PRA2: Com realitzar la neteja i l'anàlisi de dades?

Àlex Franco Granell; Roger Esteban Fabró

Gener 2023

Contents

1. Descripció del dataset

Hem decidit investigar quins factors sociodemogràfics poden influir sobre la incidència de càncer als Estats Units. Per fer aquest anàlisi s'han escollit tres datasets que permeten obtenir la taxa de mortalitat per càncer en cada comtat del país juntament amb diversos paràmetres demogràfics. Concretament utilitzem els datasets de "Cancer Mortality & Incidence Rates: (Country LVL)", les dades dels EEUU de "Demographics & observation for pandemic escalation", i un dataset d'usafacts.org que recull la població dels diversos comtats del país. Els dos primers es troben disponibles a kaggle i contenen les dades bàsiques per als anàlisis fets posteriorment.

Els objectius específics dels anàlisis realitzats són els següents:

- Avaluar si la taxa de mortalitat depèn de si la població del comtat viu majoritàriament en zones rurals o urbanes.
- Investigar quines variables sociodemogràfiques dels comtats correlacionen amb la seva taxa de mortalitat per càncer.
- Generar un model de regressió lineal per a la mortalitat per càncer als comtats segons els valors de les variables sociodemogràfiques de més interès.

2. Integració i selecció de dades

A continuació, carreguem les dades originals i mostrem els primers registres per pantalla.

Dades de *covid_county_population_usafacts.csv*:

```
dpoblacio <- read.csv('./CSVs Originals/covid_county_population_usafacts.csv', sep=',')
head(dpoblacio) %>% knitr::kable()
```

countyFIPS	County.Name	State	population
0	Statewide Unallocated	AL	0
1001	Autauga County	AL	55869
1003	Baldwin County	AL	223234
1005	Barbour County	AL	24686
1007	Bibb County	AL	22394
1009	Blount County	AL	57826

Dades de *death.csv*:

```
dcancerdeath <- read.csv('./CSVs Originals/death.csv', sep=',')
head(dcancerdeath)[,1:5]  %>% knitr::kable()
```

index	County	FIPS	Met.Objective.of.45.5...1.	Age.Adjusted.Death.Rate
0	United States	0	No	46
1	Perry County, Kentucky	21193	No	125.6
2	Powell County, Kentucky	21197	No	125.3
3	North Slope Borough, Alaska	2185	No	124.9
4	Owsley County, Kentucky	21189	No	118.5
5	Union County, Florida	12125	No	113.5

Dades de *us-county.csv*:

```
dusparam <- read.csv('./CSVs Originals/us-county.csv', sep=',')
head(dusparam)[,1:9]  %>% knitr::kable()
```

fips	state	county	Confirmed	Deaths	Smokers	Obesity	Food.Environment.index	Exercise
1001	Alabama	Autauga	19	1	18.08156	33.3	7.2	69.130124
1003	Alabama	Baldwin	78	1	17.48903	31.0	8.0	73.713549
1005	Alabama	Barbour	10	0	21.99998	41.7	5.6	53.166770
1007	Alabama	Bibb	17	0	19.11420	37.6	7.8	16.251364
1009	Alabama	Blount	15	0	19.20867	33.8	8.4	15.634486
1011	Alabama	Bullock	6	0	22.89466	37.2	4.3	2.501374

A continuació modifiquem individualment els datasets per seleccionar i reanomenar les variables d'interès i integrem les dades en una sola dataframe, *uscancer*.

```r
d1 <- dcancerdeath %>%
  # Eliminem les dades a nivell estatal
  filter(index != 0) %>%
  # Generem les columnes County i State
  separate(County, c("county","state"), sep=", ") %>%
  # Seleccionem i modifiquem les variables d'interès
  transmute(FIPS,
            county = str_remove_all(county, " County"),
            state,
            met_obj_reduction = Met.Objective.of.45.5...1.,
            age_adj_deathrate = Age.Adjusted.Death.Rate,
            avg_deaths_year = Average.Deaths.per.Year,
            trend_recent_deaths = Recent.Trend..2.,
            trend_5y_deaths = Recent.5.Year.Trend..2..in.Death.Rates #,
            )

d2 <- dpoblacio %>%
  # Seleccionem les columnes de l'ID del county i la població
  transmute(FIPS = countyFIPS,
            population) %>%
  # Eliminem els registres a nivell estatal
  filter(FIPS != 0)

d3 <- dusparam %>%
  transmute(FIPS = fips,
            smokers = Smokers,
            obesity = Obesity,
            food_env_index = Food.Environment.index,
            exercise = Exercise,
            overcrowding = overcrowding,
            diabetics = Diabetics,
            insuf_sleep = Insufficient.Sleep,
            traffic_vol = Traffic.Volume,
            above_65 = X65..Above.Population,
            rural_pop = Rural.Population)

uscancer <- left_join(d1, d2, by="FIPS") %>%
  left_join(d3, by="FIPS")
```

```r
head(uscancer)[,1:6]  %>% knitr::kable()
```

FIPS	county	state	met_obj_reduction	age_adj_deathrate	avg_deaths_year
21193	Perry	Kentucky	No	125.6	43
21197	Powell	Kentucky	No	125.3	18
2185	North Slope Borough	Alaska	No	124.9	5
21189	Owsley	Kentucky	No	118.5	8
12125	Union	Florida	No	113.5	19
21147	McCreary	Kentucky	No	111.1	22

3. Neteja de dades

3.1. Gestió de valors perduts

En primer lloc, explorem les dades mitjançant `str`.

```
str(uscancer)
```

```
## 'data.frame':    3140 obs. of  19 variables:
##  $ FIPS               : int  21193 21197 2185 21189 12125 21147 21131 21159 21165 21109 ...
##  $ county             : chr  "Perry" "Powell" "North Slope Borough" "Owsley" ...
##  $ state              : chr  "Kentucky" "Kentucky" "Alaska" "Kentucky" ...
##  $ met_obj_reduction  : chr  "No" "No" "No" "No" ...
##  $ age_adj_deathrate  : chr  "125.6" "125.3" "124.9" "118.5" ...
##  $ avg_deaths_year    : chr  "43" "18" "5" "8" ...
##  $ trend_recent_deaths: chr  "stable" "stable" "**" "stable" ...
##  $ trend_5y_deaths    : chr  "-0.6" "1.7" "**" "2.2" ...
##  $ population         : int  25758 12359 9832 4415 15237 17231 9877 11195 6489 13329 ...
##  $ smokers            : num  24.5 23.9 23.9 26.1 23.3 ...
##  $ obesity            : num  41 32.6 40.1 46.3 36.9 39.8 41.4 40.5 32.9 43.5 ...
##  $ food_env_index     : num  7.3 7.5 6.9 6.9 6.5 6.6 6.7 7.2 7.8 7.1 ...
##  $ exercise           : num  78.8 86.4 100 86.7 21.6 ...
##  $ overcrowding       : num  3.3 2.82 29.95 2.38 2.31 ...
##  $ diabetics          : num  16 17.6 7.6 11.4 13.8 14.3 15.8 15.5 22.9 17.5 ...
##  $ insuf_sleep        : num  40.9 39.1 33.6 39.8 38.7 ...
##  $ traffic_vol        : num  61.97 36.8 10.14 1.39 11.57 ...
##  $ above_65           : num  16.97 15.87 6.82 19.61 14.97 ...
##  $ rural_pop          : num  74.1 67.1 59.3 100 67.4 ...
```

A continuació convertim les variables *met_obj_reduction* i *trend_recent_deaths* en factors i les variables *age_adj_deathrate*, *avg_deaths_year* i *trend_5y_deaths* en variables numèriques. També substituïm els valors * per `NA` i avaluem la quantitat de NAs al dataset mitjançant *ColSums* i *VIM::aggr()*.

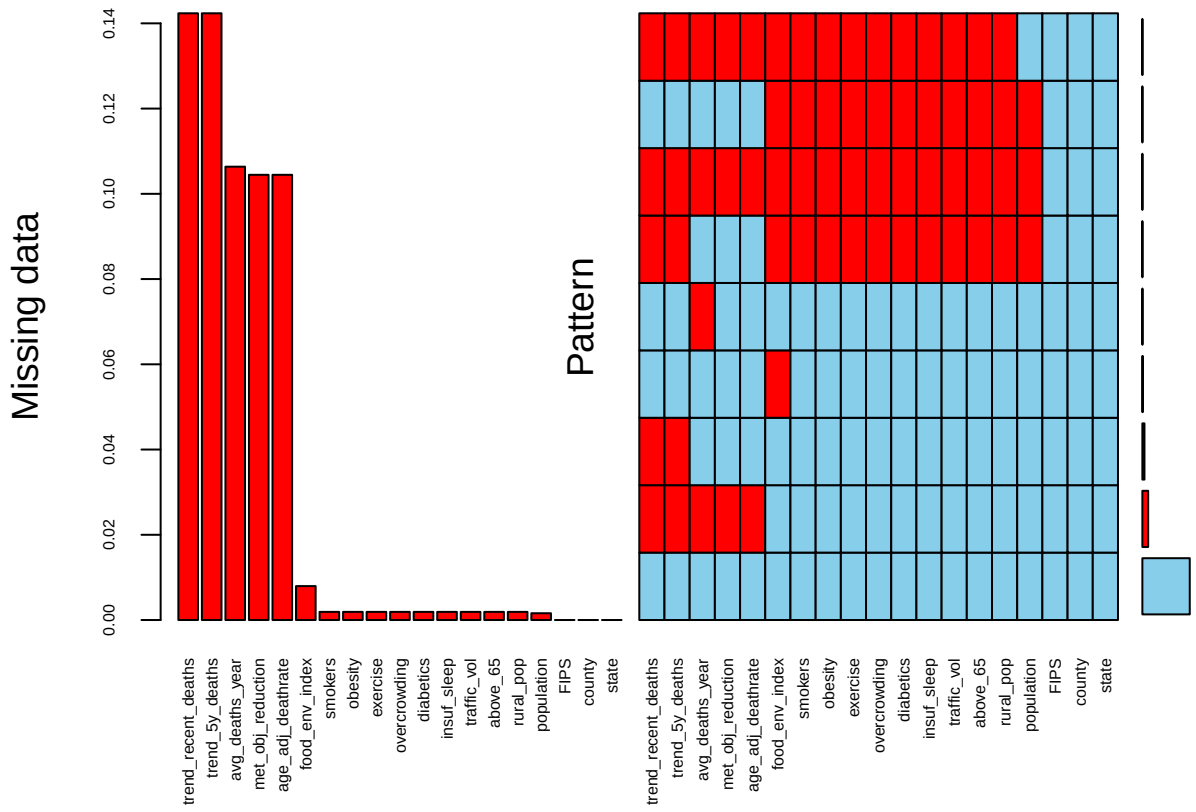
```
uscancer <- uscancer %>%
  mutate(met_obj_reduction = factor(met_obj_reduction,
                                    levels = c("Yes", "No")),
         trend_recent_deaths = factor(trend_recent_deaths,
                                      levels = c("rising", "stable","falling")),
         age_adj_deathrate = as.numeric(age_adj_deathrate),
         avg_deaths_year = as.numeric(avg_deaths_year),
         trend_5y_deaths = as.numeric(trend_5y_deaths)
         )
```

```
colSums(is.na(uscancer))
```

```
##                FIPS              county               state   met_obj_reduction
##                   0                   0                   0                 328
##   age_adj_deathrate     avg_deaths_year trend_recent_deaths     trend_5y_deaths
##                 328                 334                 447                 447
##          population             smokers             obesity      food_env_index
##                   5                   6                   6                  25
##            exercise        overcrowding           diabetics         insuf_sleep
```

```
##                   6                   6                   6                   6
##         traffic_vol            above_65           rural_pop
##                   6                   6                   6
```

```
aggr(uscancer, numbers=TRUE, sortVars=TRUE, labels=names(uscancer),
cex.axis=.5, gap=0, ylab=c("Missing data","Pattern"))
```




```
##
##  Variables sorted by number of missings:
##             Variable       Count
##  trend_recent_deaths 0.142356688
##      trend_5y_deaths 0.142356688
##      avg_deaths_year 0.106369427
##    met_obj_reduction 0.104458599
##    age_adj_deathrate 0.104458599
##       food_env_index 0.007961783
##              smokers 0.001910828
##              obesity 0.001910828
##             exercise 0.001910828
##         overcrowding 0.001910828
##            diabetics 0.001910828
##          insuf_sleep 0.001910828
##          traffic_vol 0.001910828
##             above_65 0.001910828
##            rural_pop 0.001910828
```

```
##           population 0.001592357
##                 FIPS 0.000000000
##               county 0.000000000
##                state 0.000000000
```

Observem que, respecte les dades de mortalitat per càncer, no tenim informació completa sobre el rati de mortalitat per càncer ajustada per edat (*age_adj_deathrate*) en 328 comtats, sobre la mitjana de morts per any (*avg_deaths_year*) en 334 comtats (entre es quals alguns amb elevada població com San Francisco o Los Ángeles), i tampoc tenim informació sobre les tendències en la mortalitat (*trend_recent_deaths* i *trend_5y_deaths*) en 447 comtats. Per ser una mètrica normalitzada que facilita la comparació entre comtats, centrarem els anàlisis subsegüents en les dades de rati de mortalitat per càncer ajustada per edat (*age_adj_deathrate*), que seleccionem com a variable d'interès. Observem que els 328 comtats pels què la variable *age_adj_deathrate* no té informació representen un 0.39% de la població total dels EEUU, tal i com indica la taula de sota. Per tant, prioritzant l'exactitud de les dades, hem decidit eliminar els registres que no tenen aquesta informació (`NAs` de *age_adj_deathrate*).

```
uscancer %>%
  mutate(canc_data = ifelse(is.na(age_adj_deathrate), "No", "Yes")) %>%
  group_by(canc_data) %>%
  dplyr::summarise(population = sum(population, na.rm=T),
            perc_total_pop = (sum(population, na.rm=T)/sum(uscancer$population, na.rm=T))*100) %>%
  knitr::kable()
```

canc_data	population	perc_total_pop
No	1273436	0.3880975
Yes	326849297	99.6119025

En paral · lel, tenint en compte les variables demogràfiques, que tenen menys registres incomplets, observem que a banda dels 328 comtats sense dades de *age_adj_deathrate*, 3 comtats no tenen cap dada demogràfica disponible. Per això decidim eliminar aquests registres. A més, 19 registres addicionals no disposen de dades sobre la variable *food_env_index*. Aquesta variable és un indicador de la proximitat a menjar saludable així com de la capacitat econòmica per adquirir-ne de la població. Decidim assumir que els valors podrien ser semblants entre comtats que comparteixin altres característiques del dataset, i per tant s'ha recorregut a una imputació d'aquests 19 valors perduts mitjançant la funció *kNN* del paquet *VIM*.

Finalment, observem que ja no queden valors perduts en les dades.

```
uscancer_clean <- uscancer %>%
  # Eliminem les variables de mortalitat per càncer que no ens interessen
  select(-avg_deaths_year, -trend_recent_deaths, -trend_5y_deaths) %>%
  # Eliminem els registres buits per age_adj_deathrate i
  # els que no tenen cap dada demogràfic (obesity n'és una)
  filter(!is.na(age_adj_deathrate),
         !is.na(obesity)) %>%
  # Imputem els valors NA de food_env_index amb k-Nearest Neighbours
  VIM::kNN(variable = "food_env_index") %>%
  # Eliminem la columna que indica quins registres són imputats amb kNN
  select(-food_env_index_imp)

colSums(is.na(uscancer_clean))
```

```
##              FIPS            county             state met_obj_reduction
```

```
##                 0                 0                 0                 0
## age_adj_deathrate        population           smokers           obesity
##                 0                 0                 0                 0
##    food_env_index          exercise      overcrowding         diabetics
##                 0                 0                 0                 0
##       insuf_sleep       traffic_vol          above_65         rural_pop
##                 0                 0                 0                 0
```

3.2. Gestió de valors extrems

Seguidament, avaluem la presència de valors extrems en les dades. Ja que disposem de múltiples variables i ens interessaria poder detectar outliers tenint en compte múltiples dimensions, hem decidit començar avaluant els outliers segons la distància de Mahalanobis per fer-nos una idea de quins podrien ser els casos extrems. Al codi de sota, es computa aquesta distància i es mostren els 30 comtats amb els valors de distància de Mahalanobis més elevats.

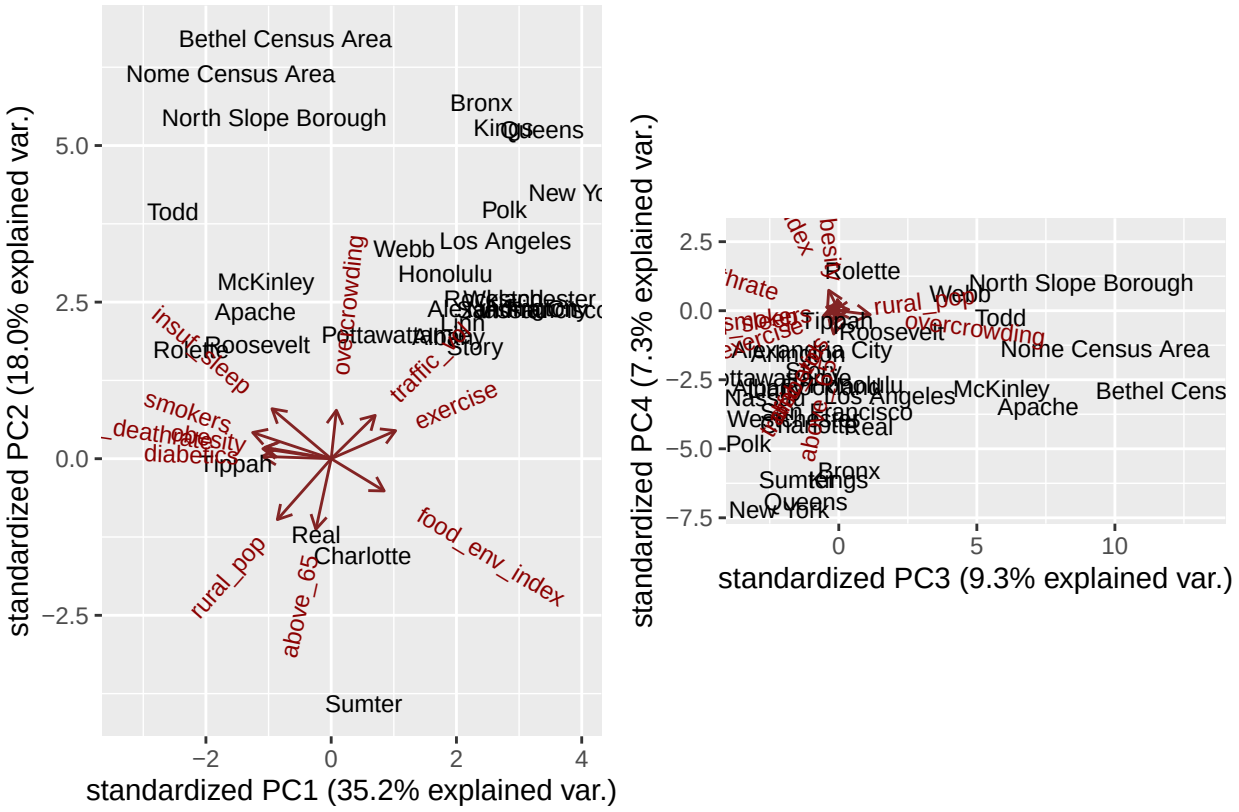
```
# Dataframe de les variables numèriques contínues de uscancer_clean
uscancer_cvars <- uscancer_clean %>%
  select(-FIPS, -county, -state,
         -population, -met_obj_reduction)

# Obtenim les posicions dels outliers per ordre decreixent de
# distància de Mahalanobis
m.dist.order <- order(mahalanobis(uscancer_cvars,
                                  colMeans(uscancer_cvars),
                                  cov(uscancer_cvars)),
                      decreasing=TRUE)

# Obtenim el nom dels outliers per ordre
m.outliers <- uscancer_clean$county[m.dist.order]

# Mostrem els top 30 outliers
m.outliers[1:30]
```

```
##  [1] "Bethel Census Area"  "North Slope Borough" "Queens"
##  [4] "New York"            "Nome Census Area"    "Polk"
##  [7] "Bronx"               "Kings"               "Todd"
## [10] "Sumter"              "Linn"                "Westchester"
## [13] "Nassau"              "Pottawattamie"       "Apache"
## [16] "Rolette"             "McKinley"            "Los Angeles"
## [19] "San Francisco"       "Roosevelt"           "Honolulu"
## [22] "Story"               "Webb"                "Arlington"
## [25] "Albany"              "Tippah"              "Alexandria City"
## [28] "Rockland"            "Charlotte"           "Real"
```

Seguidament, mostrem com es distribueixen aquests 30 comtats més extrems sobre els 4 primers components principals.

```
uscancer.pca <- prcomp(as.matrix(uscancer_cvars),
                       center = TRUE,
                       scale. = TRUE)
```

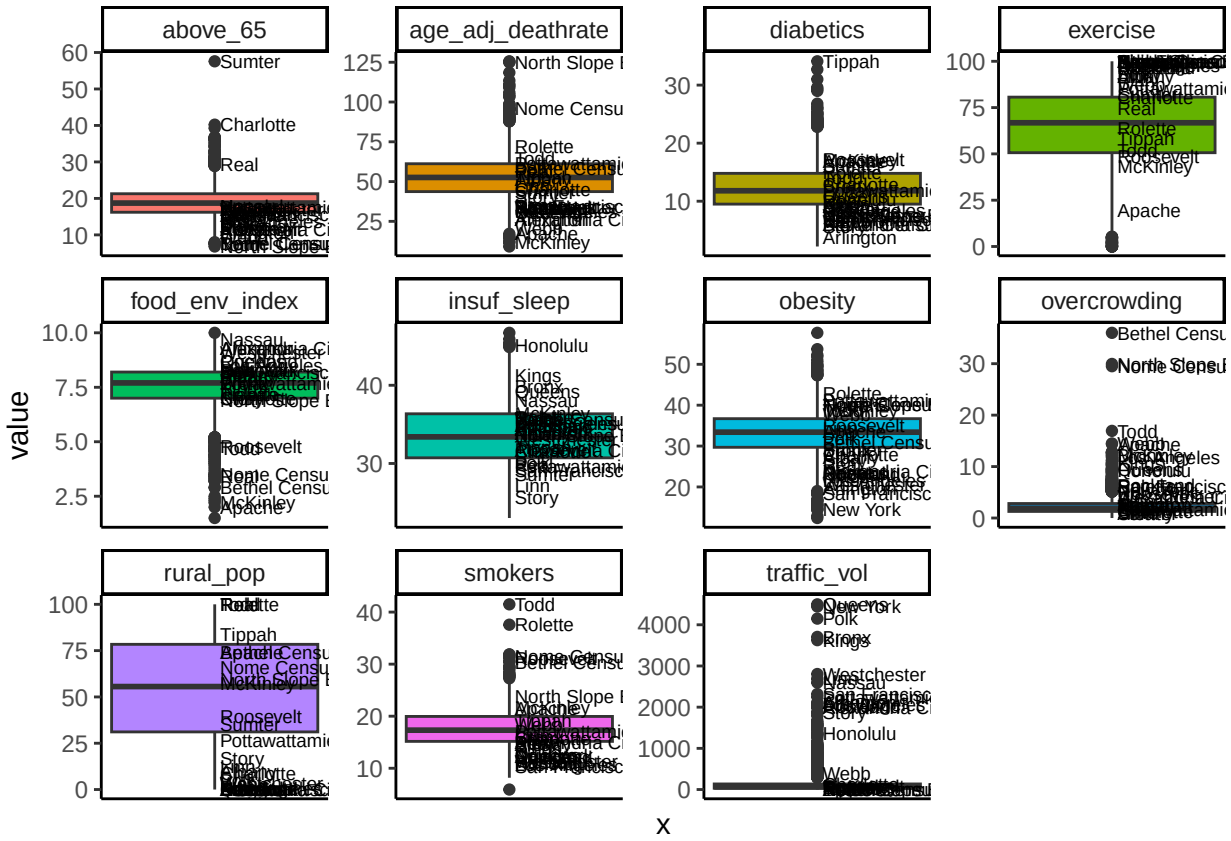
```
labels_top30_outliers <- uscancer_clean$county
labels_top30_outliers[-m.dist.order[1:30]] <- NA

# Representem els top 30 outliers sobre els 6 primers components principals
cowplot::plot_grid( ncol=2, nrow=1,
  ggbiplot(uscancer.pca,
           alpha=0.2,
           labels = labels_top30_outliers,
           choices=1:2
           )
  ,
  ggbiplot(uscancer.pca,
           alpha=0.2,
           labels = labels_top30_outliers,
           choices=3:4
           )
)
```



A sota mostrem la distribució de cada variable contínua en format boxplot, i també s'hi indiquen les posicions dels 30 comtats amb valors de Mahalanobis més elevats.

```
uscancer_cvars %>%
  mutate(top30_outliers = labels_top30_outliers) %>%
  gather(var, value, -top30_outliers) %>%
  ggplot(aes(x=1,
             y=value,
```

```
            fill=var)) +
geom_boxplot() +
geom_text(aes(x=1+0.02,
              y=value,
              label=top30_outliers),
          size=2.5,
          hjust = 0
          ) +
facet_wrap(~var, scales="free") +
theme_classic() +
theme(axis.ticks.x = element_blank(),
      axis.text.x = element_blank(),
      legend.position = "none")
```



De les dades anteriors, destaquem 4 variables amb valors molt extrems:

1. *above_65*: Observem que hi ha comtats que presenten una proporció de persones majors de 65 anys molt elevada, essent el màxim el comtat de Sumter, Florida, amb 57.6% de la població major de 65 anys. Considerem que aquests valors extrems són correctes ja que tenen una explicació sociodemogràfica: corresponen a comtats que són llocs de residència populars per a gent jubilada. Per tant, els mantenim en el dataset d'estudi.

2. *overcrowding*: Es tracta d'una variable que identifica el percentatge de població que viu en espais amb una quantitat excessivament elevada de persones. En aquest cas observem valors elevats esperables per a comtats en grans ciutats (com Nova York, San Francisco o Los Angeles). Però observem que diversos comtats rurals presenten valors extrems (entre ells Bethel Census Area, North Slope Borough i Nome

Census Area a Alaska). Tot i ser sorprenent d'entrada, sembla que l'overcrowding en regions rurals és un problema real, especialment associat a pobresa i a una població predominantment de natius americans. Per tant, decidim mantenir aquests valors extrems.

3. *smokers*: En aquest cas observem que els comtats outliers amb taxes molt elevades de fumadors (per sobre del 35% de la població) s'associen a poblacions majoritàriament de natius americans, mentre que els valors molt baixos es troben majoritàriament a l'estat de Utah, possiblement associats a població de religió mormona. El comtat de Utah, a l'estat de Utah, té un 5% de fumadors entre una població on un 82% són mormons. Atenent a aquests fets, decidim mantenir aquestes dades.

4. *traffic_vol*: els valors més extrems corresponen a grans ciutats, especialment a Nova York. Per tant, els considerem dades vàlides i també decidim mantenir aquests valors.

Observant els valors extrems de les altres variables sociodemogràfiques, trobem que totes són explicables segons les particularitats de cada comtat, com ara la taxa de diabètics del 34% de la població al comtat de Tippah, Mississippi. Per tant, tots els valors extrems trobats són explicables a causa de la diversitat existent entre regions del país, i els mantenim per als anàlisis subsegüents.

4. Anàlisi de les dades

Una vegada tenim les dades netejades, ordenades i seleccionades procedirem a l'anàlisi d'aquestes. Concretament hem decidit estudiar-les a través de tres perspectives: Primerament avaluarem l'existència de diferències en la taxa de mortalitat per càncer segons si es tracta de comtats urbans o rurals; després estudiarem si les diferents variables sociodemogràfiques de què disposem correlacionen amb la taxa mortalitat per càncer; i finalment tractarem de crear un model que puga predir la mortalitat del càncer segons les diferents variàbles donades.

4.1. Estudi de la mortalitat per càncer entre comtats rurals i urbans

Com hem comentat, en aquest apartat estudiarem la mortalitat del càncer segons si es tracta de comtats rurals o urbans. Aquest estudi és interessant perquè es comenta habitualment que els entorns urbans afavoreixen la mortalitat per càncer. Per tant és interessant saber si la mitjana poblacional dels comtats urbans és major a la mitjana dels entorns rurals.

Per fer aquest anàlisi, primerament discretitzem les dades en una nova columna, *isurban*, on guardarem els comtats que tinguen una *rural_pop* major a 50 com a 1, i la resta com a 0. Interpretem que els comtats amb 1 equivalen a entorns urbans i els que tenen valor 0 a entorns rurals.

```
# Discretització de la columna rural_pop en una nova variable
uscancer_clean$is_urban <- uscancer_clean$rural_pop
uscancer_clean$is_urban[uscancer_clean$rural_pop >= 50] <- 0
uscancer_clean$is_urban[uscancer_clean$rural_pop < 50] <- 1
```

```
## Hi ha 1592 comtats rurals i 1217 comtats urbans.
## Els comtats urbans suposen el 43.33 % dels registres,
## mentre que els comtats urbans representen el 56.67 %.
```

Una vegada tenim la variable *is_urban* discretitzada, cal estudiar la normalitat i l'homoscedasticitat de les dades de mortalitat per càncer per determinar quin test aplicar. Primerament estudiarem la normalitat de les dades amb el test de Shapiro-Wilk:

```r
# Comprovem la normalitat de les dades:
shapiro.test(uscancer_clean$age_adj_deathrate)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  uscancer_clean$age_adj_deathrate
## W = 0.98345, p-value < 2.2e-16
```

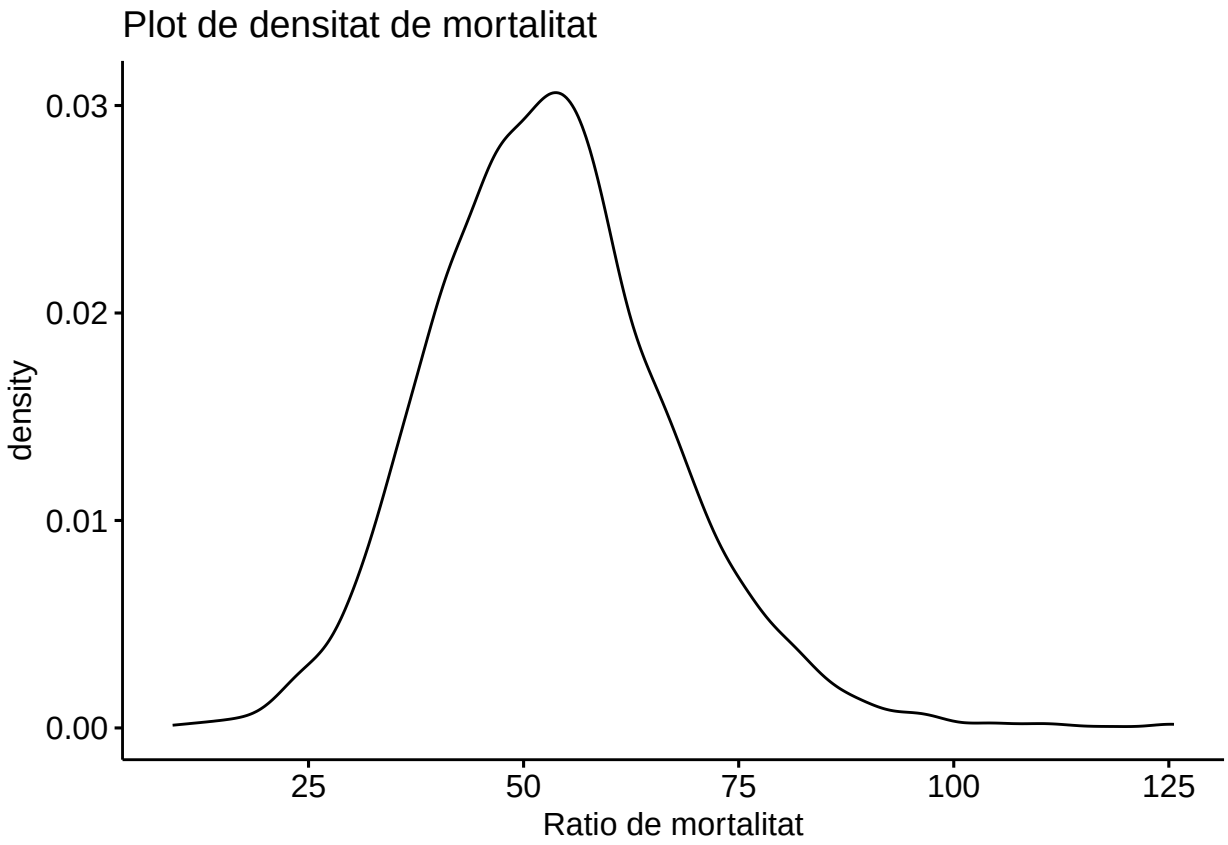
Observem que el test de Shapiro-Wilk indica que les dades no segueixen una distribució normal, ja que el p-valor és inferior al nivell de significació del 5%. Provarem a fer una transformació per millorar la normalitat:

```r
# Després d'haver fet la distribució més endavant, sabem que té una desviació
# lleugera positiva. Per tant la millor transformació és l'arrel quadrada.
dades_trans <- sqrt(uscancer_clean$age_adj_deathrate)
shapiro.test(dades_trans)
```

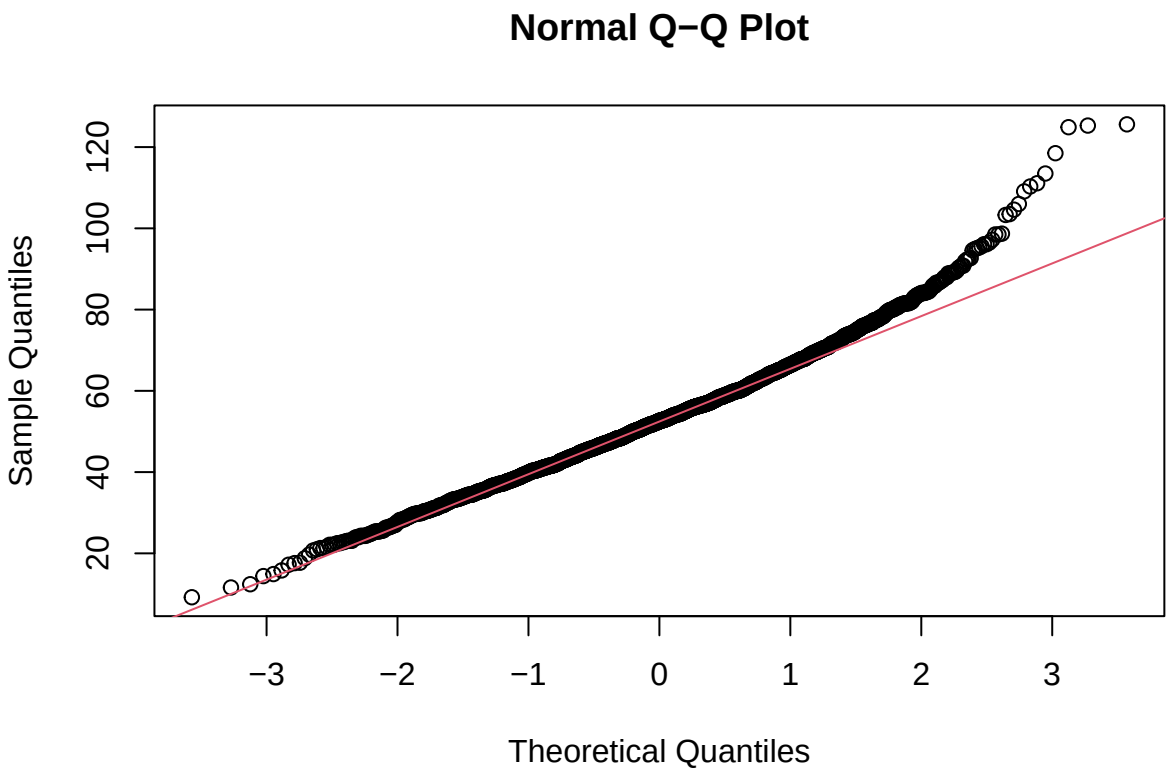
```
##
##  Shapiro-Wilk normality test
##
## data:  dades_trans
## W = 0.99583, p-value = 4.636e-07
```

Tot i que obtenim una millora en la normalitat de les dades, seguim amb un p-valor inferior a 0.05. Com que la transformació no ha estat prou per arreglar la normalitat de les dades, la descartarem per evitar complicar l'interpretació dels tests posteriors. Al veure que no es pot millorar la normalitat amb test formals, hem decidit visualitzar la distribució de la mortalitat per veure si s'aproxima a una distribució normal. A les gràfiques de sota s'hi pot veure que la nostra distribució s'aproxima a una distribució normal amb una cua cap a la dreta, i això fa que només presenti una lleugera desviació del patró d'una distribució normal en el marge dret del Q-Q plot. Atenent a aquestes dades i segons el teorema central del límit, decidim assumir que la mitjana mostral resultant tendirà a una distribució normal i aplicarem un test paramètric per a la comparació de les mitjanes de mortalitat entre comtats rurals i urbans.

```r
ggdensity(uscancer_clean$age_adj_deathrate,
          main = "Plot de densitat de mortalitat",
          xlab = "Ratio de mortalitat")
```




```
qqnorm(uscancer_clean$age_adj_deathrate)
qqline(uscancer_clean$age_adj_deathrate,col=2)
```

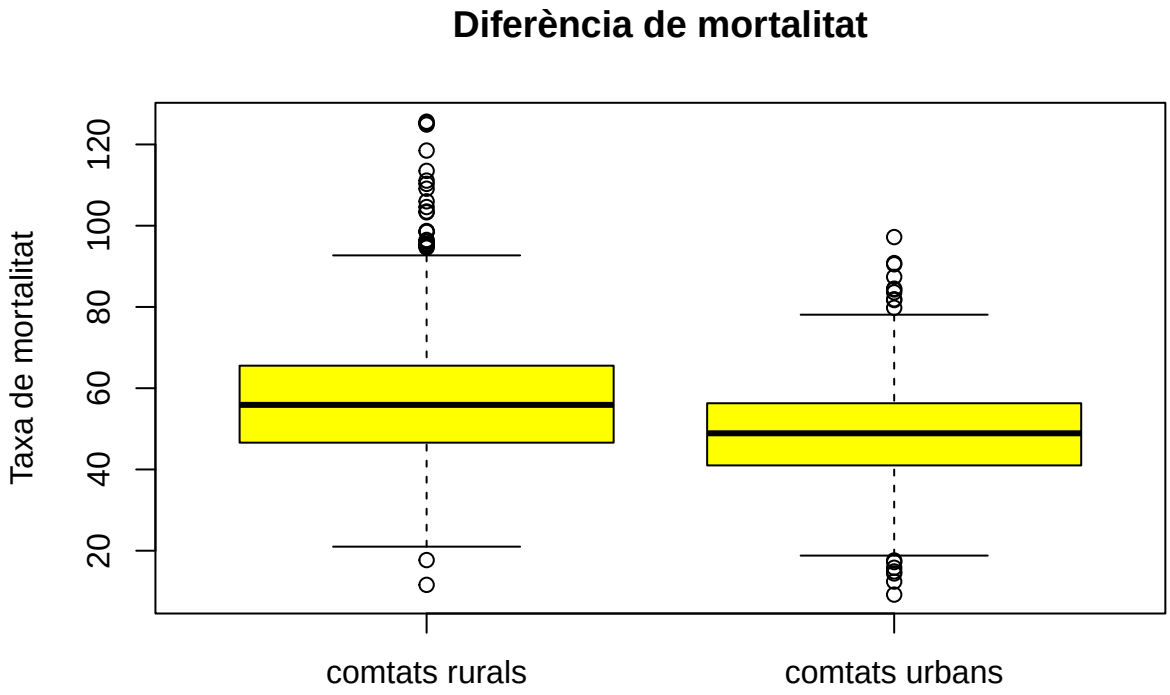


A continuacio avaluem si els dos nivells de la variable *is_urban* que volem comparar tenen una variància igual per la ràtio de mortalitat. Comprovem l'homoscedasticitat amb el codi següent, i representem les distribucions dels valors de la variància en comtats rurals i urbans mitjançant un boxplot:

```
# Separació de les dades
rural <- uscancer_clean[!(uscancer_clean$is_urban == 1),]
urban <- uscancer_clean[!(uscancer_clean$is_urban == 0),]


var.test(rural$age_adj_deathrate,
         urban$age_adj_deathrate,
         conf.level = 0.95)
```

```
##
##  F test to compare two variances
##
## data:  rural$age_adj_deathrate and urban$age_adj_deathrate
## F = 1.577, num df = 1591, denom df = 1216, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.418357 1.751982
## sample estimates:
## ratio of variances
##           1.576964
```

```
boxplot(rural$age_adj_deathrate,
        urban$age_adj_deathrate,
        col="yellow",ylab="Taxa de mortalitat",
        main="Diferència de mortalitat",
        names=c("comtats rurals","comtats urbans"))
```



A través d'una inspecció visual, i del var.test, podem concloure que ambdues poblacions tenen variàncies diferents. Al var.test, veiem que el p-valor és inferior al nivell de significació 0.05, fet que ens fa rebutjar la hipòtesi nul · la d'homoscedasticitat.

Ara, finalment, aplicarem un test per comparar dues poblacions independents, que assumim que segueixen una distribució normal, amb variàncies desconegudes i diferents. La nostra hipòtesi nul · la és que les mitjanes són iguals per a ambdues poblacions; i la nostra hipòtesi alternativa és que la mitjana dels entorns urbans és diferent a la rural. És a dir:

H0 = y(urban) = y(rural)

H1 = y(urban) != y(rural)

```
t.test(urban$age_adj_deathrate,
       rural$age_adj_deathrate,
       alternative="two.sided",
       var.equal=FALSE,
       conf.level=0.95)
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  urban$age_adj_deathrate and rural$age_adj_deathrate
## t = -15.628, df = 2802.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.776383 -6.819618
## sample estimates:
## mean of x mean of y
##  48.76902  56.56702
```

Amb una confiança del 95% podem rebutjar la hipòtesi nul · la i concloure que la mitjana del ràtio de mortalitat en entorns rurals és major que la mitjana als entorns urbans És a dir: y(rural) > y(urban). Extraiem aquesta conclusió perquè el p-valor és menor que la significància escollida, 0.05, i per tant s'ha de descartar l'hipòtesi nul · la i acceptar l'alternativa. Observant els valors de les mitjanes i el boxplot anterior, concloem que els comtats rurals tenen valors de mortalitat per càncer superiors als urbans.

Cal destacar que aquest resultat seria contrari a allò que sospitàvem inicialment, fet que ens condueix a pensar que altres factors podrien emmascarar l'impacte ambiental de les ciutats sobre la mortalitat per càncer: algun d'aquests factors (que no hem avaluat aquí per no tenir-ne les dades) podria ser la diferència en el nivell econòmic entre comtats o la facilitat d'accés a centres hospitalaris.

4.2 Estudi de la correlació entre mortalitat per càncer i variables sociodemogràfiques

Seguidament, ens interessa determinar si existeix una correlació entre *age_adj_deathrate* i les 9 variables sociodemogràfiques *smokers*, *obesity*, *food_env_index*, *exercise*, *overcrowding*, *diabetics*, *insuf_sleep*, *traffic_vol* i *rural_pop*. Hem exclòs la variable *above_65* ja que *age_adj_deathrate* és una mètrica que ja té en compte l'edat de la població.

```
# Guardem els valors de mortalitat en l'objecte mortality
mortality = uscancer_clean$age_adj_deathrate

# Convertim uscancer_clean a un format llarg de dataframe per fer els
# tests i plots alhora per a totes les variables
uscancer_clean_long <- uscancer_clean %>%
  select(smokers, obesity, food_env_index, exercise, overcrowding,
         diabetics, insuf_sleep, traffic_vol, rural_pop) %>%
  gather(variable, value)
```

En primer lloc, necessitem comprovar la normalitat en la distribució dels valors d'aquestes variables, i ho fem mitjançant el test de Shapiro-Wilk.

```
# Fem servir dplyr per avaluar si les variables d'interès s'ajusten a una
# distribució normal
uscancer_clean_long  %>%
  group_by(variable) %>%
  dplyr::summarise(shapiro_pval = shapiro.test(value)$p.value) %>%
  mutate(shapiro_pval = formatC(shapiro_pval, format = "e", digits = 2)) %>%
  knitr::kable()
```

variable	shapiro_pval
diabetics	1.68e-22
exercise	1.40e-24
food_env_index	3.04e-33
insuf_sleep	8.77e-05
obesity	2.54e-06
overcrowding	1.07e-59
rural_pop	6.79e-30
smokers	4.33e-19
traffic_vol	9.76e-71

En tots els casos, el p-valor associat al test de Shapiro-Wilk és inferior a 0.05, fet que ens porta a rebutjar la hipòtesi nul · la de normalitat de les dades. Això ens condueix a aplicar el test no paramètric de Spearman per avaluar el grau de dependència entre aquestes variables i la mortalitat per càncer. A continuació, realitzem un test de Spearman per cada variable versus la taxa de mortalitat per càncer, i apliquem la correcció de Benjamini-Hochberg per a comparacions múltiples.

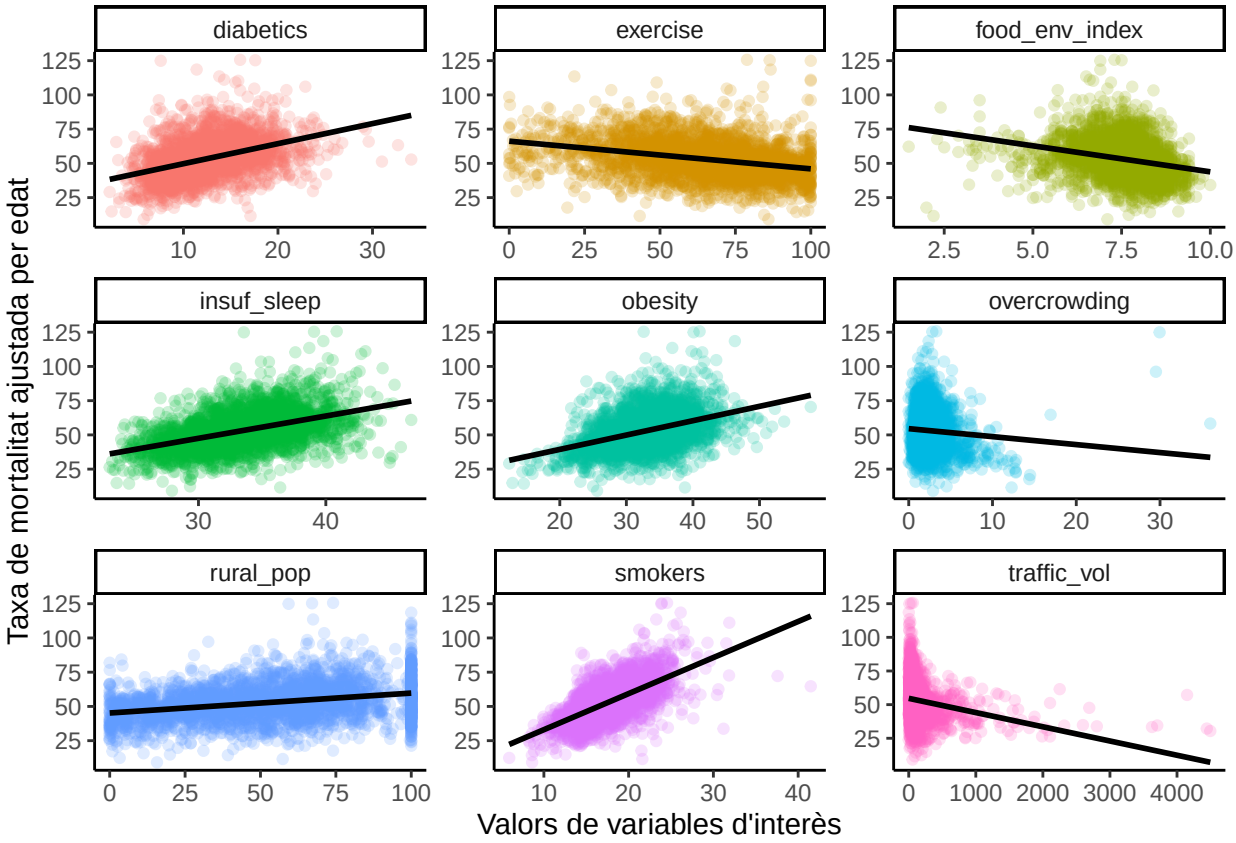
```
# Fem servir dplyr per avaluar la correlació entre la mortalitat i les variables
# d'interès
uscancer_clean_long %>%
  group_by(variable) %>%
  dplyr::summarise(spearman_pval = cor.test(value, mortality,
                                            method = "spearman")$p.value,
                   spearman_r = cor.test(value, mortality,
                                         method = "spearman")$estimate) %>%
  mutate(pval_adjusted = p.adjust(spearman_pval, method = "BH") ) %>%
  mutate(spearman_pval = formatC(spearman_pval, format = "e", digits = 2),
         spearman_r = round(spearman_r, 2),
         pval_adjusted = formatC(pval_adjusted, format = "e", digits = 2)) %>%
  knitr::kable()
```

variable	spearman_pval	spearman_r	pval_adjusted
diabetics	2.99e-152	0.47	8.96e-152
exercise	1.68e-82	-0.35	3.03e-82
food_env_index	1.90e-80	-0.35	2.85e-80
insuf_sleep	2.19e-161	0.48	9.87e-161
obesity	1.57e-108	0.40	3.53e-108
overcrowding	6.13e-01	-0.01	6.13e-01
rural_pop	4.10e-62	0.31	4.61e-62
smokers	0.00e+00	0.65	0.00e+00
traffic_vol	6.14e-67	-0.32	7.89e-67

Addicionalment, visualitzem aquestes correlacions mitjançant dotplots on hi afegim una recta de regressió seguint un model lineal (y ~ x).

```
ggplot(uscancer_clean_long %>%
         mutate(mortality = rep(mortality,9)),
       aes(x=value,
           y=mortality,
           color=variable)) +
```

```
  geom_point(alpha=0.2) +
  geom_smooth(method="lm", color="black", se=FALSE) +
  facet_wrap(~variable, scales = "free") +
  theme_classic() +
  theme(legend.position="none") +
  labs(y = "Taxa de mortalitat ajustada per edat",
       x = "Valors de variables d'interès")
```



Atenent als resultats anteriors, podem concloure que les variables *smokers*, *diabetics*, *insuf_sleep*, *obesity* i *rural_pop* presenten una correlació positiva significativa amb la taxa de mort per càcner ajustada per edat (p-valor<0.05). Considerant els coeficients de correlació de Spearman obtinguts (*spearman_r* a la taula anterior), observem que la variable *smokers* és la que presenta la correlació més forta i positiva amb *age_adj_deathrate* (r=0.65), seguida de *insuf_sleep* (r=0.48) i *diabetics* (r=0.47).

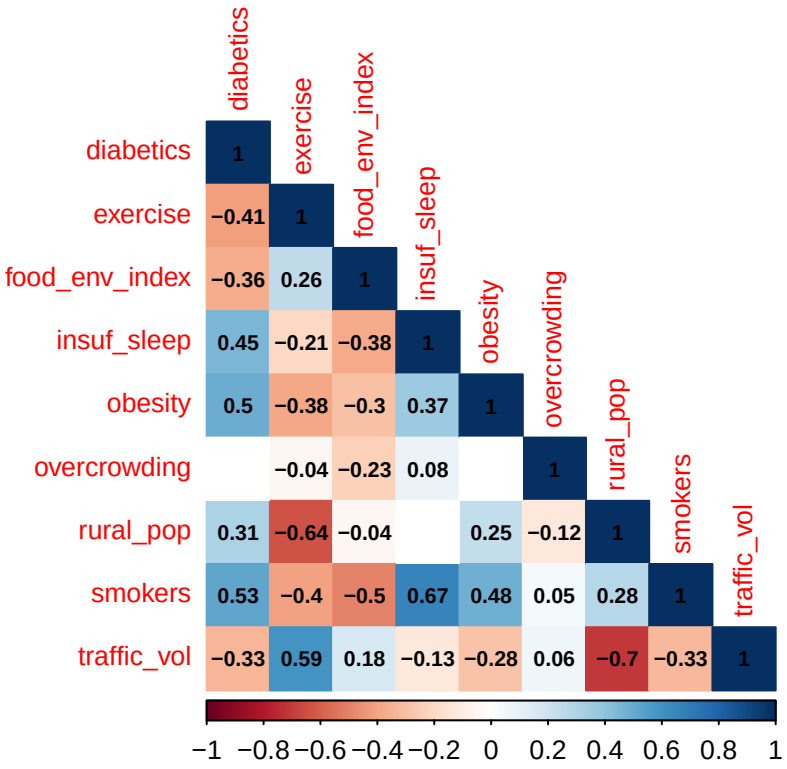
Per contra, les variables *exercise*, *food_env_index* i *traffic_vol* presenten una correlació negativa significativa amb la mortalitat per càncer als comtats, encara que els seus coeficients de correlació indiquen que no són correlacions gaire fortes. A més, s'observa que l'overcrowding (quantitat de persones convivint en un mateix espai per sobre dels màxims recomanables) no correlaciona significativament amb la taxa de mortalitat per càncer (p-valor>0.05), fet que té sentit en no ser una malaltia infecciosa transmissible.

Val a dir que la majoria de resultats coincideixen amb allò esperat. Val a dir que, de nou, observem com el percentatge de població rural correlaciona positivament amb la taxa de mortalitat per càncer, mentre que el volum de trànsit correlaciona negativament. Això ens fa pensar que la variable *traffic_vol* podria estar directament correlacionada amb la presència de ciutats als comtats.

Al següent apartat ens interessaria utilitzar algunes de les variables sociodemogràfiques anteriors per a generar un model de regressió lineal múltiple amb *age_adj_deathrate* com a variable dependent. Amb

aquest aobjectiu, decidim partir de les 8 variables que hem observat que correlacionen significativament amb *age_adj_deathrate*. Addicionalment, a continuació avaluem si algunes d'aquestes 8 variables correlacionen entre elles per tal de detectar i evitar un potencial problema de duplicació d'informació.

```
# Generem una matriu amb le svariables d'interès
mx <- uscancer_clean_long %>%
  mutate(id = rep(1:(nrow(.)/9),9)) %>%
  spread(variable, value) %>%
  select(-id) %>%
  as.matrix()

# Obtenim els coeicients de correlació i els p-valors i els representem
# gràficament amb una matriu de correlacions
corr.res = rcorr(mx, type = "spearman")
corr.res$P[is.na(corr.res$P)] <- 0

corrplot(corr.res$r,
         method = "color",
         type = 'lower',
         addCoef.col = 'black',
         p.mat=corr.res$P,
         insig = "blank",
         number.cex=.7,
         tl.cex=.8)
```



De les correlacions entre les variables anteriors, destaquem que *exercise* correlaciona positivament i amb certa

força amb *traffic_vol* i negativament amb *rural_pop*, *rural_pop* i *traffic_vol* correlacionen negativament i *smokers* correlaciona positivament amb *insuf_sleep*.

Tenint això en compte, decidim generar un model de regressió lineal múltiple amb les variables *diabetics*, *smokers*, *food_env_index*, *obesity* i *rural_pop*.

4.3 Model de regressió per a la mortalitat per càncer

A continuació creem un model de regressió lineal múltple amb *diabetics*, *smokers*, *food_env_index*, *obesity* i *rural_pop* com a variables explicatives i *age_adj_deathrate* com a variable dependent.

```
uscancer_rlm <- lm(age_adj_deathrate ~ diabetics + smokers +
                     food_env_index + obesity + rural_pop,
                   data=uscancer_clean)
summary(uscancer_rlm)
```

```
##
## Call:
## lm(formula = age_adj_deathrate ~ diabetics + smokers + food_env_index +
##     obesity + rural_pop, data = uscancer_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.713  -6.219  -0.012   6.002  58.443
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -8.164663   2.739471  -2.980 0.002904 **
## diabetics       0.380575   0.061241   6.214 5.92e-10 ***
## smokers         2.330308   0.075933  30.689  < 2e-16 ***
## food_env_index  1.024689   0.229522   4.464 8.34e-06 ***
## obesity         0.160310   0.045449   3.527 0.000427 ***
## rural_pop       0.044670   0.007051   6.335 2.75e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.44 on 2803 degrees of freedom
## Multiple R-squared:  0.4485, Adjusted R-squared:  0.4475
## F-statistic: 455.9 on 5 and 2803 DF,  p-value: < 2.2e-16
```

Observem que la fórmula de la recta de regressió de la taxa de mortalitat per càncer davant les variables *diabetics*, *smokers*, *food_env_index*, *obesity* i *rural_pop* és la següent:

$$Y = -8.16 + 0.38X_{diabetics} + 2.33X_{smokers} + 1.02X_{food\_env\_index} + 0.16X_{obesity} + 0.04X_{rural\_pop}$$

Tots els coeficients estimats són diferents de zero amb un p-valor < 0.05. Atenent al coeficient de determinació de 0.4485, podem dir que un 44.85% de la variància dels valors de la mostra és explicada per la recta de regressió. Per tant, es tracta d'un model relativament pobre i amb capacitat de millora, possiblement perquè les variables disponibles són només una part dels múltiples factors que determinen la incidència i mortalitat del càncer en cada comtat.

D'altra banda, per determinar si hi ha col · linearitat entre les variables que composen el model, calculem el FIV (factor d'inflació de la variància).

```
faraway::vif(uscancer_rlm)
```

```
##      diabetics        smokers food_env_index        obesity      rural_pop
##       1.550392       1.760966       1.390442       1.530755       1.175527
```

Observem que els valors de FIV són relativament baixos (propers a 1). Per tant, les variables introduïdes al model no serien redundants entre elles.

A continuació, com que la variable *insuf_sleep* és una de les que s'ha assoiat amb un coeficient de correlació de Spearman més elevat, ens interessaria avaluar un nou model de regressió lineal on s'introduís aquesta variable:

```
uscancer_rlm2 <- lm(age_adj_deathrate ~ diabetics + smokers +
                      food_env_index + obesity + rural_pop +
                      insuf_sleep,
                    data=uscancer_clean)
summary(uscancer_rlm2)
```

```
##
## Call:
## lm(formula = age_adj_deathrate ~ diabetics + smokers + food_env_index +
##     obesity + rural_pop + insuf_sleep, data = uscancer_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.522  -6.039  -0.048   5.940  59.829
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -16.60295    3.06812  -5.411 6.78e-08 ***
## diabetics        0.30462    0.06218   4.899 1.02e-06 ***
## smokers          2.05966    0.08805  23.391  < 2e-16 ***
## food_env_index   1.08632    0.22835   4.757 2.06e-06 ***
## obesity          0.14327    0.04526   3.165  0.00157 **
## rural_pop        0.05771    0.00734   7.861 5.37e-15 ***
## insuf_sleep      0.40430    0.06776   5.967 2.73e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.37 on 2802 degrees of freedom
## Multiple R-squared:  0.4554, Adjusted R-squared:  0.4543
## F-statistic: 390.6 on 6 and 2802 DF,  p-value: < 2.2e-16
```

```
faraway::vif(uscancer_rlm2)
```

```
##      diabetics        smokers food_env_index        obesity      rural_pop
##       1.618215       2.397153       1.393292       1.536876       1.289817
##    insuf_sleep
##       1.930405
```

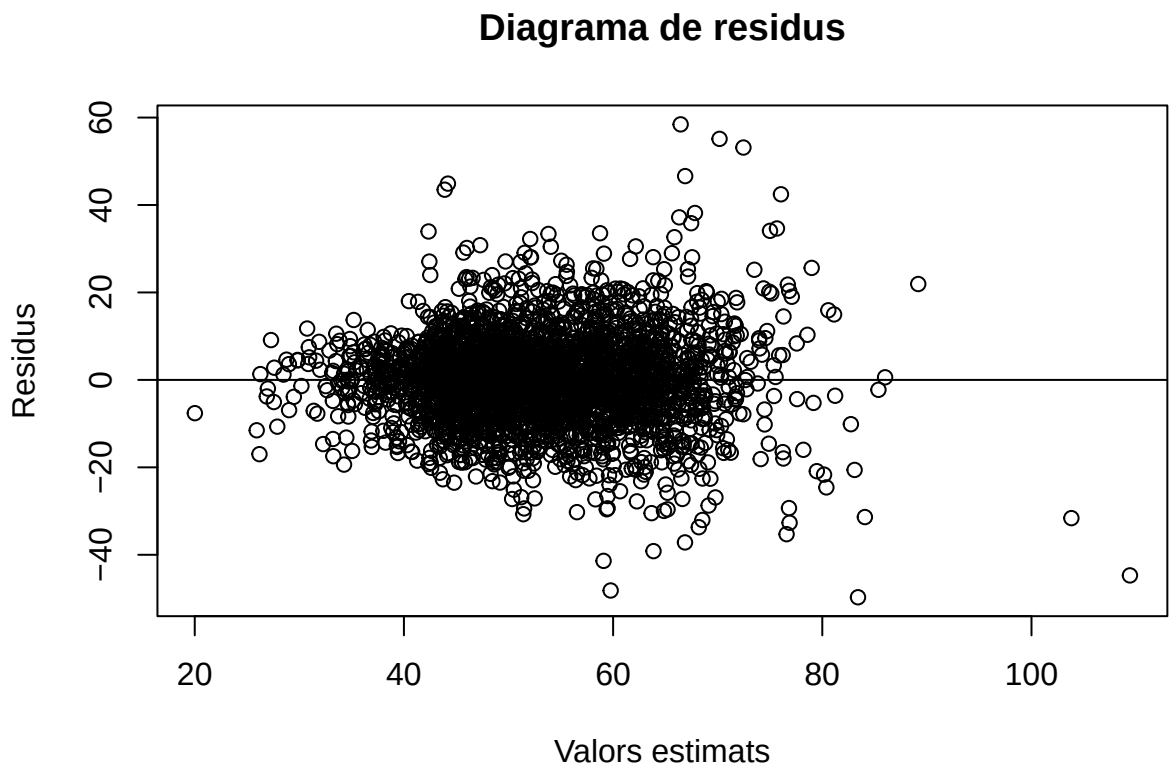
De nou, observem que tots els coeficients estimats són diferents de zero amb un p-valor < 0.05. Atenent al coeficient de determinació de 0.4554, podem dir que un 45.54% de la variància dels valors de la mostra

és explicada per la recta de regressió. Es tracta d'un benefici marginal respecte el 44.85% de la variància explicada pel model sense la variable *insuf_sleep*. A banda, els valors de FIV segueixen sent relativament baixos, encara que la introducció de *insuf_sleep* ha fet augmentar el FIV de *smokers*, fet que podria indicar una certa col · linearitat (que ja intuíem per la correlació positiva trobada anteriorment entre les variables).

Finalment, realitzem una diagnosi del primer model de regressió múltiple creat mitjançant l'elaboració de dos gràfics: un amb els valors ajustats davant dels residus (que ens permetrà veure si la variància és constant) i un segon gràfic quantil-quantil que compara els residus del model amb els valors d'una variable que es distribueix normalment(QQ plot).

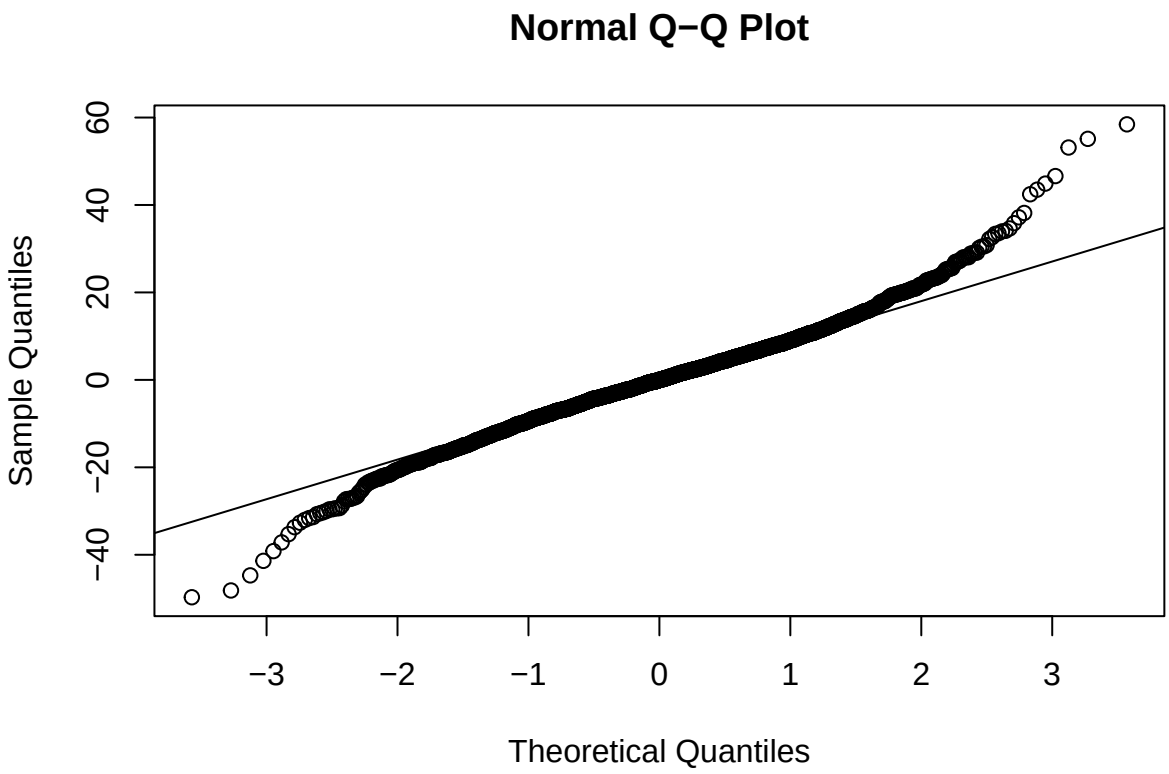
En primer lloc, el gràfic dels valors ajustats enfront dels residus permet veure l'adequació del model. En aquest cas, tal i com es veu al gràfic de sota, hi ha un patró aleatori força homogeni al voltant del residu 0, fet que indicaria que la variància seria constant i el model de regressió lineal múltiple podria ser adequat per a aquestes dades.

```
plot(residuals(uscancer_rlm)~fitted.values(uscancer_rlm), main = "Diagrama de residus",
xlab = "Valors estimats", ylab = "Residus")
abline(0,0) #Generem una línia horitzontal en i=0
```



Seguidament, el Q-Q plot permet identificar que els quantils dels residus del model s'ajusten força als d'una distribució normal per als valors centrals, encara que els valors extrems tendeixen a desviar-se lleugerament d'aquest patró.

```
qqnorm(residuals(uscancer_rlm))
qqline(residuals(uscancer_rlm))
```



5. Representació dels resultats

Les taules i gràfiques s'han aportat al llarg de la pràctica.

6. Resolució del problema. Conclusions.

Atenent als resultats obtinguts, podem afirmar que hem realitzat una anàlisi de l'impacte de certs factors sociodemogràfics sobre les taxes de mortalitat per càncer als Estats Units, partint de la informació associada a cada comtat del país. Concretament, podem concloure el següent respecte els 3 sub-objectius especificats a l'apartat 1:

- Els comtats amb una major proporció de població en entorns rurals presenten taxes de mortalitat per càncer més elevades.
- Les variables que correlacionen positivament amb la taxa de mortalitat per càncer als comtats són les proporcions de població fumadora, amb diabetis, amb obesitat, amb son insuficient i en entorns rurals. Per contra, s'ha observat una correlació negativa entre la mortalitat per càncer i les taxes de gent que s'exercita, que té un accés més fàcil a menjar saludable i el volum de trànsit dels comtats.
- S'ha generat un model de regressió lineal amb la mortalitat per càncer com a variable dependent i amb les variables explicatives de la proporció de diabètics, fumadors, obesos, població rural i facilitat d'accés a menjar saludable. S'ha confirmat que totes les variables són significatives, encara que només un 44.85% de la variància en els valors de la mostra és explicada pel model.

En general, observem que bona part de les variables sociodemogràfiques estudiades tenen una associació amb la mortalitat per càncer als comtats. Tot i així, possiblement calgui tenir en compte altres factors (com ara factors genètics, o variables sociodemogràfiques addicionals) per generar un model amb un millor rendiment.