Anàlisi bivariant

PID_00268322

Jordi Mas Elias

Temps mínim de dedicació recomanat: 4 hores

Universitat Oberta de Catalunya

Jordi Mas Elias

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats pel professor: Jordi Mas Elias (2019)

Primera edició: setembre 2019

Av. Tibidabo, 39-43, 08035 Barcelona
Realització editorial: FUOC

Índex

Introducció..... 5

1. Relació entre variables categòriques..... 9
1.1. Visualització amb taula 9
1.2. Visualització gràfica 10
1.3. Quantificar 12

2. Relació entre numèrica i categòrica..... 15
2.1. Visualitzar 16
2.2. Quantificar 18

3. Relació entre variables numèriques..... 20
3.1. Visualitzar 21
3.2. Quantificar 23
3.3. Modelar 26

4. Regressió múltiple..... 30
4.1. Quantificar 30
4.2. Cas d'estudi 31

Resum..... 34

Exercicis d'autoavaluació..... 35

Solucionari..... 37

Glossari..... 38

Bibliografia..... 39

Annex..... 40

Introducció

L'anàlisi bivariant s'ocupa de l'anàlisi estadística amb dues variables. A vegades és fàcil confondre les funcions de l'anàlisi bivariant amb l'anàlisi univariant, ja que aquesta última també admet sovint l'anàlisi amb dues variables. Distingirem entre una i l'altra de la següent manera: a l'anàlisi bivariant el nostre objectiu és establir relacions entre variables mentre que a l'anàlisi univariant no tenim aquest propòsit. Amb univariant, podem treballar amb més d'una variable simplement perquè fem comparacions entre subgrups i volem descriure similituds i diferències d'aquesta variable entre subgrups (Babbie, 2013). Per exemple, podem mirar quin és el PIB per càpita en varis continents sense voler dir que estiguem insinuant que viure en un continent determinat tingui una relació directa amb la riquesa d'un país. En canvi, amb l'anàlisi bivariant sí que tenim la intenció de buscar com dues variables es relacionen entre si. En altres paraules, amb l'anàlisi univariant no anem més lluny de la comparabilitat si treballem amb dues variables, mentre que en el bivariant estem intentant suggerir que hi ha algun tipus de causalitat. Quan parlem, doncs, d'associacions o efectes entre variables, només ho farem mitjançant les tècniques de l'anàlisi bivariant, que ens permetrà suggerir des d'un punt de vista estadístic si el comportament d'una variable està parcialment determinat per l'altra (King i altres, 1994).

Establir associacions entre dues variables és una pràctica freqüent en el dia a dia. Acostumem a dir que la contaminació té un efecte sobre l'escalfament global, que els països democràtics solen ser més rics o que tenir una ideologia de dretes influeix a l'hora de votar un partit determinat. En altres paraules, creiem que si coneixem els valors de la primera variable (nivells de contaminació, democràcia o ideologia) podrem saber alguna cosa sobre els valors de la segona variable (escalfament global, riquesa o vot a partit) perquè creiem que estan relacionades. De la primera variable, que és la que pensem que causa l'efecte, en direm **variable independent** i la il·lustrarem amb el símbol x. A la segona variable, que és la que pensem que està afectada per la variable independent, l'anomenarem **variable dependent** i la il·lustrarem amb el símbol y.

Altres denominacions de les variables

Hi ha altres maneres de referir-se a les variables independent i dependent. A la variable independent se l'anomena variable explicativa, mentre que a la variable dependent se la denomina variable explicada o de resposta.

Per visualitzar i quantificar estadísticament una associació bivariant utilitzarem una tècnica diferent segons si les variables independent i dependent són numèriques o categòriques. En la taula 1 resumim la tècnica utilitzada segons el tipus de variable i de si ocupa la posició d'independent o dependent. Si tant la variable independent com la variable dependent són categòriques, utilitzarem una taula de contingència. Si, en canvi, tant la variable independent com la variable dependent són numèriques, utilitzarem la regressió i la correlació.

Si la variable independent és categòrica i la dependent és contínua, farem servir la diferència de mitjanes. Finalment, si la variable independent és contínua i la variable dependent és categòrica, farem una regressió logística.

Taula 1. Tècniques d'anàlisi bivariant

		Variable independent	
		Categòrica	**Numèrica**
Variable dependent	**Categòrica**	Taula de contingència	(Regressió logística)
	Numèrica	Diferència de mitjanes	Regressió i correlació

En aquest mòdul no explicarem la regressió logística. Les pàgines següents descriuran les altres tres tècniques d'anàlisi bivariant:

- la taula de contingència,
- la diferència de mitjanes i
- la regressió i correlació.

Cada apartat mostra primer com representar visualment una relació bivariant segons si les variables independent i dependent són numèriques o categòriques, i a continuació explica la manera de quantificar la relació.

Encara que les tècniques utilitzades canviïn segons els tipus de variable, totes les combinacions tenen uns requisits similars per a indicar si dues variables tenen relació.

1) El primer requisit, que és el que tractarem amb més profunditat en aquestes pàgines, és que l'associació observada tingui **suport estadístic**. A vegades podem observar associacions pocs consistents entre variables. Per exemple, si tirem una moneda i surt dues vegades cara, seria una inferència dèbil fer el supòsit que sempre sortirà cara. Des d'un punt de vista estadístic, basat en les teories de la probabilitat, per fer inferències a partir d'una mostra necessitem un nombre prou elevat de casos i/o que la variació de la variable dependent quan varia la independent sigui prou gran. Per tant, la relació observada ha de ser prou consistent com perquè no sigui gaire probable pensar que realment no hi ha relació en el fenomen observat.

2) El segon requisit es refereix a la **temporalitat**. És a dir, per lògica, la variable independent s'ha d'haver produït abans o més o menys al mateix temps que la variable dependent.

3) I finalment, per a poder dir que variable té un efecte sobre un altre no solament cal que l'estadística i la temporalitat ens donin la raó, sinó que la **teoria** també ens la doni. Aquest és el requisit més difícil d'aconseguir i per raons d'espai no treballarem de manera específica aquesta part en aquest mòdul. No obstant això, cal tenir clar que per a suggerir que dues variables tenen relació

caldrà que aquesta relació sigui lògica i plausible i per això, és important haver descartat les altres alternatives que, teòricament, poden explicar el fenomen que volem estudiar.

La necessitat de suport teòric en l'anàlisi

La relació entre riquesa i democràcia és un bon exemple per a explicar la necessitat de teoria. Si les dades ens diuen que els països democràtics són més rics que els països no democràtics, això no vol dir que hi hagi un efecte entre ser una democràcia i ser ric. Podria ser que l'efecte fos a la inversa: que ser ric et faci ser més democràtic. O bé que hi hagi una altra variable que desconeixem que afecti la riquesa i la democràcia a la vegada, però que democràcia i riquesa no s'afectin entre elles. Trobareu aquesta idea ampliada a Halperin i Heath (2016, pàg. 369-370).

1. Relació entre variables categòriques

La **taula de contingència** és la manera de representar una relació bivariant quan les variables independent i dependent són categòriques. Normalment representarem els valors de la variable independent a l'eix horitzontal i els valors de la dependent a l'eix vertical, de manera que cada cel·la de dins la taula representa el recompte total d'observacions que cauen a cada combinació entre nivells de les dues variables. Un cop tinguem a cada cel·la el nombre de freqüències, calcularem els percentatges de cada freqüència sobre el total de la columna.

Per il·lustrar la taula de contingència utilitzarem en aquesta secció el paquet *unvotes* (Voeten, 2017), que conté informació sobre les votacions a l'Assemblea General de Nacions Unides (AGNU) separades en tres paquets. El primer que farem és instal·lar i carregar el paquet i demanarem l'estructura dels tres marcs de dades.

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

El paquet *unvotes*

El marc de dades `un_votes` té un registre de totes les votacions a l'AGNU segons si cada estat va votar a favor, en contra o es va abstenir. El marc de dades `un_roll_calls` té informació més concreta sobre cada votació, com el número de resolució i una descripció del contingut de la votació. El marc de dades `un_roll_calls_issues` separa les votacions segons la temàtica.

En aquest apartat, utilitzarem la taula de contingència per respondre a la pregunta següent sobre la informació continguda al paquet *unvotes*: com sabeu, Sudan del Sud és l'últim país que ha ingressat a l'Organització de les Nacions Unides (ONU) després d'aconseguir la independència el 2011. Partim de la premissa que Sudan del Sud s'ha volgut diferenciar respecte de Sudan en política internacional i volem comprovar la nostra hipòtesi mirant si aquests dos països han votat diferent a l'AGNU a partir de la independència de Sudan del Sud.

1.1. Visualització amb taula

En la taula de contingència situarem dues variables categòriques:

- La independent és *country*, situada a les columnes, que pot adoptar els valors Sudan o Sudan del Sud.

- La dependent és *vote*, situada a les files, que pot prendre els valors Sí o No.

Treure nivells dels factors

Com que la variable *vote* és un factor, quan eliminem tots els valors d'una de les categories (en aquest cas la categoria *abstain*) no ens elimina el nivell corresponent. Per a eliminar el nivell, haurem d'incloure la funció `droplevels()`. Proveu com es visualitza la taula de contingència si traieu aquesta funció del codi.

Creiem que cada país té un patró de vot diferent, de manera que si coneixem els valors de la variable independent *country* podrem saber alguna cosa sobre els valors de la dependent *vote*. Per això, en el codi següent hem filtrat les dades pels dos països que volem mirar i pel número de votació 5117 a partir de la qual va participar Sudan del Sud a l'AGNU. També hem eliminat les abstencions

i hem demanat només les votacions relacionades amb el desenvolupament econòmic. Per a visualitzar la taula de contingència, només cal indicar els dos vectors categòrics com a arguments de la funció `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 cel·la de la taula que hem generat representa una combinació del recompte total d'observacions entre categories de les dues variables. Sembla que Sudan ha votat més vegades i més a favor que no pas Sudan del Sud, ja que ha votat de manera favorable en 16 ocasions i en contra 3 vegades. Sudan del Sud ha votat de manera favorable en 11 ocasions i en contra en 3 ocasions.

Sumar files o columnes

Si volem veure la suma de les files o les columnes de la taula de contingència, haurem d'introduir la funció anterior a dins de la funció `rowSums()` o `colSums()`. R ens retornarà un vector numèric amb la suma de cada fila o columna.

Encara que des d'un punt de vista descriptiu és bo veure quantes observacions tenim a cada combinació, l'anàlisi d'una taula de contingència la farem de manera més acurada amb proporcions. Si sabem els percentatges ens serà més fàcil comparar els valors entre columnes i distingir les probabilitats que la variable dependent prengui un valor determinat si coneixem els valors de la independent. Per veure les proporcions utilitzarem la funció `prop.table()` amb l'argument `margin = 2`, que ens permetrà veure els percentatges en columnes. També hem multiplicat la taula per 100 per a veure el resultat en tants per cent i hem utilitzat `round()` per a arrodonir els valors a un decimal.

```
> 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
```

Proporcions per columnes, files o totals

La millor manera de veure una taula de contingència és amb les proporcions per columnes. No obstant això, també podem veure les proporcions per files si introduïm l'argument `margin = 1`. Si no introduïm cap argument, veurem el percentatge de cada cel·la sobre el total de cel·les de la taula.

Vista amb proporcions, la taula de contingència ens dona un altre angle de les dades en comparació amb la taula de contingència vista en freqüències. Des de 2011, Sudan del Sud tendeix a votar les resolucions sobre desenvolupament econòmic més favorablement que no pas Sudan. Els resultats, doncs, sembla que de moment no contradiuen la nostra hipòtesi de partida.

1.2. Visualització gràfica

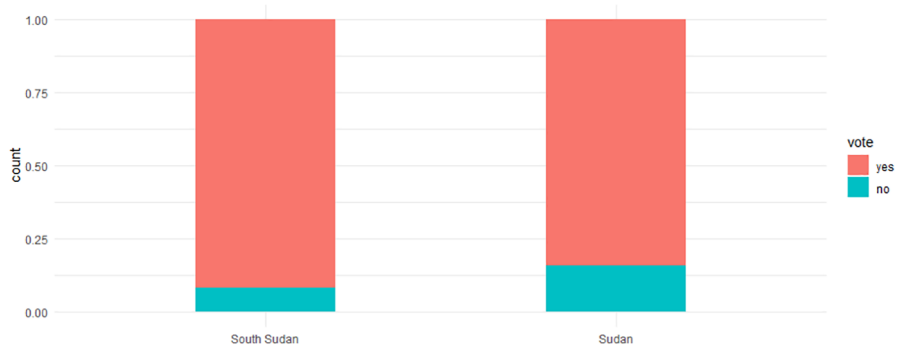
El diagrama de barres és la manera més habitual de visualitzar gràficament les taules de contingència. Si utilitzem el paquet *ggplot2* farem servir la geometria `geom_bar()`. En primer lloc visualitzarem el recompte d'observacions.

Proveu, en el codi següent, de canviar la posició per defecte de la geometria `position = "stack"` per `position = "dodge"`, que ens mostrarà horitzontalment les diferents categories de cada columna.

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

Si volem visualitzar la taula de contingència en proporcions, només caldrà que canviem la posició de la geometria del codi anterior. En aquest cas, situarem la `position = "fill"`.

Figura 1. Visualització gràfica d'una taula de contingència



En el codi d'aquesta taula també hem disminuït l'amplada de les barres amb l'argument `width = 0.4` i hem eliminat el nom de l'eix horitzontal amb `xlab(NULL)`.

Diferència de vot amb les administracions Carter i Reagan

Com va canviar el vot dels Estats Units a Nacions Unides entre les administracions Carter i Reagan? En aquest codi teniu la resposta.

```
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")
```

Fixeu-vos com hem filtrat les votacions per a poder separar la mostra en dos períodes, el que va governar Carter i el que va governar Reagan.

1.3. Quantificar

Ja hem explicat que quantificar una relació va més enllà de descriure i comparar dues variables (d'això ja se n'encarrega l'anàlisi univariant). És important remarcar que un dels objectius crucials de la quantificació bivariant és saber si les dades donen suport o no a la nostra hipòtesi inicial, que acostuma a ser que les dues variables estan relacionades. En cas que els resultats confirmin la hipòtesi mitjançant els procediments que veurem a continuació, és molt important no córrer: això no voldrà dir que la hipòtesi sigui correcta. Si és correcta o no mai no ho podrem saber del cert. L'únic que voldrà dir és que els resultats no semblen indicar que les variables són independents entre si i per tant no tenen relació. Aquestes dues frases anteriors semblen un embolic, però tenen molt de sentit. Per això és important que entenguem el significat de la **hipòtesi nul·la**, que és el que realment comprovem estadísticament en una anàlisi bivariant. La hipòtesi nul·la acostuma a ser la hipòtesi contrària a la nostra hipòtesi inicial, també dita **hipòtesi alternativa**. És a dir, si diem que hi ha relació entre *a* i *b*, la hipòtesi nul·la que testarem és que no hi ha relació entre *a* i *b*.

Hipòtesi nul·la

Per a entendre la hipòtesi nul·la, hem de partir de la base que mai no podem estar segurs que el que volem provar sigui cert. Moltes troballes de grans científics que en el seu temps van ser un gran descobriment han estat falsificades per científics més contemporanis. Això indica que hem d'anar molt en compte amb la nostra recerca i que, en lloc de provar que dues variables tenen relació, sempre és més prudent descartar el contrari: que les dues variables no tenen relació. En l'anàlisi de dades, doncs, sempre comprovarem la hipòtesi nul·la, i els resultats ens permetran descartar o no que no hi hagi diferència des d'un punt de vista estadístic.

Quan quantifiquem, doncs, el que fem és mirar d'aportar arguments estadístics per a desmuntar la hipòtesi nul·la que no hi ha relació entre les variables. El primer argument que podem aportar és la diferència entre els percentatges de la taula de contingència: observem un 7,46% de diferència en el percentatge de vot afirmatiu respecte del vot negatiu en temes de desenvolupament entre Sudan del Sud i Sudan. Aquest argument és més aviat descriptiu i estadísticament no és suficient, encara que no deixi de ser un nombre que ens pugui ajudar a explicar les diferències entre categories de la relació que analitzem.

Un segon argument que podem aportar, més associat amb l'anàlisi explicativa, és la força de la relació entre les dues variables. Direm que una relació entre variables categòriques és molt forta quan per mitjà dels valors de la variable independent podem saber amb molta exactitud els valors que prendrà la variable dependent. Hi ha diversos mètodes per a mesurar la força de la relació, com el test de Cramer o el de Lambda. La majoria d'aquests mètodes, que es diuen formalment **coeficients d'associació**, retornen un valor proper a 1 si la relació té molta força i un valor proper a 0 si la relació no té força. Per a utilitzar aquestes eines, haurem d'instal·lar i carregar el paquet *DescTools*. En el codi següent hem demanat el test de Cramer amb `CramerV()`, on introduïm a dins de la funció un argument per cada variable.

Exemples de la força de la relació

Si tots els països pobres del món fossin autocràcies i tots els països rics fossin democràcies, llavors diríem que la relació entre nivell de desenvolupament i règim polític és molt forta. És a dir, si sabem els valors de la variable *desenvolupament* podrem endevinar amb tota seguretat els valors de la variable *democràcia*. Això, però, rarament passa, sobretot, en ciències socials. Sempre tindrem casos de democràcies pobres i autocràcies riques, així com també tindrem casos de persones ideològicament de dretes que votin partits d'esquerres o d'anys que els nivells de contaminació han augmentat, però en canvi la temperatura del planeta ha disminuït.

```
library(DescTools)

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

L'operació ens haurà retornat un coeficient d'associació de Cramer proper a 0,11. Això vol dir que si sabem els valors de x tenim un 11% de probabilitats d'encertar els valors de y. Certament, no és una probabilitat d'encert gaire alta.

Vota diferent Israel quan es tracten temes de l'Orient Mitjà?

Trobareu un coeficient d'associació de Cramer més alt que en l'exemple anterior si mireu l'orientació del vot d'Israel en temes relacionats amb l'Orient Mitjà en comparació amb la resta de temes de l'AGNU. Primer hem creat l'objecte `israel_un` i després hem creat una taula de contingència i demanat 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)
```

Fixeu-vos en la taula de contingència i el coeficient d'associació de Cramer. En aquest cas, la relació té molta més força perquè coneixent els valors de la variable independent (si el tema és sobre l'Orient Mitjà o no) tindrem prop d'un 33% de possibilitats d'encertar la orientació del vot d'Israel.

A part de la força de la relació, l'altra mesura rellevant per a quantificar una relació és saber fins a quin punt és significatiu el nostre resultat. Tornant a l'exemple anterior de la moneda, les conclusions que puguem treure sobre el comportament d'una moneda si la tirem dues vegades a l'aire seran poc significatives comparades amb les conclusions que puguem treure sobre la mateixa moneda si la tirem cent vegades. Amb l'exemple anterior dels dos Sudans ens passa una cosa semblant. Suposem que demà hi ha una votació a l'ONU i Sudan del Sud vota en contra d'una resolució sobre desenvolupament econòmic. Això canviaria els percentatges de Sudan del Sud a la taula de contingència d'una manera prou significativa: els percentatges passarien a ser aproximadament del 85% a favor i del 15% en contra, exactament els mateixos percentatges que té Sudan. Si només un sol cas addicional enderroca la nostra hipòtesi que Sudan del Sud s'ha volgut diferenciar respecte de Sudan en política internacional, no podem estar gaire convençuts que la diferència observada en la nostra anàlisi sigui generalitzable de manera prou significativa.

Per a ajudar-nos a decidir si el resultat obtingut és prou significatiu o no, l'estadística té diverses tècniques, que s'anomenen proves de significació, que ens ajuden a decidir si donem per bona la diferència observada en les nostres dades. En altres paraules, un **test de significació** ens diu quina probabilitat hi ha que la hipòtesi nul·la sigui certa i que conseqüentment no hi hagi realment cap diferència en els nostres resultats si els extrapoléssim a una població de casos més àmplia.

Per a saber-ne més

Per a una revisió a fons de la significació i altres temes estadístics en ciències socials explicats en aquest mòdul podeu consultar Agresti i Finlay (2009), Babbie (2013), Halperin i Heath (2016) i Johnson i altres (2007).

El test de significació més habitual que s'utilitza per a taules de contingència és la **prova del Khi-quadrat de Pearson**. Aquest test el fem amb la funció `chisq.test()` i ens haurem de fixar en el p-valor (*p-value*) de la darrera línia que retorna la consola. Aquest valor ens diu la probabilitat que la hipòtesi nul·la sigui certa.

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

Segons el test de significació que acabem d'aplicar, el 0,95 que hem obtingut voldrà dir que hi ha un 95% per cent de possibilitats que la hipòtesi nul·la sigui certa. En ciències socials, hi ha la convenció que els testos de significació necessiten un p-valor inferior a 0,05 per a rebutjar la hipòtesi nul·la. Un p-valor més baix que 0,05 voldrà dir que les probabilitats que realment no hi hagi cap diferència són molt baixes, inferiors a un 5%, proporció que els científics consideren suficient per a rebutjar la hipòtesi nul·la. En el nostre exemple, la hipòtesi nul·la té més d'un 95% de probabilitats de ser certa. És un nombre massa alt, ja que necessitem menys d'un 5% per descartar-la i poder acceptar la nostra hipòtesi alternativa. Direm que no podem assegurar amb un 95% de confiança que la nostra hipòtesi sigui certa. Per tant, en aquest cas, haurem d'acceptar la hipòtesi nul·la.

Test de Fisher com a alternativa

Un altra test de significació per a taules de contingències és el test de Fisher. Podem demanar-lo amb la funció `fisher.test()` i en el retorn a la consola també podrem observar un p-valor.

Voten diferent els Estats Units si en el títol de la resolució hi apareix la paraula USSR?

Ens hem preguntat si el comportament de vot dels Estats Units s'escapa de l'habitual quan hi apareix la paraula *USSR* al títol de la resolució. Per això hem utilitzat el paquet *stringr* per a crear una nova columna, que sigui veritat quan apareix USSR i fals quan no hi apareix. En acabar hem demanat la taula de contingència i el test de significació.

```
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)
```

Podem observar que la diferència en el vot d'Estats Units quan a la resolució apareix USSR i quan no hi apareix és significativa amb un 95% de confiança. En altres paraules, la probabilitat que realment no hi hagi diferència en una població més elevada de casos és inferior al 5%.

2. Relació entre numèrica i categòrica

Quan vulguem mirar l'associació entre una variable independent categòrica i una dependent numèrica utilitzarem la **diferència de mitjanes**. Per a il·lustrar aquesta tècnica, en aquesta secció partim de la hipòtesi que les democràcies i les no democràcies voten diferent a l'ONU. El que no tenim clar és quins règims polítics voten més favorablement les resolucions a l'AGNU, però per ara només comprovarem la hipòtesi que hi ha diferència entre la mitjana de vot favorable en les democràcies en comparació amb les no democràcies.

Tenim, doncs, una variable independent categòrica, si un país és una democràcia o no, i una dependent numèrica, la mitjana de vot afirmatiu de cada país per any. Per a comprovar la nostra hipòtesi, utilitzarem i ajuntarem les bases de dades *unvotes* i *psData*. La segona base de dades, acrònim de Political Science Data, conté la funció `DDGet()` amb la qual podem extreure la classificació dicotòmica de democràcia de Cheibub i altres (2010). Per això instal·larem i carregarem la llibreria *psData* i importarem les variables *democracy* i *un_continent_name* amb el codi següent.

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

L'objectiu és construir un sol marc de dades que separi els països per any, que digui si cada any el país era una democràcia o no ho era i quina és la proporció anual de vots favorable de cada país. Primer hem netejat els marcs de dades que necessitem de *unvotes*, transformant la variable *year* i convertint en binària la variable *vote*, de manera que adoptarà el valor 1 quan el vot sigui favorable i el valor 0 en qualsevol altre cas. Un cop convertida en binària, hem demanat la mitjana agrupada per any i país per a obtenir la proporció de vot favorable sobre els vots totals.

```
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))
```

Fixem-nos en el marc de dades `un_year_mean` que hem creat. Ara ja tenim la taula neta amb les dades de *unvotes* que necessitem. El segon pas és unir-hi el marc de dades `DDdata` per mitjà de les columnes que tenen en comú: per codi de país i per any. A continuació seleccionem les columnes que ens interessen i transformem la variable binària *democracy* en les categories "Democracy" i "No Democracy".

```
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"))
```

Evitar unir les dades per nom de país

Unir marcs de dades pel nom del país és perillós, perquè no podem estar segurs que els noms estiguin tipificats igual. Per exemple, hi ha marcs de dades que anomenen els Estats Units com United States i n'hi ha que els anomenen United States of America. Això ens crearia dues categories diferents quan, realment, ens referim a la mateixa. Sempre serà millor, doncs, unir marcs de dades per mitjà d'un codi estandarditzat de país o variables estables com l'any.

El marc de dades `un_dd_year` està format per les variables necessàries per a visualitzar i quantificar una diferència de mitjanes. Tenim una variable amb l'any, una altra amb la mitjana de cada país per any i una altra que indica si el país era o no una democràcia.

2.1. Visualitzar

Hi ha diverses maneres de visualitzar una diferència de mitjanes. Podem, per exemple, utilitzar un diagrama de barres en el qual l'alçada de les barres representi la mitjana de cada categoria. Dues eines també comunes per a representar visualment la diferència de mitjanes són el diagrama de caixes i el diagrama de dispersió. En aquest apartat utilitzarem el diagrama de dispersió perquè creiem que és una eina amb més possibilitats visuals i que ens permet representar més variables al gràfic. Si el codi següent no us funciona, proveu de carregar el paquet *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()
```

Amb el codi anterior hem creat el diagrama de dispersió amb *jittering* que observem en la figura 2 i que conté la variable categòrica *democracy* a l'eix de les *x* i la numèrica *mean* a l'eix de les *y*. Hem volgut saber si el primer any que van començar les votacions a l'AGNU, el 1946, hi havia diferència entre l'orientació del vot entre les democràcies i no democràcies. Per això hem demanat un sumari estadístic de la mitjana que pren la variable dependent per a cada categoria de la variable independent, representat amb un punt vermell. També hem demanat un estadístic anomenat l'**error típic de la mitjana**, representat amb línies vermelles, que ens dona l'interval on podem assegurar que es troba realment la mitjana amb un 95% de confiança.

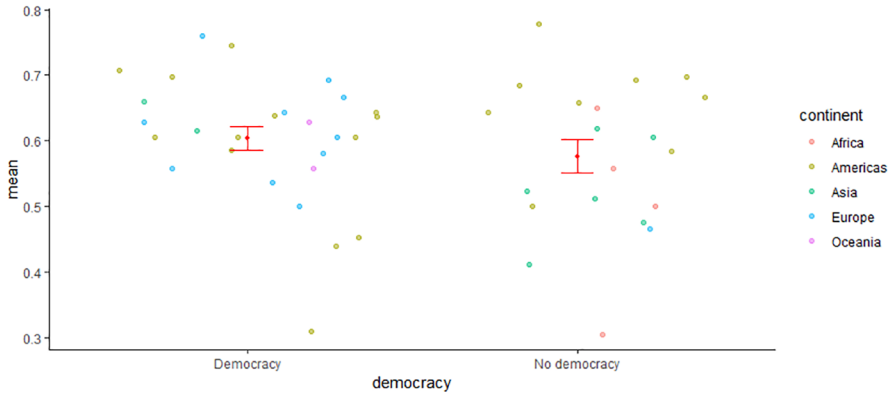
Error típic de la mitjana

L'error típic de la mitjana ens ajuda a estimar en quins intervals es pot trobar realment la mitjana en la població segons un nivell de confiança determinat. Com menys casos tinguem i com més dispersos estiguin aquests casos al voltant de la mitjana, més gran serà l'error típic i per tant menys precisió tindrem per saber on es troba la mitjana. Per a obtenir l'interval on es troba la mitjana amb un 95% de confiança haurem de sumar dos errors típics a la mitjana i restar-li dos errors típics.

Evitar els diagrames de barres

Si substituïm l'alçada barra per un punt tindrem la mateixa informació i haurem gastat menys tinta. És per això que els diagrames de barres no són una bona eina per a visualitzar la diferència de mitjanes. Utilitzarem sempre diagrames de caixes, que ens permeten veure més descriptius al gràfic, o diagrames de dispersió, que ens permeten observar la distribució de la variable dependent i afegir més variables al gràfic si ho creiem oportú. En alguns GitHub a internet trobareu comparacions entre tipus de representacions visuals.

Figura 2. Diferència de mitjana de vot per règim polític el 1946



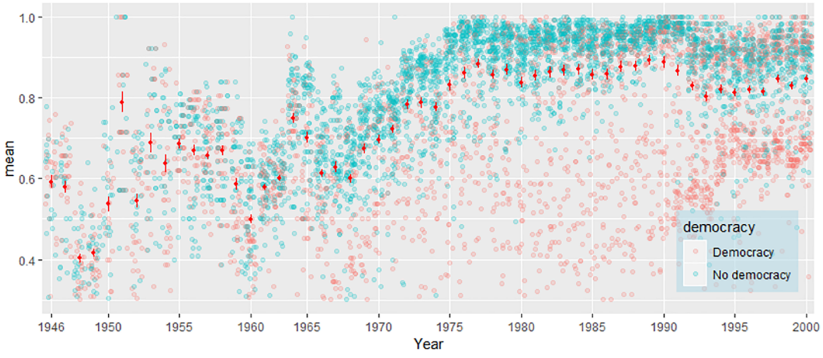
Fixem-nos que les democràcies votaven de mitjana lleugerament més a favor que les no democràcies en les resolucions de l'AGNU el 1946. Però fixem-nos també que els intervals de confiança sembla que s'encavalquen. Per tant, encara que apreciem una lleugera diferència entre totes dues categories, aquesta diferència no és estadísticament prou significativa. Això vol dir que, si hi hagués hagut moltes més votacions el 1946, és massa probable, segons els nostres estàndards científics, que la diferència que acabem d'observar realment no existís en una població més àmplia de casos.

Hem generat la figura 2 filtrant el codi per l'any 1946. Proveu de modificar el codi filtrant per altres anys o per tota la mostra d'anys i mireu si hi ha diferència de mitjana entre el vot de democràcies i no democràcies a l'AGNU. Intenteu contestar si els resultats donen suport o no a la nostra hipòtesi inicial.

Canvis de tendència en el temps

La diferència de mitjanes s'utilitza especialment per a veure canvis de tendència al llarg del temps. Per exemple, podem veure quins anys podem dir que hi ha hagut canvi significatiu de tendència en el vot a l'AGNU i quins anys no podem assegurar que el canvi observat sigui generalitzable amb un nivell de confiança del 95%. En la figura 3 podem observar en quins anys hi ha hagut un canvi prou significatiu de tendència en el vot i en quins anys no.

Figura 3. Anàlisi longitudinal amb diferència de mitjanes



Fixem-nos en els punts blaus i vermells. Tot sembla indicar que la majoria d'anys les no democràcies voten més a favor de les resolucions de l'AGNU que no pas les democràcies. Podeu consultar en l'annex del mòdul el codi d'aquesta figura.

2.2. Quantificar

Un dels avantatges de visualitzar la diferència de mitjanes és que podem observar les línies vermelles que representem els intervals de confiança i, per tant, podem obtenir una intuïció visual prou clara dels nivells de significació sense necessitat de quantificar. No obstant això, és pertinent ser precisos en les nostres anàlisis i tenir també l'estimació de la diferència de mitjanes quantificada numèricament. Aquesta quantificació ens ha de permetre respondre a dues preguntes:

1) La primera és si hi ha diferència estadística en la mitjana de la variable dependent segons les categories de la variable independent. Per a trobar si la resposta és afirmativa o negativa haurem de mirar si podem rebutjar o no la hipòtesi nul·la.

2) La segona pregunta és saber quanta diferència hi ha entre una mitjana i l'altra des d'un punt de vista estadístic, per la qual cosa també ens haurem de moure amb intervals de confiança per a determinar on es troba la mitjana en la població real.

Per a respondre a les dues preguntes utilitzarem la funció `t.test()` on introduirem els vectors de les dues variables que estem analitzant a la figura 2. La primera pregunta es respon amb el p-valor. Un p-valor inferior a 0,05 significarà que podem rebutjar la hipòtesi nul·la i afirmar amb un 95% de confiança que la diferència de mitjanes observada existeix en la realitat. La segona resposta es respon amb l'interval de confiança al 95%, que s'obté a partir de sumar i restar dos errors típics a la mitjana.

```
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
```

Com podem veure en el resultat general per a aquest codi, tenim un p-valor molt superior al 0,05. És a dir, tal com ja ens havia confirmat visualment el gràfic de la figura 2, no podem assegurar que la diferència entre mitjanes no

sigui zero. Per tant, en resposta a la primera pregunta, no podem assegurar amb un 95% de confiança que hi hagi diferència en orientació de vot entre democràcies i no democràcies el 1946.

En la segona pregunta volíem saber quina era la diferència real estimada entre les mitjanes. La resposta la trobem en els nombres -0,13 i 0,16. Aquests nombres indiquen l'interval en què estimem que es mourà la diferència de mitjanes en la població real amb un 95% de confiança. És a dir, amb la mostra de resultats que tenim, podem apostar amb un 95% d'encert que la diferència de mitjanes en la població real es trobaria en l'interval següent: a un extrem, podria ser que les democràcies hagin votat afirmativament un 16% més que les no democràcies, però a l'altre extrem podria ser que les no democràcies hagin votat afirmativament un 13% més que les democràcies. Com que no ho podem saber del cert, l'únic que podem dir és que si apostem que la mitjana real està entre aquests dos intervals tenim un 95% de probabilitats d'encertar.

Fixeu-vos que entra a dins dels nostres pronòstics la possibilitat que la diferència de mitjanes sigui zero. En aquesta situació, no podem rebutjar la hipòtesi nul·la, ja que és massa probable que sigui certa. Per a contrastar-ho amb un cas en què sí que hi ha diferència, proveu de demanar el registre de vot entre els anys 1995 i 2000.

Codi del registre de vot

Si no us en sortiu a l'hora de demanar aquest registre de vot, trobareu el codi de visualització en l'annex.

El resultat de demanar el *t-test* en aquest cas és el següent:

```
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 aquest cas, el p-valor és inferior a 0,05 i per tant sí que podem rebutjar la hipòtesi nul·la. Sembla, doncs, que entre els anys 1995 i 2000 hi ha diferència prou significativa entre el tipus de règim polític i l'orientació del vot. Concretament, podem estimar amb un 95% de confiança que la diferència real en la població se situa entre 0,23 i 0,29. En altres paraules, el vot de les no democràcies a l'AGNU és afirmatiu, de mitjana, entre un 23% i un 29% en relació amb les democràcies. En aquest cas, el 0% no figura entre els nostres pronòstics. Com que la diferència de mitjanes no pot ser zero, podem rebutjar la hipòtesi nul·la.

3. Relació entre variables numèriques

Quan volem explicar la relació entre dues variables numèriques, necessitem diversos instruments, tant visuals com quantitatius. Visualment, la manera de representar gràficament aquestes dues variables és mitjançant un diagrama de dispersió, un gràfic de dues dimensions en què les coordenades de cada punt representen el valor que pren cada observació segons les variables independent i dependent. Situarem sempre la variable independent x a l'eix horitzontal i la variable dependent y a l'eix vertical. El diagrama de dispersió s'acostuma a complementar amb la **recta de regressió**, una tècnica tant visual com quantitativa que ens permet dibuixar la línia que millor s'ajusta a la relació que observem en el gràfic i quantificar-la. Per a quantificar la relació també utilitzarem el **coeficient de correlació**, que descriu de manera quantitativa la força i la direcció de la relació, i el coeficient de determinació, que ja ens és familiar.

En conjunt, amb les tècniques esmentades podem arribar a saber cinc aspectes principals de l'associació entre dues variables numèriques. Aquests aspectes els resumim aquí i els ampliem en els apartats següents:

1) Forma: La forma es refereix al dibuix que fa la relació entre x i y. En direm lineal si aquesta forma dibuixa una línia més o menys recta. Si, en canvi, aquest dibuix pren una forma corba, en direm no lineal o quadràtica.

2) Força: La força ens indica com de perfecta és la forma que fan els punts al gràfic. Si, per exemple, les coordenades de x i de y descriuen una línia recta perfecta, direm que la força de la relació és molt alta. A mesura que la línia imaginària que formen els punts es desdibuixa, tindrem uns punts més dispersos i una relació amb poca força.

3) Direcció: Direm que la direcció és positiva quan les dues variables es mouen en la mateixa direcció. És a dir, quan una incrementa els seus valors l'altra també ho fa. Direm que la relació és negativa quan les dues variables es mouen en direcció oposada. És a dir, quan els valors d'una variable incrementen, els valors de l'altra disminueixen. Si no apreciem una direcció positiva o negativa, direm que és una relació plana.

4) Significació: La relació entre x i y és estadísticament significativa si les probabilitats que no hi hagi relació, és a dir que la línia sigui plana, siguin inferiors al 5%.

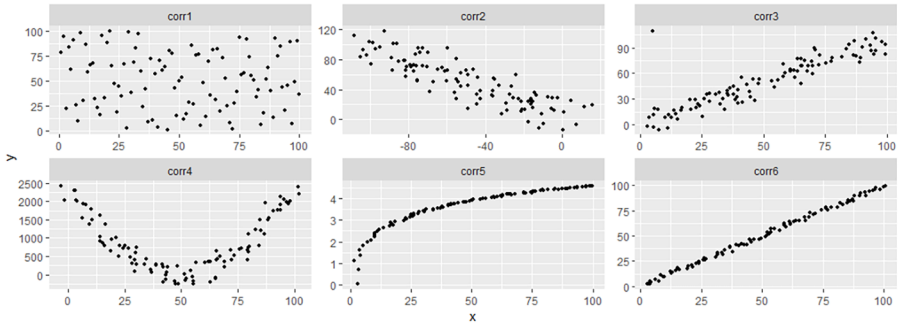
5) Casos extrems: Direm que la relació té casos extrems si hi ha observacions que estan molt allunyades de la forma que pren la relació.

3.1. Visualitzar

La manera de visualitzar una relació entre variables numèriques és per mitjà d'un diagrama de dispersió, que podem cridar amb R per mitjà del paquet *ggplot2* i la geometria `geom_point()`. A la figura 4 veiem sis exemples de relacions entre variables amb diferent forma, força, direcció i presència de casos extrems:

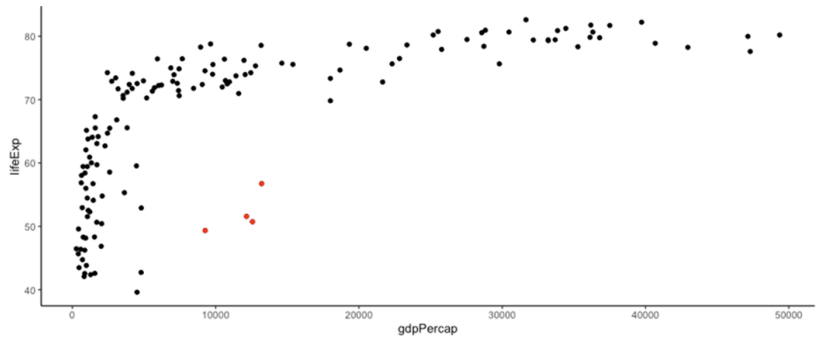
- en el gràfic *corr1* observem una relació sense forma, sense força i de direcció plana;
- en el cas de *corr2* té una relació lineal, de força dèbil i direcció negativa;
- *corr3* té una forma lineal, de força mitjana, direcció positiva i que té la presència d'un cas extrem;
- en el cas de *corr4* és una relació no lineal, de força mitja;
- *corr5* té una relació no lineal, molt forta i positiva;
- *corr6* té una relació lineal, molt forta i positiva.

Figura 4. Visualització de la correlació entre variables



Cas real de relació no lineal amb casos extrems

Un dels fenòmens més investigats en els estudis de desenvolupament és la relació existent entre el PIB per càpita d'un país i la seva esperança de vida. Com observem en les dades de 2007, la relació és no lineal, forta i positiva. També podríem considerar la presència d'alguns casos extrems que es troben amb més de 10.000 dòlars per càpita i esperança de vida al voltant de 50 anys, que hem marcat en vermell. Sabríeu trobar quins són aquests quatre casos? Trobareu el codi als annexos.



La línia de regressió és una tècnica visualment molt utilitzada per a representar la forma i la direcció de la relació entre dues variables numèriques. Hi ha diferents mètodes per a representar aquesta relació, entre elles el mètode **Ordinary Least Squares (OLS)** i el mètode **Locally Estimated Scatterplot Smoothing**

(LOESS). La principal diferència entre totes dues és que mentre la línia OLS és una línia completament recta, la LOESS s'intenta ajustar a la distribució dels punts. Nosaltres utilitzarem preferentment la línia OLS.

La característica principal de la línia OLS és que intenta minimitzar amb una línia recta la suma de la distància al quadrat entre la línia i els punts. En altres paraules, és la recta que de mitjana passa més a prop de tots els punts del gràfic. Qualsevol altra línia que vulguem traçar no aconseguirà minimitzar aquesta distància. La línia OLS es representa amb la funció `geom_smooth()` juntament amb la geometria del diagrama de dispersió. A dins de la funció haurem d'especificar el mètode lm (*linear model*) i preferiblement posarem l'error típic `se` (*standard error*) a FALSE, com mostrem en aquest codi creat a partir del marc de dades *gapminder*.

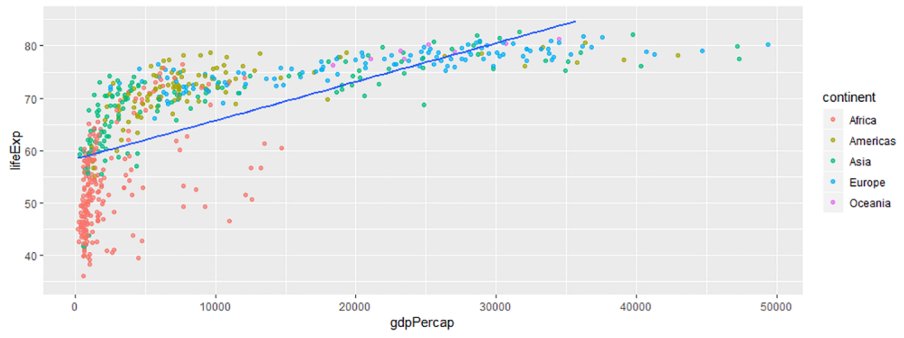
Dibuixar una línia manualment

Podem provar de dibuixar manualment una línia en el gràfic amb la geometria `geom_abline()`. En el primer argument dins de la geometria indicarem l'*intercept* i en el segon indicarem el pendent.

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

En aquest codi, representat a la figura 5, hem inclòs l'estètic de color al diagrama de dispersió però a la recta de regressió OLS, perquè no volem que generi una línia per a cada continent. Podem veure que la relació entre PIB per càpita i esperança de vida dibuixa una forma més aviat corba, positiva i més aviat forta.

Figura 5. Diagrama de dispersió amb recta OLS



Línies horitzontals i verticals customitzades

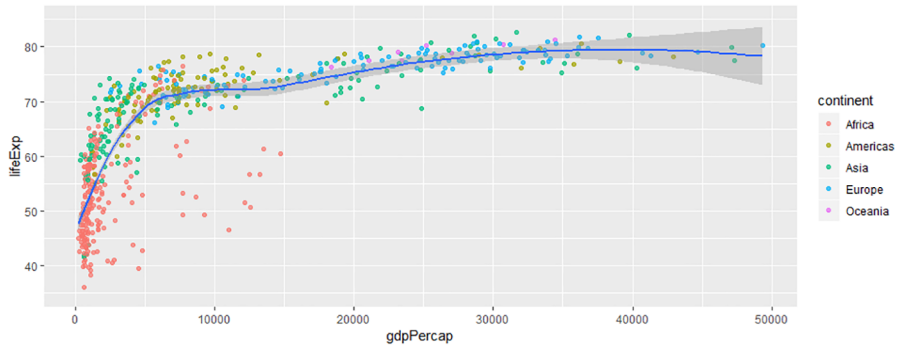
Les geometries `geom_vline()`, `geom_hline()` i `geom_abline()` permeten col·locar una línia al gràfic amb la posició que indiquem. Les dues primeres són línies paral·leles en l'eix vertical i l'horitzontal respectivament, mentre que la darrera permet establir un pendent amb l'argument *slope*.

La línia LOESS és una eina utilitzada també en anàlisi de regressió per a veure la relació entre dues variables numèriques. Aquesta tècnica és especialment útil en línies temporals, ja que ens permet fer previsions de tendència. Per a veure la LOESS al gràfic anterior, només cal que eliminem l'argument `method = "lm"`, ja que `geom_smooth()` ja ens dibuixa una LOESS per defecte. També podeu indicar-ho explícitament amb `method = "loess"`. En la figura 6, visualitzem una LOESS amb els atributs que té per defecte, entre ells l'interval de confiança al 95% i la sensibilitat de la línia:

Sensibilitat de la línia i error típic

Podem canviar la sensibilitat de la línia amb l'argument `span`, on podem indicar un nombre entre 0 i 1. Proper a 0, la sensibilitat serà més alta i la línia tindrà canvis més abruptes. Proper a 1, la línia serà més harmònica. L'argument `se` ens marcarà l'interval de confiança de la línia al 95% de confiança.

Figura 6. Diagrama de dispersió amb línia LOESS



Podem comprovar que la LOESS estableix una línia suau que es va adaptant segons la situació dels punts al gràfic. Les zones grises marquen l'interval de confiança de la línia. Podeu veure que l'interval serà més estret en les zones dels gràfic on hi hagi més punts i estiguin més junts, mentre que l'interval serà més gran en les zones on menys punts hi hagi i més separats estiguin.

3.2. Quantificar

Amb un diagrama de dispersió ja podem intuir visualment alguns paràmetres com la força i la direcció d'una relació entre variables numèriques, però la quantificació ens serà més útil per a fer-nos preguntes més precises com per exemple: «com de forta?» Per a poder quantificar aquests paràmetres tenim principalment dues mesures numèriques:

- el coeficient de correlació,
- el coeficient de determinació.

El **coeficient de correlació**, que sovint veurem il·lustrat amb el símbol *r*, és la manera que utilitzarem per a quantificar la força i la direcció d'una relació lineal entre dues variables numèriques. Quantificar relacions per mitjà del coeficient de correlació ens permet comparar numèricament les relacions entre elles . Aquest coeficient és un nombre que varia entre -1 a 1. El signe ens indica si la relació és positiva o negativa. La magnitud del número correspon amb la força de la relació. Per tant, un número proper a 1 indicarà una relació molt forta i positiva. Un número proper a -1 indicarà una relació molt forta i negativa, mentre que un número proper a 0 indicarà que no hi ha relació entre totes dues variables.

Karl Pearson, l'inventor

També sentirem a parlar d'*r* com el coeficient de correlació de Pearson (o originalment *Pearson product-moment correlation*). Karl Pearson és un dels matemàtics que han contribuït més a l'evolució de l'estadística. Amb el seu coeficient de correlació, Pearson pretenia trobar una manera de calcular la desviació respecte la mitjana a *x* i *y*. Podeu trobar a internet molts vídeos que expliquen la lògica matemàtica que hi ha al darrere d'*r*. Memoritzar-ne la fórmula no és important, perquè ja ens ho fa R, però sí que tenir clar d'on surt *r* ens ajudarà a reforçar la intuïció que hi ha al darrere.

Tornem-nos a fixar en la figura 4 i intentem deduir el coeficient de correlació de cada una de les relacions. Després podem comprovar si ens hi hem acostat mitjançant la funció `cor()`, que calcula el coeficient de correlació a partir de dos arguments: una variable en el primer argument i l'altra variable en el segon argument. No importa l'ordre en què especifiquem les variables, ja que les relacions són simètriques i ens donaran el mateix coeficient. A continuació hem imprès els coeficients de correlació de cada un dels sis gràfics. Com veiem,

la relació més forta i també positiva la trobem a *cor6*. També són relacions fortes i positives *cor3* i *cor5*. En el cas de *cor1* i *cor4* no hi ha relació entre les variables mentre que a *cor2* la relació és forta i 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
```

És important tenir en compte que el coeficient de correlació no pot capturar amb precisió la força d'una relació no lineal. És per això que en el cas de *cor5*, on els punts descriuen una línia corba quasi perfecta, no trobem un coeficient de correlació tant elevat com a *cor3* o *cor6*. Per a incrementar el coeficient en el cas d'aquestes relacions quadràtiques, podem buscar, per exemple, el logaritme neperià d'una de les variables.

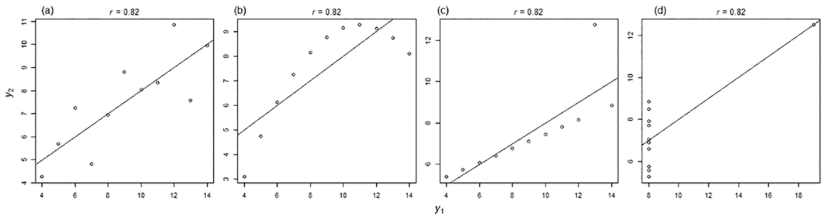
D'una manera més general, podem aplicar el coeficient de correlació per a veure com estan relacionats en tres continents diferents el PIB per càpita i l'esperança de vida en diversos anys. A continuació, hem demanat un sumari amb el coeficient de correlació per sis anys diferents a Europa, Àfrica i Àsia.

```
> 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 codi també hem demanat un sumari del nombre d'observacions. Així podem saber quants casos tenim a cada correlació i inferir fins a quin punt poden ser significatius aquests resultats. Com més casos tinguem, més significatives acostumen a ser les observacions. En totes les combinacions de la taula la relació entre variables és positiva encara que sempre ha estat més forta a Europa que en els altres continents. A Àfrica aquesta correlació era especialment baixa a mitjan segle passat, però ha augmentat lleugerament al llarg del temps.

Francis Anscombe i el problema de quantificar relacions

Quantificar relacions lineals també pot tenir inconvenients, tal com va detectar el professor d'estadística Francis Anscombe amb una base de dades sintètica construïda el 1973. Hi va il·lustrar com relacions completament diferents podien tenir el mateix nombre de punts, la mateixa mitjana i desviació típica per x i y i la mateixa correlació. Fixem-nos en els quatre diagrames de dispersió següents, obtinguts per mitjà del marc de dades `anscombe`, on tots quatre tenen una r de 0,82.

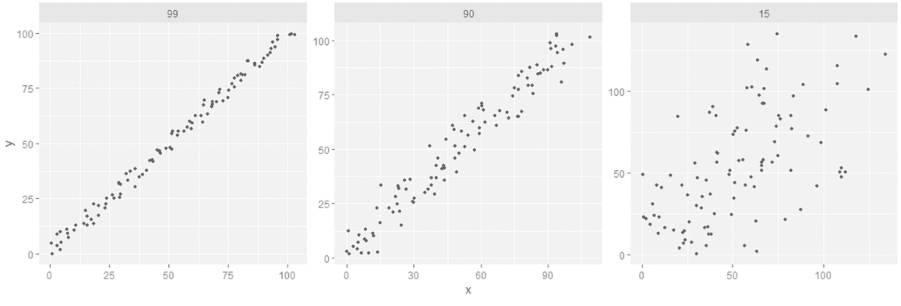


La conclusió més important que podem treure d'aquests quatre gràfics és que encara que *r* ens pugui ajudar a quantificar numèricament una relació entre dues variables, sempre serà pertinent visualitzar la relació per a evitar que els nombres ens facin una mala passada. Per exemple, en el darrer cas veiem que no hi ha cap mena de relació entre *x* i *y*, però un sol cas extrem ens distorsiona el coeficient de correlació. Podeu trobar aquest marc de dades en els paquets de base d'R.

La segona manera de poder quantificar la relació entre dues variables és per mitjà del **coeficient de determinació**, representat amb R^2, que és un número entre 0 i 1 que ens determina la proporció de la variació en la variable dependent que es pot explicar amb la variació dels valors de la independent. La sort que tenim és que R^2 és molt fàcil de calcular. Només cal elevar al quadrat el coeficient de correlació. El que és més complicat, però, és comprendre el significat d'aquest nombre.

Per a entendre millor com interpretar R^2, podem considerar que el seu valor respon a una pregunta: quin percentatge de la variabilitat de *y* podem explicar si sabem la variabilitat de *x*? Fixem-nos en la figura 7. A l'esquerra veiem una correlació quasi perfecta. Sabent com varien els valors de *x* podríem endevinar molt acuradament els valors de *y*. Per exemple, si ens diuen que el valor de *x* és 64, sabem que el valor de *y* també estarà al voltant de 64 i no estarem gens lluny d'equivocar-nos. Si ens pregunten com de segurs estem d'encertar, entre 0 i 1, direm que estem un 99% convençuts que no ens equivocarem. L'R^2 és 0,99 perquè, sabent la variabilitat de *x*, podem explicar un 99% de la variabilitat de *y*. Al gràfic de la dreta, en canvi, ens serà molt més difícil d'encertar els valors de *y* si coneixem els valors de *x*. Si ens diuen que la *x* és 90, podrem estimar amb molta menys precisió els valors de *y*. Diríem que la nostra R^2 estarà al voltant de 0,15.

Figura 7. Coeficient de determinació de tres correlacions



Compte a l'hora d'interpretar R^2

Hem d'anar en compte a l'hora d'interpretar R^2. Un coeficient de determinació baix no vol dir necessàriament que tinguem un mal model i menys en ciències socials. Sovint podem tenir una R^2 baixa però aquest petit percentatge d'explicació pot ser una aportació estadísticament significativa per un problema complex.

Quan ens referim a R^2, parlarem sempre en percentatges i ens expressarem d'una manera tècnica. Per exemple, per a descriure el gràfic del mig de la figura anterior, direm que «un 90% de la variabilitat de la variable dependent pot ser explicada per la independent».

Hi ha dues maneres de trobar el coeficient de determinació. La primera és elevar el coeficient de correlació al quadrat, com mostrem en el codi següent. Podeu trobar els codis de la figura 7 en l'annex.

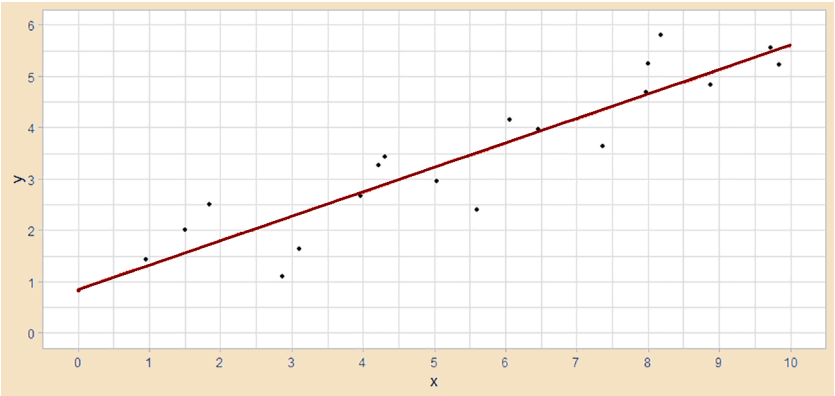
```
cor(x, y)^2
```

La segona manera que tenim per a trobar R^2 és a partir del model de regressió, destinat a construir models lineals, que expliquem en l'apartat següent.

3.3. Modelar

Després d'aprendre a visualitzar i quantificar una correlació, en aquest apartat veurem com crear un model de regressió a partir de les dades disponibles. Modelar significa crear una fórmula matemàtica que ens permeti predir el valor d'una variable si sabem el valor de l'altra. Això ho podem fer amb la recta de regressió OLS, que, com ja hem explicat, té una doble virtut visual i quantitativa. Visualment, el mètode OLS ens mostra en un gràfic quina és la manera més ajustada de descriure la relació entre dues variables numèriques mitjançant una línia recta. De manera quantitativa, la recta OLS ens permet predir quin valor prendria *y* quan la recta passa pels valors de *x*. Fixem-nos en la figura 8. La recta ens ajuda a fer una predicció del comportament de les variables *x* i *y* del marc de dades model, que trobareu disponible amb codi en l'annex. La recta ens diu, per exemple, que si *x* té un valor de 2,5, *y* hauria de tenir un valor proper a 2.

Figura 8. Model lineal OLS



En lloc de fer aquestes estimacions de manera visual, podem quantificar la recta OLS i obtenir-ne el model matemàtic a partir de dos paràmetres: la inclinació de la recta de regressió i el valor que pren *y* quan *x* és zero.

Interpretar correctament la inclinació

Per a interpretar correctament el valor de la inclinació caldrà tenir molt clar quines són les unitats de *x* i de *y*. Si per exemple la variable *x* són anys i *y* són dòlars per càpita, ho llegirem com «cada any (unitat de *x*) hi ha un canvi de dòlars per càpita (unitat de *y*) del nombre que marqui la inclinació».

1) Per **inclinació** ens referim a quant creix o decreix *y* si augmentem *x* en una unitat. Si ens fixem en el gràfic anterior, quan *x* passa de 0 a 1, *y* canvia aproximadament en mig punt, ja que el valor de *y* passa d'una mica menys d'1 a una mica menys d'1,5. Llavors, la inclinació, que és la mateixa per a tota la recta de regressió, serà aproximadament de 0,5.

2) Del valor que pren *y* quan *x* és 0 en direm **constant** o ***intercept***. A primer cop d'ull podem veure com la constant d'aquesta regressió és una mica inferior a 1, ja que és el punt per on la recta talla l'eix de les *y* quan la *x* és igual a 0.

El nou tipus d'objecte *lm*

La funció `lm()` ens retorna un tipus d'objecte que no hem vist fins ara de classe *lm*, com podem observar amb la funció `class()`. Si demanem el `typeof()`, veurem que és una llista, un altre tipus d'objecte com seria un vector o un marc de dades. Per a transformar aquest objecte en vector posarem tota l'operació dins de la funció `coef()`.

Per a determinar amb precisió aquests dos paràmetres haurem d'utilitzar la funció `lm()`, on indicarem com a primer argument el nom de la variable dependent seguit d'una titlla i el nom de la dependent. El segon argument és el marc de dades. Si obtenim el model de regressió del marc de dades `model`, veiem que no ens equivocàvem gaire. La inclinació és de 0,475 i l'*intercept* és de 0,854.

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

Amb aquest resultat ja tenim les dues dades que necessitem per a interpretar el model. Si ens fixem en l'*intercept*, direm que quan la *x* és 0, *y* és 0,854. Si ens fixem en la inclinació, direm que la *y* creix en 0,475 unitats per cada increment d'una unitat d'*x*. Aquests dos valors els podem utilitzar per a crear un model predictiu que ens permetrà conèixer una estimació dels valors de *y* si sabem el valor de *x*. En el codi següent, podem predir els valors de *y* amb una fórmula senzilla, en què multipliquem el pendent pel valor de *x* i hi sumem la constant.

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

Podem provar el nostre model predictiu substituint la *x* per diferents nombres. Per exemple, si la substituïm per 5, la *y* serà 3,229 com comprovarem si entrem el codi `0.854 + 5 * 0.475` a la consola. Si la *x* és 10, la *y* serà 5,604.

A part de l'*intercept* i el pendent, hi ha molta més informació que podem obtenir a partir del model lineal. En el codi següent hem creat primer l'objecte `lm_model` i després hem generat la informació addicional a partir de la funció `summary()`. R no solament ens genera l'*intercept* i el pendent, sinó que també

ens proporciona informació del model de regressió com les dades utilitzades, l'especificació del model, els residus, l'error estàndard associat, la significació estadística o l'R^2 múltiple i l'R^2 ajustat.

```
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 les línies següents només descriurem alguns dels elements d'aquest sumari. A l'inici del sumari ens apareix la formula que hem utilitzat, mentre que a la columna *Estimate* de la taula de coeficients podem veure l'*intercept* i el pendent. El sumari també ens mostra l'R^2 (*R-squared*), amb la diferència que aquí veiem l'R^2 múltiple i l'R^2 ajustat. Si busquem l'R^2 de la manera que hem explicat anteriorment per mitjà de `cor(model$x, model$y)^2`, veurem que el seu resultat coincideix amb l'R^2 múltiple, que és la versió que utilitzarem en anàlisi bivariant per a interpretar l'R^2.

L'element que ens faltava veure, i que ja hem vist en les altres tècniques d'anàlisi bivariant, és el **test de significació** estadística, una convenció matemàtica que ens permet descartar que els efectes observats en una relació lineal siguin fruit de l'atzar. És a dir, quan observem una relació entre dues variables, sigui positiva o negativa, ens hem de preguntar si és possible que la relació observada no sigui real quan extrapolem la nostra mostra a la realitat. Si no és real, vol dir que hi ha una probabilitat superior al 5% que el pendent de la recta de regressió sigui zero i que per tant la hipòtesi nul·la sigui certa. Sabem que, en ciències socials, normalment s'estableix un valor inferior a 0,05 per a rebutjar la hipòtesi nul·la i així descartar l'efecte de l'atzar en els nostres resultats. La significació estadística la trobem indicada a la columna amb nom *Pr(>|t|)* de la taula de coeficients. Vegem la taula novament:

```
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
```

Exemple de significació estadística

Suposem que hem viatjat dues vegades de la nostra vida a Màlaga i les dues vegades ha plogut. Si, d'acord amb la nostra experiència, afirmem que 'a Màlaga sempre hi plou', quina validesa tindria aquesta afirmació? És molt possible que si incrementem el nombre d'experiències a Màlaga la nostra afirmació sigui més difícil de sustentar, ja que trobarem algun dia que no plogui. Per tant, és molt possible que la nostra afirmació inicial hagi estat fruit de l'atzar.

En la taula de coeficients ens fixarem en el valor de *x* a la taula *Pr(>|t|)*. Veiem que el nombre que ens apareix és 0,000000068, la qual cosa compleix el requisit que el valor en qüestió sigui inferior a 0,05. Per tant podem descartar la hipòtesi nul·la i assegurar amb un 95% de confiança que els resultats observats no són fruit de l'atzar. Si els valors són superiors a 0,05, direm que la relació no és estadísticament significativa.

Interpretar els asteriscos

Per facilitar la lectura dels resultats, tenim la significació dels coeficients assenyalats amb asteriscos. Si, com el nostre cas, la significació està entre 0 i 0,001, s'indicarà amb ***. Si la significació està entre 0,001 i 0,01 s'indicarà amb ** mentre que si està entre 0,01 i 0,05 s'indicarà amb *. En aquests tres casos, direm que la relació entre *x* i *y* és estadísticament significativa.

4. Regressió múltiple

Com dèiem en la introducció d'aquest mòdul, podem argumentar que dues variables estan associades si compleixen diversos requisits. En les anteriors seccions hem après, principalment, a donar arguments estadístics per mitjà de la quantificació, per exemple, de la força i el nivell de significació d'una relació. Aquests arguments ens permeten dir si és possible que la relació observada succeeixi també en la realitat i quina és la magnitud d'aquesta relació. Un altre requisit important, però, és contraposar la nostra hipòtesi amb altres hipòtesis alternatives. I aquesta és la gran virtut de la **regressió múltiple**, que és una tècnica d'anàlisi multivariant que permet afegir noves variables independents al model de regressió. Aquesta tècnica ens ajuda a considerar l'efecte que poden tenir altres variables, com *z* o *w*, en la nostra afirmació que *x* té un efecte sobre *y*. Així, la regressió múltiple ens permet fer afirmacions més robustes sobre la relació entre *x* i *y*, perquè podem comprovar si *x* i *y* tenen relació fins i tot quan controlem els efectes de *x* per a altres variables independents com podrien ser *z* o *w*. Aquesta lògica l'expliquem tot seguit i acabarem la secció amb un exemple aplicat que relaciona el tipus de règim i el vot afirmatiu a l'ONU.

4.1. Quantificar

Si sabem com es construeix un model de regressió simple, la regressió múltiple és molt senzilla ja que només cal afegir les variables addicionals amb el signe + a la fórmula `lm()`. En el codi següent hem afegit la variable categòrica binària *z* a la relació i hem demanat directament un sumari. Hem exclòs de la impressió la part de residus.

```
> 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 d'aquest sumari, on trobem inclòs la nova variable *z*, és molt semblant a la lectura en l'anàlisi amb dues variables. En primer lloc, la constant o *intercept* mostra quin valor pren *y* quan el valor de tota la resta de variables

explicatives és 0. És a dir, quan *x* pren el valor 0 i *z* pren el valor 0, el valor de *y* serà de 0,94. En segon lloc, el coeficient de la variable numèrica *x* ens diu que, per cada unitat que augmentem de *x*, la *y* augmentarà 0,48 i deixarà constant la resta de variables independents. Aquesta relació és estadísticament significativa. En tercer lloc, el coeficient de la variable categòrica *z* ens explica l'efecte que té sobre *y* augmentar els valors de *z* de 0 a 1 deixant constant la resta de variables independents. Això vol dir que quan *z* passa de 0 a 1, y disminueix 0,20 en el nostre model. La relació és negativa i no és estadísticament significativa, per la qual cosa no podem assegurar que aquesta relació de *z* sobre *y* existeixi si la generalitzem en una població més àmplia.

En darrer lloc, en anàlisi multivariant ens fixarem en l'R^2 ajustat. Veiem com un 81% de la variabilitat de *y* pot ser explicada per les dues independents del model. Si comparem aquests resultats amb l'exercici anterior, veiem com la introducció de *z* en el nostre model no ens ha aportat més capacitat d'explicar *y*. No és estadísticament significativa ni tampoc no ens ha fet augmentar de manera substancial el coeficient de determinació, més aviat al contrari. Per això mateix, valdrà la pena descartar *z* del nostre model. Abans, però, veurem com quedaria la fórmula si decidíssim incloure-hi la variable.

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

És important entendre que la regressió múltiple ha de servir essencialment com un mètode que permet rebutjar amb més solidesa la hipòtesi nul·la. És a dir, aquest mètode ens permet dir que l'efecte observat de la variable independent principal sobre la variable dependent es continua produint amb un interval de confiança superior al 95% fins i tot quan controlem per l'efecte d'altres variables explicatives rivals.

4.2. Cas d'estudi

En els apartats anteriors hem pogut observar, en contra de les nostres expectatives, que justament són les no democràcies les que voten més afirmativament a l'ONU i no pas les democràcies. La figura 3 ens donava una pista sobre aquesta tendència. En aquest apartat comprovarem la hipòtesi que les no democràcies tendeixen a votar més favorablement les resolucions de l'AGNU en relació amb les no democràcies. Per això, hem dissenyat una anàlisi de regressió múltiple on contrastem la nostra hipòtesi amb altres hipòtesis rivals, com trobareu explicat en la taula 2. La hipòtesi que volem comprovar és que ser una democràcia (*x*) té un efecte negatiu sobre el vot afirmatiu a l'AGNU (*y*). Pensem, però, que potser això està determinat pel nivell de desenvolupament del país (*w*). Les democràcies acostumen a ser riques i les autocràcies acostumen a ser pobres, per la qual cosa podria ser que el fet de ser ric és el que faci votar de manera menys favorable a l'AGNU i no pas el fet de ser una democràcia. Així, en l'anàlisi multivariant controlarem els efectes de la nostra

hipòtesi principal pels efectes de la primera hipòtesi alternativa, el nivell de desenvolupament, mesurat pel PIB per càpita. Creiem que un PIB per càpita més gran afectarà negativament el vot afirmatiu a l'AGNU.

Com a segona hipòtesi alternativa hem considerat que el vot afirmatiu a l'AGNU (*y*) pot estar condicionat per la proporció de democràcies a l'assemblea (*z*). Així, els anys que hi hagi més democràcies a l'ONU el vot afirmatiu serà més desfavorable i els anys que n'hi hagi menys el vot afirmatiu serà més favorable. Per tant, creiem que la proporció anual més elevada de democràcies a l'AGNU tindrà una incidència negativa en el vot afirmatiu de les democràcies.

Taula 2. Anàlisi multivariant de vot a l'AGNU

Variable	Mesura i font	Codificació	Hipòtesi
Variable dependent (*y*)	Proporció de vot afirmatiu a l'AGNU (*unvotes*)	Proper a 1, més vot afirmatiu, proper a zero, menys vot afirmatiu (*vote*)	
Variable independent principal (*x*)	Classificació dicotòmica de democràcia (*psData*)	1 = Democràcia, 0 = No democràcia (*democracy*)	Més democràtic, menys vot afirmatiu
Variable independent alternativa (*w*)	Desenvolupament del país (*WDI*)	PIB per càpita PPP, preus constants 2011 (*gdpcap*)	Més PIB per càpita, menys vot afirmatiu
Variable independent alternativa (*z*)	Proporció de democràcies a l'AGNU (*unvotes*)	Proper a 1, més proporció de democràcies, proper a 0, menys proporció (*unga_prop*)	Més proporció de democràcies, menys vot afirmatiu

Per a poder fer aquesta anàlisi multivariant hem creat el marc de dades `un_wdi`, disponible en l'annex, amb les variables *vote*, *democracy*, *gdpcap* i *unga_prop*. A continuació, hem demanat el sumari del model lineal. Podem veure com *democracy* té una relació negativa i estadísticament significativa amb la variable dependent quan controlem els efectes de les altres variables del model. Això permet dir que hi ha relació amb un 95% de confiança i ens permet descartar la hipòtesi nul·la. També veiem que el *gdpcap* té una relació negativa i estadísticament significativa amb la variable dependent, controlant els efectes de les altres variables. La variable *unga_prop* també té una relació negativa, tal com esperàvem, però en aquest cas no té una relació estadísticament significativa amb *y*.

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

Residuals:
     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
```

El més important d'aquesta anàlisi és que la nostra hipòtesi continua sent certa amb un nivell suficient de confiança fins i tot quan la contrastem amb dues hipòtesis rivals. Per tant, l'afirmació que les no democràcies tendeixen a votar més favorablement les resolucions de l'AGNU que no pas les no democràcies és més robusta després d'aquest anàlisi multivariant. També podem dir, observant l'R^2, que el nostre model explica prop d'un 19% de la variabilitat de la variable dependent. En anàlisi multivariant mirarem l'R^2 ajustat. També és molt important saber llegir la columna *Estimate*, que ens diu l'efecte de cada variable independent sobre la dependent. En el codi següent traduïm aquesta columna a la fórmula matemàtica que ens permet tenir un model predictiu.

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

Segons el nostre model, en un hipotètic país amb valor zero a la variable democràcia (per tant, que sigui una no democràcia), amb el PIB per càpita de valor zero (per tant, absolutament pobre) i en una situació amb zero democràcies a l'AGNU, aquest hipotètic país votarà afirmatiu un 98% de les vegades. En canvi, si canviem el règim del país a democràtic (la variable *x* passa de 0 a 1) i deixem les altres variables constants, aquest hipotètic país votarà a favor un 88% de les vegades, ja que en el nostre model hi ha un canvi del 10% en el vot segons el tipus de règim. Quant al nivell de desenvolupament, a mesura que el PIB per càpita incrementa, la proporció de vot afirmatiu disminueix. Concretament, la proporció de vot afirmatiu disminueix un 2% per cada 10.000 dòlars addicionals per càpita que tingui el país. Finalment, com més democràcies hi hagi a l'AGNU, menys votarà a favor el país. Traduït en nombres: d'una hipotètica situació d'una AGNU amb només democràcies a una hipotètica situació d'una AGNU amb només no democràcies, el vot hipotètic disminueix un 10%. Aquesta diferència, però, no és estadísticament significativa.

Resum

Hi ha una frase atribuïda al primer ministre britànic Benjamin Disraeli i popularitzada pel cèlebre escriptor Mark Twain que diu:

«Hi ha tres tipus de mentides: mentides, mentides podrides i l'estadística.»

L'estadística és una eina molt poderosa per a explicar associacions entre variables, però també és una eina molt perillosa si no interpretem la seva utilitat correctament. Al cap i a la fi, l'únic que ens suggereix aquesta metodologia és que sembla que les dades ens donin la raó, però això no vol dir que tinguem la raó. Pensem-ho amb un exemple. Podem dir que a mesura que un país esdevé més democràtic acostuma a generar també més riquesa. Aquesta afirmació pot ser confirmada amb les tècniques d'anàlisi de dades que hem après en aquest mòdul. El que no podem afirmar amb seguretat, però, és si és la democràcia que causa la riquesa o bé si és la riquesa que causa la democràcia. En la nostra anàlisi multivariant, on haurem inclòs diverses hipòtesis alternatives a la nostra hipòtesi principal, tampoc podem dir del cert que no hi hagi altres hipòtesis rivals i plausibles que no hàgim inclòs en el model i que puguin fer trontollar les nostres conclusions. És per això que l'anàlisi de dades estadística és un primer pas, necessari però no suficient, per a respondre a les preguntes que ens fem.

Per a saber-ne més

El llibre *Models, Numbers, and Cases* de Sprinz i Wolinsky-Nahmias (2004, pàg. 129-226) conté un repàs destacable de metodologia quantitativa en estudis internacionals.

Mansfield i Pevehouse (2008) van preparar un resum per al *The Oxford Handbook of International Relations*.

En aquest sentit, la interpretació de les dades té un paper fonamental, en especial en estudis basats en l'observació de fenòmens com els estudis internacionals. Hem d'anar molt en compte a no suggerir erròniament que la correlació entre dos fenòmens implica que un causi l'altre. Tampoc no cal arribar al punt d'obsessionar-se amb observar només els nivells de significació estadística i obviar altra informació substantiva (Braumoeller i Sartori 2004, pàg. 131). Les dades ens poden donar resultats significatius però amb efectes estadísticament minúsculs simplement pel fet que disposem d'una població enorme de casos. És per això que hem d'apreciar també les bondats que té la visualització de les dades, que ens pot ajudar a observar trets substantius que ens ajudin a formular noves preguntes rellevants.

Correlacions absurdes

Podeu veure alguns exemples de correlacions absurdes en aquest web (http://www.tylervigen.com/spurious-correlations), que mostra que el nombre de persones que s'ofegaven en una piscina anualment entre 1999 i 2009 estava correlacionat amb el nombre de pel·lícules anuals en què apareixia Nicolas Cage.

En resum, aquest mòdul ha explicat algunes de les principals tècniques estadístiques per a l'anàlisi de dades com un instrument útil per a explicar fenòmens. Però tant important és utilitzar bones tècniques, sustentades matemàticament, com utilitzar bones raons, sustentades teòricament, per mirar de demostrar un argument. És per això que una interpretació humil de les dades i una comunicació clara i transparent ajudaran a afegir valor a les nostres troballes com a analistes de dades.

Exercicis d'autoavaluació

Per a un millor l'aprenentatge, intenteu fer mentalment el màxim d'exercicis possible, sense utilitzar R.

1. Dedueix la tècnica que utilitzarem segons aquestes dues variables.

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

2. Dedueix la tècnica que utilitzarem segons aquestes dues variables.

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

3. Demana la taula de contingència d'aquestes variables.

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

4. Demana la taula de contingència amb les proporcions a les files per a aquestes variables.

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

5. Canvia el codi següent per a visualitzar les proporcions de la taula de contingència.

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

6. Quin dels resultats següents no pot retornar un test de Cramer.

```
0.45, 0.15, 1.10
```

7. Hem demanat el test de Khi-quadrat de Pearson. Quin dels p-valors següents és estadísticament significatiu?

```
0.01, 0.96, 0.06
```

8. Demana la diferència de mitjanes entre aquestes dues variables.

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

9. Canvia els atributs de la geometria per demanar una recta LOESS.

```
geom_smooth()
```

10. Indica quina de les correlacions següents és la més forta.

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

11. Quin R-quadrat tindrem amb el coeficient de correlació següent?

```
1
```

12. Digues quin és el valor de y quan la x és zero en aquesta equació.

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

13. Quin és l'efecte en y per cada unitat que augmentem de x?

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

14. Indica els asteriscos que veurem amb el p-valor següent.

```
0.002
```

15. Demana l'anàlisi multivariant de les variables següents.

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

Solucionari

1. `Taula de contingència`

2. `Regressió lineal / correlació`

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 s'ha de canviar res)

10. `-0.921`

11. `1`

12. `450`

13. `-0.3`

14. `**`

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

Glossari

`chisq.test()` Demana la prova del Khi-quadrat de Pearson.

`colSums()` Suma les columnes d'una taula de contingència.

`cor()` Retorna el coeficient de correlació de dues variables.

`CramerV()` Demana el test de Cramer V per a taules de contingència (*DescTools*).

`t.test()` Demana el t-test per a diferència de mitjanes.

`lm()` Retorna el model de regressió lineal simple o múltiple, segons el nombre de variables que utilitzem.

`prop.table()` Mostra les proporcions d'una taula de contingència.

`rowSums()` Suma les files d'una taula de contingència.

`table()` Mostra una taula de contingència de dues variables categòriques.

Bibliografia

Agresti, A.; Finlay, B. (2009). *Statistical Methods for the Social Sciences*. Nova 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*. A: D. F. Sprinz; Y. Wolinsky-Nahmias, (eds.). *Models, Numbers, and Cases: Methods for Studying International Relations*. Ann Arbor, MI: University of Michigan Press (pàg. 129–151).

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

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

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.

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

Annex del mòdul

Codi 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)))
```

Codi de l'apartat 2.2

```
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()
```

Codi 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("taula", "codi"), sep = "_") %>%
  group_by_at(vars(-value)) %>%
  mutate(row_id=1:n()) %>% ungroup() %>%
  spread(key=codi, value=value) %>%
  select(-row_id)

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

Codi de l'apartat 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()
```

Codi 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("codi", "R2"), sep = "_") %>%
  group_by_at(vars(-value)) %>%
  mutate(row_id=1:n()) %>% ungroup() %>%
  spread(key=codi, 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")
```

Codi 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)))
```

Codi de l'apartat 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))
```