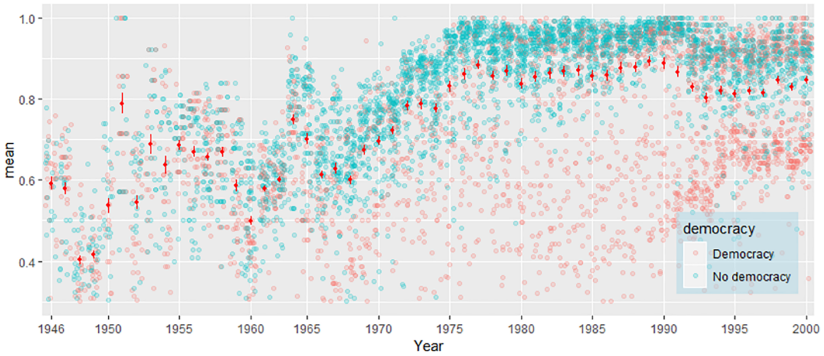
Fijémonos en que las democracias votaban de media ligeramente más a favor que las no democracias en las resoluciones de la AGNU en 1946. Pero fijémonos también en que los intervalos de confianza parece que se encavalquen. Por lo tanto, aunque apreciamos una ligera diferencia entre las dos categorías, esta diferencia no es estadísticamente lo bastante significativa. Esto quiere decir que, si hubiera habido muchas más votaciones en 1946, es demasiado probable, según nuestros estándares científicos, que la diferencia que acabamos de observar realmente no existiera en una población más amplia de casos.

Hemos generado la figura 2 filtrando el código por el año 1946. Probad de modificar el código filtrando por otros años o por toda la muestra de años y mirad si hay diferencia de media entre el voto de democracias y no democracias en la AGNU. Intentad contestar si los resultados apoyan o no a nuestra hipótesis inicial.

Cambios de tendencia en el tiempo

La diferencia de medias se utiliza especialmente para ver cambios de tendencia a lo largo del tiempo. Por ejemplo, podemos ver qué años podemos decir que ha habido un cambio significativo de tendencia en el voto en la AGNU y qué años no podemos asegurar que el cambio observado sea generalizable con un nivel de confianza del 95%. En la figura 3 podemos observar en qué años ha habido un cambio lo bastante significativo de tendencia en el voto y en qué años no.

Figura 3. Análisis longitudinal con diferencia de medias



Fijémonos en los puntos azules y rojos. Todo parece indicar que la mayoría de años las no democracias votan más a favor de las resoluciones de la AGNU que las democracias.

2.2. Cuantificar

Una de las ventajas de visualizar la diferencia de medias es que podemos observar las líneas rojas en que representamos los intervalos de confianza y, por lo tanto, podemos obtener una intuición visual bastante clara de los niveles de significación sin necesidad de cuantificar. Sin embargo, es pertinente ser precisos en nuestros análisis y tener también la estimación de la diferencia de medias cuantificada numéricamente. Esta cuantificación nos tiene que permitir responder a dos preguntas:

1) La primera es si hay diferencia estadística en la media de la variable dependiente según las categorías de la variable independiente. Para encontrar si la respuesta es afirmativa o negativa, tendremos que mirar si podemos rechazar o no la hipótesis nula.

2) La segunda pregunta es saber cuánta diferencia hay entre una media y la otra desde un punto de vista estadístico, por lo cual también nos tendremos que mover con intervalos de confianza para determinar dónde se encuentra la media en la población real.

Para responder a las dos preguntas utilizaremos la función `t.test()` donde introduciremos los vectores de las dos variables que estamos analizando en la figura 2. La primera pregunta se responde con el p-valor. Un p-valor inferior a 0,05 significará que podemos rechazar la hipótesis nula y afirmar con un 95% de confianza que la diferencia de medias observada existe en la realidad. La segunda respuesta se responde con el intervalo de confianza en el 95%, que se obtiene a partir de sumar y restar dos errores típicos a la media.

```
un_dd_year_1946 <- un_dd_year %>%
  filter(year == 1946) %>%
  mutate(democracy = if_else(democracy == "Democracy", 1, 0))


> t.test(un_dd_year_1946$mean, un_dd_year_1946$democracy)
data:  un_dd_year_1946$mean and un_dd_year_1946$democracy
t = 0.23618, df = 49.996, p-value = 0.8143
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1319858  0.1671619
sample estimates:
mean of x mean of y
0.5920561 0.5744681
```

Como podemos ver en el resultado general para este código, tenemos un p-valor muy superior al 0,05. Es decir, tal como ya nos había confirmado visualmente el gráfico de la figura 2, no podemos asegurar que la diferencia entre medias no sea cero. Por lo tanto, en respuesta a la primera pregunta, no podemos asegurar con un 95% de confianza que haya diferencia en orientación de voto entre democracias y no democracias en 1946.

En la segunda pregunta queríamos saber cuál era la diferencia real estimada entre las medias. La respuesta la encontramos en los números -0,13 y 0,16. Estos números indican el intervalo en que estimamos que se moverá la diferencia de medias en la población real con un 95% de confianza. Es decir, con la muestra de resultados que tenemos, podemos apostar con un 95% de acierto que la diferencia de medias en la población real se encontraría en el intervalo siguiente: en un extremo, podría ser que las democracias hayan votado afirmativamente un 16% más que las no democracias, pero en el otro extremo podría ser que las no democracias hayan votado afirmativamente un 13% más que las democracias. Como no lo podemos saber en realidad, lo único que podemos decir es que si apostamos que la media real está entre estos dos intervalos, tenemos un 95% de probabilidades de acertar.

Fijaos en que entra dentro de nuestros pronósticos la posibilidad de que la diferencia de medias sea cero. En esta situación, no podemos rechazar la hipótesis nula, puesto que es demasiado probable que sea cierta. Para contrastarlo con un caso en que sí que hay diferencia, probad de pedir el registro de voto entre los años 1995 y 2000.

Código del registro de voto

Si no encontráis la manera de pedir este registro de voto, encontraréis el código de visualización en el anexo.

El resultado de pedir *el t-test* en este caso es el siguiente:

```
data:  un_dd_year_95$mean and un_dd_year_95$democracy
t = 16.575, df = 1245.1, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.2329403 0.2954871
sample estimates:
mean of x mean of y
0.8203988 0.5561851
```

En este caso, el p-valor es inferior a 0,05 y por lo tanto sí que podemos rechazar la hipótesis nula. Parece, pues, que entre los años 1995 y 2000 hay una diferencia bastante significativa entre el tipo de régimen político y la orientación del voto. Concretamente, podemos estimar, con un 95% de confianza, que la diferencia real en la población se sitúa entre 0,23 y 0,29. En otras palabras, el voto de las no democracias en la AGNU es afirmativo, de media, entre un 23% y un 29% en relación con las democracias. En este caso, el 0% no figura entre nuestros pronósticos. Como la diferencia de medias no puede ser cero, podemos rechazar la hipótesis nula.

3. Relación entre variables numéricas

Cuando queremos explicar la relación entre dos variables numéricas, necesitamos varios instrumentos, tanto visuales como cuantitativos. Visualmente, la manera de representar gráficamente estas dos variables es mediante un diagrama de dispersión, un gráfico de dos dimensiones en que las coordenadas de cada punto representan el valor que toma cada observación según las variables independiente y dependiente. Situaremos siempre la variable independiente x en el eje horizontal y la variable dependiente y en el eje vertical. El diagrama de dispersión se suele complementar con la **recta de regresión**, una técnica tanto visual como cuantitativa que nos permite dibujar la línea que mejor se ajusta a la relación que observamos en el gráfico y cuantificarla. Para cuantificar la relación, también utilizaremos el **coeficiente de correlación**, que describe de manera cuantitativa la fuerza y la dirección de la relación, y el coeficiente de determinación, que ya nos es familiar.

En conjunto, con las técnicas mencionadas podemos llegar a saber cinco aspectos principales de la asociación entre dos variables numéricas. Estos aspectos los resumimos aquí y los ampliamos en los apartados siguientes:

1) Forma: La forma se refiere al dibujo que hace la relación entre x e y. Lo llamaremos lineal si esta forma dibuja una línea más o menos recta. Si, en cambio, este dibujo toma una forma curva, lo llamaremos no lineal o cuadrática.

2) Fuerza: La fuerza nos indica como de perfecta es la forma que hacen los puntos en el gráfico. Si, por ejemplo, las coordenadas de x y de y describen una línea recta perfecta, diremos que la fuerza de la relación es muy alta. A medida que la línea imaginaria que forman los puntos se desdibuja, tendremos unos puntos más dispersos y una relación con poca fuerza.

3) Dirección: Diremos que la dirección es positiva cuando las dos variables se muevan en la misma dirección. Es decir, cuando una incrementa sus valores la otra también lo hace. Diremos que la relación es negativa cuando las dos variables se mueven en dirección opuesta. Es decir, cuando los valores de una variable se incrementan, los valores de la otra disminuyen. Si no apreciamos una dirección positiva o negativa, diremos que es una relación plana.

4) Significación: La relación entre x e y es estadísticamente significativa si las probabilidades de que no haya relación, es decir, de que la línea sea plana, sean inferiores al 5%.

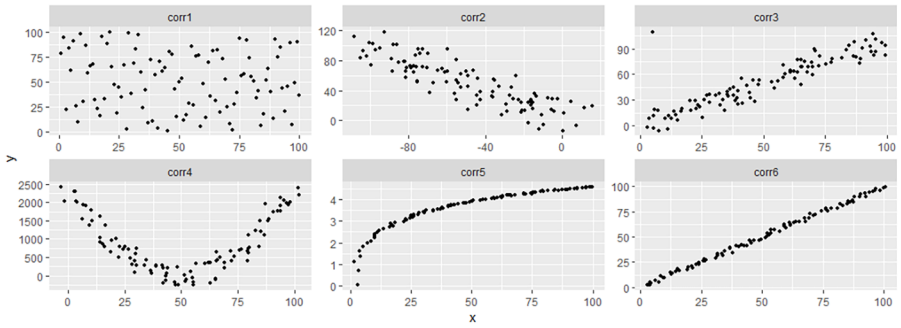
5) Casos extremos: Diremos que la relación tiene casos extremos si hay observaciones que están muy alejadas de la forma que toma la relación.

3.1. Visualizar

La manera de visualizar una relación entre variables numéricas es por medio de un diagrama de dispersión, que podemos llamar con R por medio del paquete *ggplot2* y la geometría `geom_point()`. En la figura 4 vemos seis ejemplos de relaciones entre variables con diferente forma, fuerza, dirección y presencia de casos extremos:

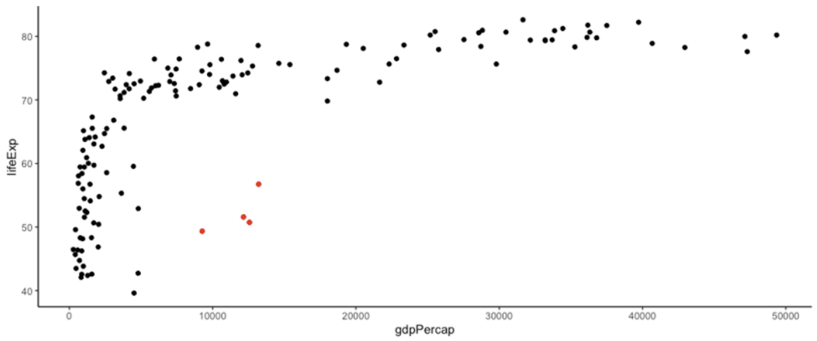
- en el gráfico *corr1* observamos una relación sin forma, sin fuerza y de dirección plana;
- en el caso de *corr2* tiene una relación lineal, de fuerza débil y dirección negativa;
- *corr3* tiene una forma lineal, de fuerza mediana, dirección positiva y que tiene la presencia de un caso extremo;
- en el caso de *corr4* es una relación no lineal, de fuerza media;
- *corr5* tiene una relación no lineal, muy fuerte y positiva;
- *corr6* tiene una relación lineal, muy fuerte y positiva.

Figura 4. Visualización de la correlación entre variables



Caso real de relación no lineal con casos extremos

Uno de los fenómenos más investigados en los estudios de desarrollo es la relación existente entre el PIB per cápita de un país y su esperanza de vida. Como observamos en los datos de 2007, la relación es no lineal, fuerte y positiva. También podríamos considerar la presencia de algunos casos extremos que se encuentran con más de 10.000 dólares per cápita y esperanza de vida alrededor de 50 años, que hemos marcado en rojo. ¿Sabríais encontrar cuáles son estos cuatro casos? Encontraréis el código en los anexos.



La línea de regresión es una técnica visualmente muy utilizada para representar la forma y la dirección de la relación entre dos variables numéricas. Hay diferentes métodos para representar esta relación, entre ellos el método **Ordi-**

nary Least Squares (OLS) y el método **Locally Estimated Scatterplot Smoothing (LOESS)**. La principal diferencia entre los dos es que, mientras que la línea OLS es una línea completamente recta, la LOESS se intenta ajustar a la distribución de los puntos. Nosotros utilizaremos preferentemente la línea OLS.

La característica principal de la línea OLS es que intenta minimizar con una línea recta la suma de la distancia al cuadrado entre la línea y los puntos. En otras palabras, es la recta que de media pasa más cerca de todos los puntos del gráfico. Cualquier otra línea que queramos trazar no conseguirá minimizar esta distancia. La línea OLS se representa con la función `geom_smooth()` junto con la geometría del diagrama de dispersión. Dentro de la función tendremos que especificar el método lm (*linear model*) y preferiblemente pondremos el error típico `se` (*standard error*) en FALSE, como mostramos en este código creado a partir del marco de datos *gapminder*.

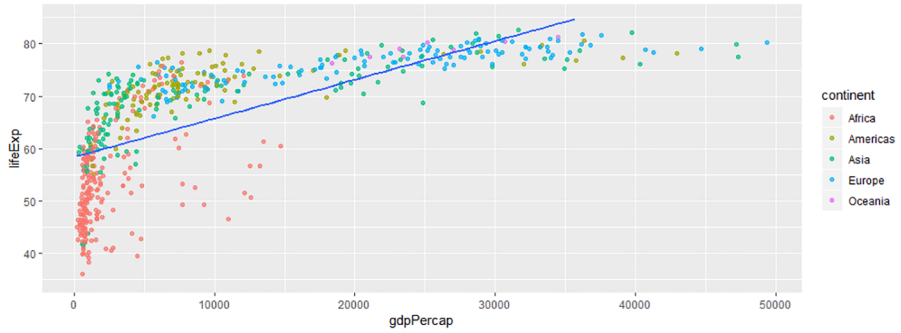
Dibujar una línea manualmente

Podemos intentar dibujar manualmente una línea en el gráfico con la geometría `geom_abline()`. En el primer argumento dentro de la geometría indicaremos la *intercept* y en el segundo indicaremos la pendiente.

```
gapminder %>% filter(year > 1990, lifeExp > 30) %>% ggplot(aes(gdpPercap, lifeExp)) +
geom_point(aes(col = continente), alpha = 0.7) + geom_smooth(method = "lm", se = FALSE)
```

En este código, representado en la figura 5, hemos incluido el estético de color en el diagrama de dispersión pero no lo hemos incluido en la recta de regresión OLS, porque no queremos que genere una línea para cada continente. Podemos ver que la relación entre el PIB per cápita y la esperanza de vida dibuja una forma más bien curva, positiva y más bien fuerte.

Figura 5. Diagrama de dispersión con recta OLS



Líneas horizontales y verticales customizadas

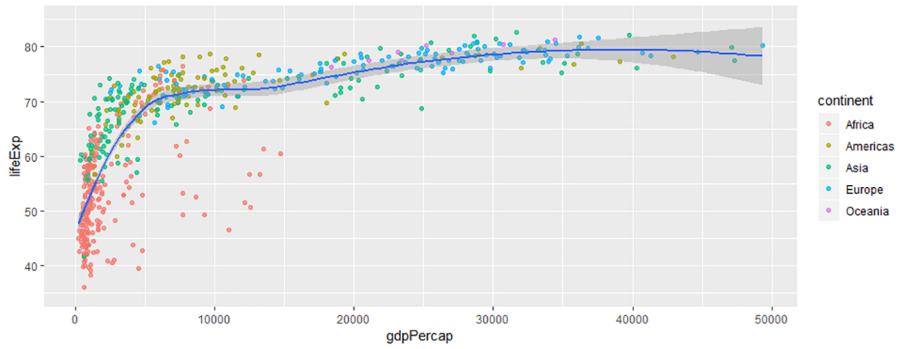
Las geometrías `geom_vline()`, `geom_hline()` y `geom_abline()` permiten colocar una línea en el gráfico con la posición que indicamos. Las dos primeras son líneas paralelas en el eje vertical y la horizontal respectivamente, mientras que la última permite establecer una pendiente con el argumento *slope*.

La línea LOESS es una herramienta utilizada también en análisis de regresión para ver la relación entre dos variables numéricas. Esta técnica es especialmente útil en líneas temporales, puesto que nos permite hacer previsiones de tendencia. Para ver la LOESS en el gráfico anterior, solo hace falta que eliminemos el argumento `method = "lm"`, puesto que `geom_smooth()` ya nos dibuja una LOESS por defecto. También podéis indicarlo explícitamente con `method = "loess"`. En la figura 6, visualizamos una LOESS con los atributos que tiene por defecto, entre ellos el intervalo de confianza al 95% y la sensibilidad de la línea:

Sensibilidad de la línea y error típico

Podemos cambiar la sensibilidad de la línea con el argumento `span`, donde podemos indicar un número entre 0 y 1. Cercano a 0, la sensibilidad será más alta y la línea tendrá cambios más abruptos. Cercano a 1, la línea será más armónica. El argumento `se` nos marcará el intervalo de confianza de la línea al 95% de confianza.

Figura 6. Diagrama de dispersión con línea LOESS



Podemos comprobar que la LOESS establece una línea suave que se va adaptando según la situación de los puntos en el gráfico. Las zonas grises marcan el intervalo de confianza de la línea. Podéis ver que el intervalo será más estrecho en las zonas de los gráficos donde haya más puntos y estén más juntos, mientras que el intervalo será más grande en las zonas donde menos puntos haya y más separados estén.

3.2. Cuantificar

Con un diagrama de dispersión ya podemos intuir visualmente algunos parámetros como la fuerza y la dirección de una relación entre variables numéricas, pero la cuantificación nos será más útil para hacernos preguntas más precisas, como por ejemplo: «¿cómo de fuerte?» Para poder cuantificar estos parámetros, tenemos principalmente dos medidas numéricas:

- el coeficiente de correlación,
- el coeficiente de determinación.

El **coeficiente de correlación**, que a menudo veremos ilustrado con el símbolo *r*, es la manera que utilizaremos para cuantificar la fuerza y la dirección de una relación lineal entre dos variables numéricas. Cuantificar relaciones por medio del coeficiente de correlación nos permite comparar numéricamente las relaciones entre ellas. Este coeficiente es un número que varía entre -1 a 1. El signo nos indica si la relación es positiva o negativa. La magnitud del número corresponde con la fuerza de la relación. Por lo tanto, un número cercano a 1 indicará una relación muy fuerte y positiva. Un número cercano a -1 indicará una relación muy fuerte y negativa, mientras que un número cercano a 0 indicará que no hay relación entre las dos variables.

Karl Pearson, el inventor

También oiremos hablar de *r* como el coeficiente de correlación de Pearson (u originalmente *Pearson product-moment correlation*). Karl Pearson es uno de los matemáticos que han contribuido más a la evolución de la estadística. Con su coeficiente de correlación, Pearson pretendía encontrar una manera de calcular la desviación respecto de la media en *x* e *y*. Podéis encontrar en internet muchos vídeos que explican la lógica matemática que hay detrás de *r*. Memorizar la fórmula no es importante, porque ya nos lo hace R, pero sí que tener claro de donde sale *r* nos ayudará a reforzar la intuición que hay detrás.

Volvámonos a fijar en la figura 4 e intentemos deducir el coeficiente de correlación de cada una de las relaciones. Después podemos comprobar si nos hemos acercado mediante la función `cor()`, que calcula el coeficiente de correlación a partir de dos argumentos: una variable en el primer argumento y la otra variable en el segundo argumento. No importa la orden en que especifiquemos las variables, puesto que las relaciones son simétricas y nos darán el mismo coeficiente. A continuación, hemos imprimido los coeficientes de co-

rrelación de cada uno de los seis gráficos. Como vemos, la relación más fuerte y también positiva la encontramos en *cor6*. También son relaciones fuertes y positivas *cor3* y *cor5*. En el caso de *cor1* y *cor4* no hay relación entre las variables, mientras que en *cor2* la relación es fuerte y negativa.

```
      cor1   cor2  cor3   cor4  cor5  cor6
     <dbl>  <dbl> <dbl>  <dbl> <dbl> <dbl>
1 -0.0563 -0.843 0.923 0.0589 0.894 0.999
```

Es importante tener en cuenta que el coeficiente de correlación no puede capturar con precisión la fuerza de una relación no lineal. Es por eso que en el caso de *cor5*, donde los puntos describen una línea curva casi perfecta, no encontramos un coeficiente de correlación tan elevado como *cor3* o *cor6*. Para incrementar el coeficiente en el caso de estas relaciones cuadráticas, podemos buscar, por ejemplo, el logaritmo neperiano de una de las variables.

De una manera más general, podemos aplicar el coeficiente de correlación para ver cómo están relacionados en tres continentes diferentes el PIB per cápita y la esperanza de vida en varios años. A continuación, hemos pedido un sumario con el coeficiente de correlación por seis años diferentes en Europa, África y Asia.

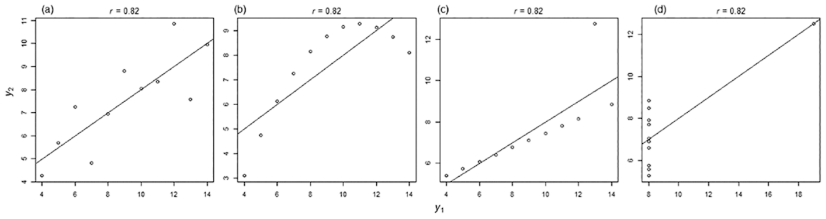
```
> gapminder %>% filter(year %in% c(1952, 1962, 1972, 1977, 1982, 1992, 2002), continent %in%
c("Europe", "Africa", "Asia")) %>% group_by(continent, year) %>%
summarize(cor = cor(gdpPercap, lifeExp), N = n()) %>% spread(year, cor)
# A tibble: 3 x 9
# Groups:   continent [3]
  continent     N `1952` `1962` `1972` `1977` `1982` `1992` `2002`
  <fct>     <int>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
1 Africa       52  0.322  0.346  0.488  0.570  0.663  0.684  0.515
2 Asia         33  0.513  0.555  0.621  0.649  0.710  0.815  0.814
3 Europe       30  0.865  0.795  0.747  0.746  0.722  0.770  0.846
```

En el código también hemos pedido un sumario del número de observaciones. Así podemos saber cuántos casos tenemos en cada correlación e inferir hasta qué punto pueden ser significativos estos resultados. Cuantos más casos tengamos, más significativas acostumbran a ser las observaciones. En todas las combinaciones de la tabla, la relación entre variables es positiva aunque siempre ha sido más fuerte en Europa que en los otros continentes. En África esta correlación era especialmente baja a mediados del siglo pasado, pero ha aumentado ligeramente a lo largo del tiempo.

Francis Anscombe y el problema de cuantificar relaciones

Cuantificar relaciones lineales también puede tener inconvenientes, tal como detectó el profesor de estadística Francis Anscombe con una base de datos sintética construida en 1973. Ilustró cómo relaciones completamente diferentes podían tener el mismo número de puntos, la misma media y desviación típica por *x* e *y* y la misma correlación. Fijémonos

en los cuatro diagramas de dispersión siguientes, obtenidos por medio del marco de datos `anscombe`, donde los cuatro tienen una *r* de 0,82.

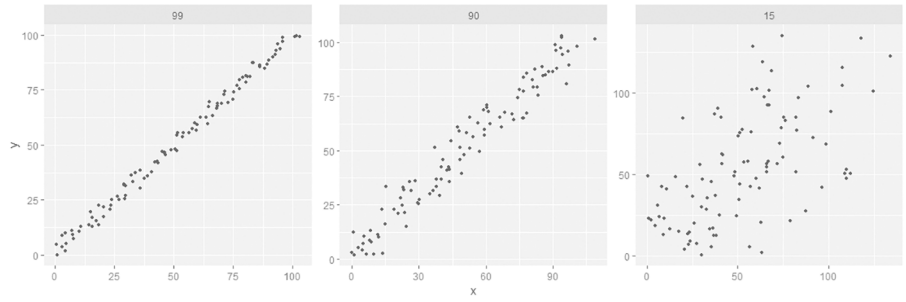


La conclusión más importante que podemos sacar de estos cuatro gráficos es que, aunque *r* nos pueda ayudar a cuantificar numéricamente una relación entre dos variables, siempre será pertinente visualizar la relación para evitar que los números nos hagan un mala pasada. Por ejemplo, en el último caso vemos que no hay ningún tipo de relación entre *x* e *y*, pero un solo caso extremo nos distorsiona el coeficiente de correlación. Podéis encontrar este marco de datos en los paquetes de base de R.

La segunda manera de poder cuantificar la relación entre dos variables es por medio del **coeficiente de determinación**, representado con R^2, que es un número entre 0 y 1 que nos determina la proporción de la variación en la variable dependiente que se puede explicar con la variación de los valores de la independiente. La suerte que tenemos es que R^2 es muy fácil de calcular. Solo hay que elevar al cuadrado el coeficiente de correlación. Lo que es más complicado, sin embargo, es comprender el significado de este número.

Para entender mejor cómo interpretar R^2, podemos considerar que su valor responde a una pregunta: ¿qué porcentaje de la variabilidad de *y* podemos explicar si sabemos la variabilidad de *x*? Fijémonos en la figura 7. A la izquierda vemos una correlación casi perfecta. Sabiendo cómo varían los valores de *x* podríamos adivinar muy cuidadosamente los valores de *y*. Por ejemplo, si nos dicen que el valor de *x* es 64, sabemos que el valor de *y* también estará alrededor de 64 y no estaremos muy lejos de equivocarnos. Si nos preguntan cómo de seguros estamos de acertar, entre 0 y 1, diremos que estamos un 0,99 convencidos de que no nos equivocaremos. La R^2 es 0,99 porque, a sabiendas de la variabilidad de *x*, podemos explicar un 99% de la variabilidad de *y*. En el gráfico de la derecha, en cambio, nos será mucho más difícil de acertar los valores de *y* si conocemos los valores de *x*. Si nos dicen que la *x* es 90, podremos estimar con mucha menos precisión los valores de *y*. Diríamos que nuestra R^2 estará alrededor de 0,15.

Figura 7. Coeficiente de determinación de tres correlaciones



Cuidado a la hora de interpretar R^2

Tenemos que ir con cuidado a la hora de interpretar R^2. Un coeficiente de determinación bajo no quiere decir necesariamente que tengamos un mal modelo y menos en ciencias sociales. A menudo podemos tener una R^2 baja, pero este pequeño porcentaje de explicación puede ser una aportación estadísticamente significativa por un problema complejo.

Cuando nos referimos a R^2, hablaremos siempre en porcentajes y nos expresaremos de una manera técnica. Por ejemplo, para describir el gráfico del medio de la figura anterior, diremos que «un 90% de la variabilidad de la variable dependiente puede ser explicada por la independiente».

Hay dos maneras de encontrar el coeficiente de determinación. La primera es elevar el coeficiente de correlación al cuadrado, como mostramos en el código siguiente. Podéis encontrar los códigos de la figura 7 en el anexo.

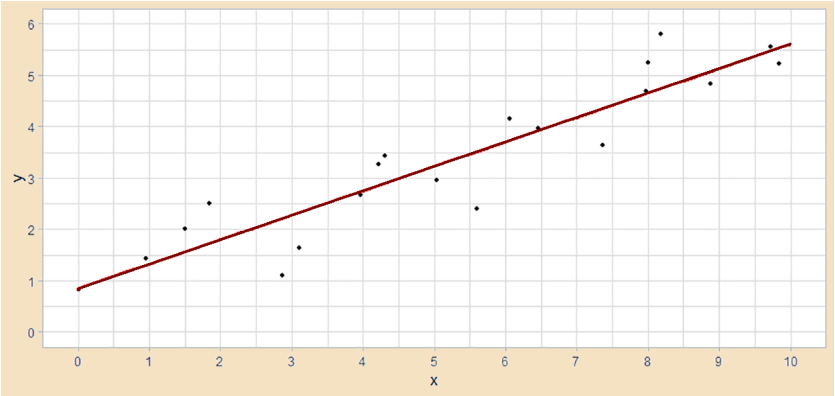
```
cor(x, y)^2
```

La segunda manera que tenemos para encontrar R^2 es a partir del modelo de regresión, destinado a construir modelos lineales, que explicamos en el apartado siguiente.

3.3. Modelar

Después de aprender a visualizar y cuantificar una correlación, en este apartado veremos cómo crear un modelo de regresión a partir de los datos disponibles. Modelar significa crear una fórmula matemática que nos permita predecir el valor de una variable si sabemos el valor de la otra. Esto lo podemos hacer con la recta de regresión OLS, que, como ya hemos explicado, tiene una doble virtud visual y cuantitativa. Visualmente, el método OLS nos muestra en un gráfico cuál es la manera más ajustada de describir la relación entre dos variables numéricas mediante una línea recta. De manera cuantitativa, la recta OLS nos permite predecir qué valor tomaría *y* cuando la recta pasa por los valores de *x*. Fijémonos en la figura 8. La recta nos ayuda a hacer una predicción del comportamiento de las variables *x* e *y* del marco de datos modelo, que encontraréis disponible con código en el anexo. La recta nos dice, por ejemplo, que si *x* tiene un valor de 2,5, *y* tendría que tener un valor cercano a 2.

Figura 8. Modelo lineal



En lugar de hacer estas estimaciones de manera visual, podemos cuantificar la recta OLS y obtener el modelo matemático a partir de dos parámetros: la inclinación de la recta de regresión y el valor que toma *y* cuando *x* es cero.

Interpretar correctamente la inclinación

Para interpretar correctamente el valor de la inclinación habrá que tener muy claro cuáles son las unidades de *x* y de *y*. Si por ejemplo la variable *x* son años e *y* son dólares per cápita, lo leeremos como «cada año (unidad de *x*) hay un cambio de dólares per cápita (unidad de *y*) del número que marque la inclinación».

1) Por **inclinación** nos referimos a cuánto crece o decrece *y* si aumentamos *x* en una unidad. Si nos fijamos en el gráfico anterior, cuando *x* pasa de 0 a 1, *y* cambia aproximadamente en medio punto, puesto que el valor de *y* pasa de algo menos de 1 a algo menos de 1,5. Entonces, la inclinación, que es la misma para toda la recta de regresión, será aproximadamente de 0,5.

2) El valor que toma *y* cuando *x* es 0 lo llamaremos **constante** o ***intercept***. A primera vista, podemos ver cómo la constante de esta regresión es un poco inferior a 1, puesto que es el punto por donde la recta corta el eje de las *y* cuando la *x* es igual a 0.

El nuevo tipo de objeto *lm*

La función `lm()` nos devuelve un tipo de objeto que no hemos visto hasta ahora de clase *lm*, como podemos observar con la función `class()`. Si pedimos el `typeof()`, veremos que es una lista, otro tipo de objeto como sería un vector o un marco de datos. Para transformar este objeto en vector, pondremos toda la operación dentro de la función `coef()`.

Para determinar con precisión estos dos parámetros, tendremos que utilizar la función `lm()`, donde indicaremos como primer argumento el nombre de la variable dependiente seguido de una virgulilla y el nombre de la dependiente. El segundo argumento es el marco de datos. Si obtenemos el modelo de regresión del marco de datos `model`, vemos que no nos equivocábamos mucho. La inclinación es de 0,475 y la *intercept* es de 0,854.

```
> lm(formula = y ~ x, data = model)
(Intercept)           x
      0.854       0.475
```

Con este resultado ya tenemos los dos datos que necesitamos para interpretar el modelo. Si nos fijamos en el *intercept*, diremos que cuando la *x* es 0, *y* es 0,854. Si nos fijamos en la inclinación, diremos que la *y* crece en 0,475 unidades por cada incremento de una unidad de *x*. Estos dos valores los podemos utilizar para crear un modelo predictivo que nos permitirá conocer una estimación de los valores de *y* si sabemos el valor de *x*. En el código siguiente, podemos predecir los valores de *y* con una fórmula sencilla, en que multiplicamos la pendiente por el valor de *x* y sumamos la constante.

```
y = intercept + x * pendiente
y = 0.854     + x * 0.475
```

Podemos probar nuestro modelo predictivo sustituyendo la *x* por diferentes números. Por ejemplo, si la sustituimos por 5, la *y* será 3,229 como comprobaremos si entramos el código `0.854 + 5 * 0.475` en la consola. Si la *x* es 10, la *y* será 5,604.

Aparte del *intercept* y la pendiente, hay mucha más información que podemos obtener a partir del modelo lineal. En el código siguiente hemos creado primero el objeto `lm_model` y después hemos generado la información adicional a partir de la función `summary()`. R no solamente nos genera la *intercept* y la pendiente, sino que también nos proporciona información del modelo de regresión como los datos utilizados, la especificación del modelo, los residuos, el error estándar asociado, la significación estadística o el R^2 múltiple y el R^2 ajustado.

```
lm_model <- lm(formula = y ~ x, data = model)
> summary(lm_model)
Call:
lm(formula = y ~ x, data = model)

Residuals:
     Min       1Q   Median       3Q      Max
-1.11730 -0.30072  0.03741  0.42807  1.07039

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.85402    0.32640   2.616   0.0181 *
x            0.47501    0.05263   9.026  6.8e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6241 on 17 degrees of freedom
Multiple R-squared:  0.8274, Adjusted R-squared:  0.8172
F-statistic: 81.47 on 1 and 17 DF,  p-value: 6.799e-08
```

En las líneas siguientes solo describiremos algunos de los elementos de este sumario. Al inicio del sumario nos aparece la fórmula que hemos utilizado, mientras que en la columna *Estimate* de la tabla de coeficientes podemos ver el *intercept* y la pendiente. El sumario también nos muestra el R^2 (*R-squared*), con la diferencia de que aquí vemos el R^2 múltiple y el R^2 ajustado. Si buscamos el R^2 de la manera que hemos explicado anteriormente por medio de `cor(model $x, model$y)^2`, veremos que su resultado coincide con el R^2 múltiple, que es la versión que utilizaremos en análisis bivariante para interpretar el R^2.

El elemento que nos faltaba ver, y que ya hemos visto en las otras técnicas de análisis bivariante, es el **test de significación** estadística, una convención matemática que nos permite descartar que los efectos observados en una relación lineal sean fruto del azar. Es decir, cuando observamos una relación entre dos variables, sea positiva o negativa, nos tenemos que preguntar si es posible que la relación observada no sea real cuando extrapolamos nuestra muestra a la realidad. Si no es real, quiere decir que hay una probabilidad superior al 5% de que la pendiente de la recta de regresión sea cero y que, por lo tanto, la hipótesis nula sea cierta. Sabemos que, en ciencias sociales, normalmente se establece un valor inferior en 0,05 para rechazar la hipótesis nula y así descartar el efecto del azar en nuestros resultados. La significación estadística la encontramos indicada en la columna con nombre *Pr(>|t|)* de la tabla de coeficientes. Veamos la tabla nuevamente:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.85402    0.32640   2.616   0.0181 *
x            0.47501    0.05263   9.026  6.8e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ejemplo de significación estadística

Supongamos que hemos viajado dos veces en la vida a Málaga y las dos veces ha llovido. Si, de acuerdo con nuestra experiencia, afirmamos que en Málaga siempre llueve, ¿qué validez tendría esta afirmación? Es muy posible que si incrementamos el número de experiencias en Málaga, nuestra afirmación sea más difícil de sustentar, puesto que encontraremos algún día que no llueva. Por lo tanto, es muy posible que nuestra afirmación inicial haya sido fruto del azar.

En la tabla de coeficientes nos fijaremos en el valor de *x* en la tabla *Pr(>|t|)*. Vemos que el número que nos aparece es 0,000000068, lo cual cumple el requisito de que el valor en cuestión sea inferior a 0,05. Por lo tanto, podemos descartar la hipótesis nula y asegurar con un 95% de confianza que los resultados observados no son fruto del azar. Si los valores son superiores a 0,05, diremos que la relación no es estadísticamente significativa.

Interpretar los asteriscos

Para facilitar la lectura de los resultados, tenemos la significación de los coeficientes señalados con asteriscos. Si, como nuestro caso, la significación está entre 0 y 0,001, se indicará con ***. Si la significación está entre 0,001 y 0,01 se indicará con ** mientras que si está entre 0,01 y 0,05 se indicará con *. En estos tres casos, diremos que la relación entre *x* e *y* es estadísticamente significativa.

4. Regresión múltiple

Como decíamos en la introducción de este módulo, podemos argumentar que dos variables están asociadas si cumplen varios requisitos. En las anteriores secciones hemos aprendido, principalmente, a dar argumentos estadísticos por medio de la cuantificación, por ejemplo, de la fuerza y el nivel de significación de una relación. Estos argumentos nos permiten decir si es posible que la relación observada suceda también en la realidad y cuál es la magnitud de esta relación. Otro requisito importante, sin embargo, es contraponer nuestra hipótesis con otras hipótesis alternativas. Y esta es la gran virtud de la **regresión múltiple**, que es una técnica de análisis multivariante que permite añadir nuevas variables independientes al modelo de regresión. Esta técnica nos ayuda a considerar el efecto que pueden tener otras variables, como *z* o *w*, en nuestra afirmación de que *x* tiene un efecto sobre *y*. Así, la regresión múltiple nos permite hacer afirmaciones más robustas sobre la relación entre *x* e *y*, porque podemos comprobar si *x* e *y* tienen relación incluso cuando controlamos los efectos de *x* para otras variables independientes como podrían ser *z* o *w*. Esta lógica la explicamos a continuación y acabaremos la sección con un ejemplo aplicado que relaciona el tipo de régimen y el voto afirmativo a la ONU.

4.1. Cuantificar

Si sabemos cómo se construye un modelo de regresión simple, la regresión múltiple es muy sencilla puesto que solo hay que añadir las variables adicionales con el signo + a la fórmula `lm()`. En el código siguiente hemos añadido la variable categórica binaria *z* a la relación y hemos pedido directamente un sumario. Hemos excluido de la impresión la parte de residuos.

```
> summary(lm(formula = y ~ x + z, data = model))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.94829    0.35723   2.655   0.0173 *
x            0.47556    0.05343   8.901 1.35e-07 ***
z           -0.20550    0.29111  -0.706   0.4904
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6335 on 16 degrees of freedom
Multiple R-squared:  0.8326, Adjusted R-squared:  0.8116
F-statistic: 39.78 on 2 and 16 DF,  p-value: 6.173e-07
```

La lectura de este sumario, donde encontramos incluida la nueva variable *z*, es muy parecida a la lectura en el análisis con dos variables. En primer lugar, la constante o *intercept* muestra qué valor toma *y* cuando el valor de todo el

resto de variables explicativas es 0. Es decir, cuando *x* toma el valor 0 y *z* toma el valor 0, el valor de *y* será de 0,94. En segundo lugar, el coeficiente de la variable numérica *x* nos dice que, por cada unidad que aumentamos de *x*, la *y* aumentará 0,48 y dejará constante el resto de variables independientes. Esta relación es estadísticamente significativa. En tercer lugar, el coeficiente de la variable categórica *z* nos explica el efecto que tiene sobre *y* aumentar los valores de *z* de 0 a 1 dejando constante el resto de variables independientes. Esto quiere decir que cuando *z* pasa de 0 a 1, disminuye 0,20 en nuestro modelo. La relación es negativa y no es estadísticamente significativa, por lo cual no podemos asegurar que esta relación de *z* sobre *y* exista si la generalizamos en una población más amplia.

En último lugar, en análisis multivariante nos fijaremos en el R^2 ajustado. Vemos cómo un 81% de la variabilidad de *y* puede ser explicada por las dos independientes del modelo. Si comparamos estos resultados con el ejercicio anterior, vemos cómo la introducción de *z* en nuestro modelo no nos ha aportado más capacidad de explicar *y*. No es estadísticamente significativa, ni tampoco nos ha hecho aumentar de manera sustancial el coeficiente de determinación, más bien al contrario. Por eso mismo, valdrá la pena descartar *z* de nuestro modelo. Antes, sin embargo, veremos cómo quedaría la fórmula si decidiéramos incluir la variable.

```
y = 0.94829 + 0.47556 * x -0.20550 * z
```

Es importante entender que la regresión múltiple tiene que servir esencialmente como un método que permite rechazar con más solidez la hipótesis nula. Es decir, este método nos permite decir que el efecto observado de la variable independiente principal sobre la variable dependiente se continúa produciendo con un intervalo de confianza superior al 95% incluso cuando controlamos por el efecto otras variables explicativas rivales.

4.2. Caso de estudio

En los apartados anteriores hemos podido observar, en contra de nuestras expectativas, que justamente son las no democracias las que votan más afirmativamente en la ONU que las democracias. La figura 3 nos daba una pista sobre esta tendencia. En este apartado testaremos la hipótesis de que las no democracias tienden a votar más favorablemente las resoluciones de la AGNU en relación con las democracias. Por eso, hemos diseñado un análisis de regresión múltiple donde contrastamos nuestra hipótesis con otras hipótesis rivales, como encontraréis explicado en la tabla 2. La hipótesis que queremos comprobar es que ser una democracia (*x*) tiene un efecto negativo sobre el voto afirmativo en la AGNU (*y*). Pensamos, sin embargo, que quizás esto está determinado por el nivel de desarrollo del país (*w*). Las democracias acostumbran a ser ricas y las autocracias acostumbran a ser pobres, por lo cual podría ser que el hecho de ser rico es el que haga votar de manera menos favorable a la AGNU y no el hecho de ser una democracia. Así, en el análisis multivariante controlaremos

los efectos de nuestra hipótesis principal por los efectos de la primera hipótesis alternativa, el nivel de desarrollo, medido por el PIB per cápita. Creemos que un PIB per cápita más grande afectará negativamente al voto afirmativo en la AGNU.

Como segunda hipótesis alternativa hemos considerado que el voto afirmativo en la AGNU (*y*) puede estar condicionado por la proporción de democracias en la asamblea (*z*). Así, los años que haya más democracias en la ONU el voto afirmativo será más desfavorable, y los años que haya menos el voto afirmativo será más favorable. Por lo tanto, creemos que la proporción anual más elevada de democracias en la AGNU tendrá una incidencia negativa en el voto afirmativo de las democracias.

Tabla 2. Análisis multivariante de voto en la AGNU

Variable	Medida y fuente	Codificación	Hipótesis
Variable dependiente (*y*)	Proporción de voto afirmativo en la AGNU (*unvotes*)	Cercano a 1, más voto afirmativo, cercano a cero, menos voto afirmativo (*vote*)	
Variable independiente principal (*x*)	Clasificación dicotómica de democracia (*psData*)	1 = Democracia, 0 = No democracia (*democracy*)	Más democrático, menos voto afirmativo
Variable independiente alternativa (*w*)	Desarrollo del país (*WDI*)	PIB per cápita PPP, precios constantes 2011 (*gdpcap*)	Más PIB per cápita, menos voto afirmativo
Variable independiente alternativa (*z*)	Proporción de democracias en la AGNU (*unvotes*)	Cercano a 1, más proporción de democracias, cercano a 0, menos proporción (*unga_prop)*	Más proporción de democracias, menos voto afirmativo

Para poder hacer este análisis multivariante, hemos creado el marco de datos `un_wdi`, disponible en el anexo, con las variables *vote*, *democracy*, *gdpcap* y *unga_prop*. A continuación, hemos pedido el sumario del modelo lineal. Podemos ver cómo *democracy* tiene una relación negativa y estadísticamente significativa con la variable dependiente cuando controlamos los efectos de las otras variables del modelo. Esto permite decir que hay relación con un 95% de confianza y nos permite descartar la hipótesis nula. También vemos que el *gdpcap* tiene una relación negativa y estadísticamente significativa con la variable dependiente, controlando los efectos de las otras variables. La variable *unga_prop* también tiene una relación negativa, tal como esperábamos, pero en este caso no tiene una relación estadísticamente significativa con *y*.

```
> summary(lm(formula = vote ~ democracy + gdpcap + unga_prop, data = un_wdi))

Call:
lm(formula = vote ~ democracy + gdpcap + unga_prop, data = un_wdi)


Residuales:
     Min       1Q   Median       3Q      Max
```

```
-0.91732 -0.07830  0.01938  0.10733  0.32181

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.805e-01  5.026e-02  19.509   <2e-16 ***
democracy   -1.042e-01  5.226e-03 -19.946   <2e-16 ***
gdpcap      -2.247e-06  1.442e-07 -15.585   <2e-16 ***
unga_prop   -1.113e-01  8.921e-02  -1.248    0.212
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1441 on 3171 degrees of freedom
Multiple R-squared:  0.1903, Adjusted R-squared:  0.1895
F-statistic: 248.4 on 3 and 3171 DF,  p-value: < 2.2e-16
```

Lo más importante de este análisis es que nuestra hipótesis continúa siendo cierta con un nivel suficiente de confianza incluso cuando la contrastamos con dos hipótesis rivales. Por lo tanto, la afirmación de que las no democracias tienden a votar más favorablemente las resoluciones de la AGNU que las democracias es más robusta después de este análisis multivariante. También podemos decir, observando el R^2, que nuestro modelo explica cerca de un 19% de la variabilidad de la variable dependiente. En análisis multivariante miraremos el R^2 ajustado. También es muy importante saber leer la columna *Estimate*, que nos dice el efecto de cada variable independiente sobre la dependiente. En el código siguiente traducimos esta columna a la fórmula matemática que nos permite tener un modelo predictivo.

```
y = 0.98 + - 0.10*democracy - 0.000002*gdpcap – 0.11*unga_prop
```

Según nuestro modelo, en un hipotético país con valor cero en la variable democracia (por lo tanto, que sea una no democracia), con el PIB per cápita de valor cero (por lo tanto, absolutamente pobre) y en una situación con cero democracias en la AGNU, este hipotético país votará afirmativo un 98% de las veces. En cambio, si cambiamos el régimen del país a democrático (la variable *x* pasa de 0 a 1) y dejamos las otras variables constantes, este hipotético país votará a favor un 88% de las veces, puesto que en nuestro modelo hay un cambio del 10% en el voto según el tipo de régimen. En cuanto al nivel de desarrollo, a medida que el PIB per cápita se incrementa, la proporción de voto afirmativo disminuye. Concretamente, la proporción de voto afirmativo disminuye un 2% por cada 10.000 dólares adicionales per cápita que tenga el país. Finalmente, cuantas más democracias haya en la AGNU, menos votará a favor el país. Traducido en números: de una hipotética situación de una AGNU con solo democracias a una hipotética situación de una AGNU con solo no democracias, el voto hipotético disminuye un 10%. Esta diferencia, sin embargo, no es estadísticamente significativa.

Resumen

Hay una frase atribuida al primer ministro británico Benjamin Disraeli y popularizada por el célebre escritor Mark Twain que dice:

«Hay tres tipos de mentiras: mentiras, mentiras podridas y la estadística.»

La estadística es una herramienta muy poderosa para explicar asociaciones entre fenómenos, pero también es una herramienta muy peligrosa si no interpretamos su utilidad correctamente. Al fin y al cabo, lo único que nos sugiere esta metodología es que parece que los datos nos den la razón, pero esto no quiere decir que tengamos la razón. Pensémoslo con un ejemplo. Podemos decir que, a medida que un país se hace más democrático, acostumbra a generar también más riqueza. Esta afirmación puede ser confirmada con las técnicas de análisis de datos que hemos aprendido en este módulo. Lo que no podemos afirmar con seguridad es si es la democracia la que causa la riqueza o bien si es la riqueza la que causa la democracia. En nuestro análisis multivariante, donde habremos incluido varias hipótesis alternativas en nuestra hipótesis principal, tampoco podemos decir en realidad que no haya otras hipótesis rivales y plausibles que no hayamos incluido en el modelo y que puedan hacer tambalear nuestras conclusiones. Es por eso que el análisis de datos estadístico es un primer paso, necesario pero no suficiente, para responder a las preguntas que nos hacemos.

Para saber más

El libro *Models, Numbers, and Cases* de Sprinz y Wolinsky-Nahmias (2004, pág. 129-226) contiene un repaso destacable de metodología cuantitativa en estudios internacionales.

Mansfield y Pevehouse (2008) prepararon un resumen para el *The Oxford Handbook of International Relations*.

En este sentido, la interpretación de los datos tiene un papel fundamental, en especial en estudios basados en la observación de fenómenos como los estudios internacionales. Hemos de tratar no sugerir erróneamente que la correlación entre dos fenómenos implica que uno cause el otro. Tampoco hay que llegar al punto de obsesionarse con observar solo los niveles de significación estadística y obviar otra información sustantiva (Braumoeller y Sartori, 2004, pág. 131). Los datos nos pueden dar resultados significativos pero con efectos estadísticamente minúsculos, simplemente, por el hecho de que disponemos de una población enorme de casos. Es por eso que tenemos que apreciar también las bondades que tiene la visualización de los datos, que nos puede ayudar a observar rasgos sustantivos que nos ayuden a formular nuevas preguntas relevantes.

Correlaciones absurdas

Podéis ver algunos ejemplos de correlaciones absurdas en este web (http://www.tylervigen.com/spurious-correlations), que muestra que el número de personas que se ahogaban en una piscina anualmente entre 1999 y 2009 estaba correlacionado con el número de películas anuales en que aparecía Nicolas Cage.

En resumen, este módulo ha explicado algunas de las principales técnicas estadísticas para el análisis de datos como un instrumento útil para explicar fenómenos. Pero tan importante es utilizar buenas técnicas, sustentadas matemáticamente, como utilizar buenas razones, sustentadas teóricamente, para

tratar de demostrar un argumento. Es por eso que una interpretación humilde de los datos y una comunicación clara y transparente ayudarán a añadir valor a nuestros hallazgos como analistas de datos.

Ejercicios de autoevaluación

Para un mejor aprendizaje, intentad hacer mentalmente el máximo de ejercicios posible, sin utilizar R.

1. Deduce la técnica que utilizaremos según estas dos variables.

```
database$black_white, database$blue_red
```

2. Deduce la técnica que utilizaremos según estas dos variables.

```
database$gdp, database$unemployment
```

3. Pide la tabla de contingencia de estas variables.

```
wdi$class, wdi$adult
```

4. Pide la tabla de contingencia con las proporciones en las filas para estas variables.

```
wdi$gender, wdi$vote
```

5. Cambia el código siguiente para visualizar las proporciones de la tabla de contingencia.

```
ggplot(wdi, aes(country, democracy)) + geom_bar()
```

6. ¿Cuál de los resultados siguientes no puede devolver un test de Cramer?

```
0.45, 0.15, 1.10
```

7. Hemos pedido el test de Khi-cuadrado de Pearson. ¿Cuál de los p-valores siguientes es estadísticamente significativo?

```
0.01, 0.96, 0.06
```

8. Pide la diferencia de medias entre estas dos variables.

```
wdi$country, wdi$vote
```

9. Cambia los atributos de la geometría para pedir una recta LOESS.

```
geom_smooth()
```

10. Indica cuál de las correlaciones siguientes es la más fuerte.

```
-0.921, 0.826, 0.656
```

11. ¿Qué R-cuadrado tendremos con el coeficiente de correlación siguiente?

```
1
```

12. Di cuál es el valor de y cuando la x es cero en esta ecuación.

```
y = 450 + x * -0.3
```

13. ¿Cuál es el efecto en y por cada unidad que aumentamos de x?

```
y = 450 + x * -0.3
```

14. Indica los asteriscos que veremos con el p-valor siguiente.

```
0.002
```

15. Pide el análisis multivariante de las variables siguientes.

```
mdata$conflict / mdata$ethnic, mdata$gdp, mdata$mspend
```

Solucionario

1. `Tabla de contingencia`

2. `Regresión lineal / correlación`

3. `table(wdi$class, wdi$adult)`

4. `prop.table(table(wdi$gender, wdi$vote), 1)`

5. `ggplot(wdi, aes(country, democracy)) + geom_bar(position = "fill")`

6. `1.10`

7. `0.01`

8. `t.test(wdi$country, wdi$vote)`

9. `geom_smooth()` (no se tiene que cambiar nada)

10. `-0.921`

11. `1`

12. `450`

13. `-0.3`

14. `**`

15. `lm(formula = conflict ~ ethnic + gdp + mspend, data = mdata)`

Glosario

`chisq.test()` Pide la prueba del Khi-cuadrado de Pearson.

`colSums()` Suma las columnas de una tabla de contingencia.

`cor()` Devuelve el coeficiente de correlación de dos variables.

`CramerV()` Pide el test de Cramer V para tablas de contingencia (*DescTools*).

`t.test()` Pide el t-test para diferencia de medias.

`lm()` Devuelve el modelo de regresión lineal simple o múltiple, según el número de variables que utilizamos.

`prop.table()` Muestra las proporciones de una tabla de contingencia.

`rowSums()` Suma las filas de una tabla de contingencia.

`table()` Muestra una tabla de contingencia de dos variables categóricas.

Bibliografía

Agresti, A.; Finlay, B. (2009). *Statistical Methods for the Social Sciences*. Nueva Jersey: Pearson.

Babbie, E. R. (2013). *The practice of social research*. Wadsworth: Cengage Learning.

Braumoeller, B. F.; Sartori, A. E. (2004). *The promise and perils of statistics in international relations*. En: D. F. Sprinz; Y. Wolinsky-Nahmias, (eds.). *Models, Numbers, and Cases: Methods for Studying International Relations*. Ann Arbor, MI: University of Michigan Press (pág. 129–151).

Cheibub, J. A.; Gandhi, J.; Vreeland, J. R. (2010). *Democracy and Dictatorship Revisited*. *Public Choice* (vol. 143, núm. 2-1, pág. 67-101).

Halperin, S.; Heath, O. (2016). *Political Research: Methods and Practical Skills*. Oxford: Oxford University Press.

Johnson, J. B.; Reynolds, H.; Mycoff, J. (2007). *Political Science Research Methods*. Washington, DC: CQ Press.

King, G.; Keohane, R. O.; Verba, S. (1994). *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press.

Mansfield, E. D.; Pevehouse, J. C. (2008). *Quantitative Approaches*. En: C. Reus-Smit,; D. Snida, (eds.). *The Oxford Handbook of International Relations*. Oxford: Oxford University Press (pág. 481–498).

Sprinz, D. F.; Wolinsky-Nahmias, Y. (2004). *Models, Numbers, and Cases: Methods for Studying International Relations*. Ann Arbor, MI: University of Michigan Press.

Voeten, E. (2017). *Data and Analyses of Voting in the UN General Assembly*. Routledge Handbook of International Organization. Routledge.

Anexo del módulo

Código de la figura 2

```
un_dd_year %>%
  filter(between(year, 1995, 2000)) %>%
  ggplot(aes(x = democracy, y = mean, col = continent)) +
  geom_jitter(alpha = 0.5) +
  stat_summary(geom = "point", fun.data = mean_se,
               col= "red", size = 1.2) +
  stat_summary(geom = "errorbar", fun.data = mean_se,
               col = "red", width = 0.1) +
  theme_classic()
```

Código de la figura 3

```
un_dd_year %>%
  filter(between(year, 1946, 2000)) %>%
  ggplot(aes(x = factor(year), y = mean, col = democracy)) +
  geom_jitter(alpha = 0.2) +
  stat_summary(fun.data = mean_se, col = "red",
               size = 0.2) +
  scale_y_continuous(limits = c(0.3, 1)) +
  scale_x_discrete(name = "Year",
    breaks = c(1946, seq(1950, 2000, 5))) +
  theme(legend.position = c(0.9, 0.2),
        legend.background = element_rect(fill=alpha("light blue", 0.5)))
```

Código de la figura 4

```
q <- seq(1,100,1)
w <- (50 - q)^2
s <- sample(100)

visual_corr <- data.frame(graph1_x = q,graph1_y = s,
           graph2_x = -jitter(seq(1,100,1), 100),
           graph2_y  = jitter(seq(1,100,1), 100),
           graph3_x = abs(c(jitter(seq(1,99,1), 50), 5)),
           graph3_y = jitter(seq(1,100,1), 50),
           graph4_x = jitter(q, 20),
           graph4_y = jitter(w, 1500),
           graph5_x = abs(jitter(q, 10)),
           graph5_y = abs(jitter(log(q), 10)),
           graph6_x = jitter(q, 10),
```

```
            graph6_y = jitter(q, 10))

visual_corr1 <- visual_corr %>%
  gather("graph", "value") %>%
  separate(graph, c("tabla", "codigo"), sep = "_") %>%
  group_by_at(vars(-value)) %>%
  mutate(row_id=1:n()) %>% ungroup() %>%
  spread(key=codigo, value=value) %>%
  select(-row_id)

visual_corr1 %>%
  ggplot(aes(x, y)) +
    geom_point() +
    facet_wrap(~ tabla, scales = "free")
```

Código del apartado 3.1

```
gpm7 <- gapminder %>%
  filter(year == 2007, between(lifeExp, 47, 60), gdpPercap > 9000)

gapminder %>% filter(year == 2007) %>% ggplot(aes(gdpPercap, lifeExp)) + geom_point() +
geom_point(data = gpm7, aes(gdpPercap, lifeExp), col = "red") + theme_classic()
```

Código de la figura 7

```
R2 <- data.frame(x_99 = abs(jitter(seq(1,100,1), 15)),
            y_99 = abs(jitter(seq(1,100,1), 15)),
            x_90 = abs(jitter(seq(1,100,1), 50)),
            y_90 = abs(jitter(seq(1,100,1), 50)),
            x_15 = abs(jitter(seq(1,100,1), 200)),
            y_15 = abs(jitter(seq(1,100,1), 200)))

R2x <- R2 %>%
  gather("graph", "value") %>%
  separate(graph, c("codigo", "R2"), sep = "_") %>%
  group_by_at(vars(-value)) %>%
  mutate(row_id=1:n()) %>% ungroup() %>%
  spread(key=codigo, value=value) %>%
  select(-row_id) %>%
  arrange(desc(R2))

R2x %>%
  ggplot(aes(x, y)) +
  geom_point(size = 1) +
  facet_wrap(~ factor(R2, levels = c("99", "90", "15")), scales = "free")
```

Código de la figura 8

```
model <- data.frame(x = c(0.96, 1.51, 1.84, 2.87, 3.10, 3.96, 4.22, 4.31, 5.03, 5.60, 6.06, 6.46,
7.36, 7.98, 8.18, 8.01, 8.88, 9.84, 9.73), y = c(1.44, 2.00, 2.51, 1.10, 1.64, 2.67, 3.26, 3.44,
2.95, 2.41, 4.16, 3.96, 3.64, 4.68, 5.81, 5.24, 4.84, 5.22, 5.56))

ggplot(model, aes(x, y)) + geom_point(size = 1) + geom_smooth(method = "lm", col = "red",
fullrange = TRUE, se = FALSE) + scale_x_continuous(breaks = c(seq(1,10))) +
scale_y_continuous(breaks = c(seq(1,6)))
```

Código del apartado 4.2

```
library(WDI)
WDIsearch("gdp per capita.*constant")
gdp_wdi <- WDI(indicator = "NY.GDP.PCAP.PP.KD", start = 1946)

un_year_vote <- un_roll_calls %>%
  separate(date, "year", extra = "drop") %>%
  inner_join(un_votes) %>%
  mutate(year = as.numeric(year),
         vote = if_else(vote == "yes", 1, 0)) %>%
  group_by(country, country_code, year) %>%
  summarize(vote = mean(vote))

un_vote_dem <- un_year_vote %>%
  inner_join(DDdata, by = c("country_code" = "iso2c", "year" = "year")) %>%
  select(country = country.x, country_code, year, vote, democracy)

un_prop <- un_vote_dem %>%
  group_by(year) %>%
  summarize(unga_prop = mean(democracy))

un_wdi <- un_vote_dem %>%
  left_join(un_prop, by = "year") %>%
  inner_join(gdp_wdi, by = c("country_code" = "iso2c", "year" = "year")) %>%
  select(country = country.x, year, vote, democracy, gdpcap = NY.GDP.PCAP.PP.KD, unga_prop) %>%
  filter(gdpcap != is.na(gdpcap))

summary(lm(formula = vote ~ democracy + gdpcap + unga_prop, data = un_wdi))
```