Practica2

Luis D. Hilario
10 de enero de 2019

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

Nos encontramos con un conjunto de datos compuesto de dos ficheros que contienen datos reales de ventas de un laboratorio farmacéutico y datos de ventas de distintas farmacias a lo largo del territorio nacional. Dichos datos han sido obtenidos del ERP de la empresa y de una consultora externa respectivamente. En concreto, tenemos:

- Fichero sell-in: contiene los datos de ventas que el laboratorio realiza a sus 13045 clientes; en particular, tenemos las ventas en número de unidades de tres productos exclusivos y las ventas en unidades monetarias del resto de productos que el laboratorio también les vende. Además, las ventas vienen desglosadas por población.

- Fichero sell-out: contiene los datos de ventas totales que una muestra de 438 farmacias realiza a sus clientes, es decir, son ventas de todos los productos que realiza cada establecimiento al cliente final.

La denominación “sell-in” y “sell-out” corresponde al origen de la información. Así, “sell-in” es un fichero de ventas internas que obtenemos del propio sistema informático de gestión del laboratorio y “sell-out” es un fichero de ventas externas, que obtenemos de una consultora externa la cual realiza un muestreo de 438 clientes de los cuales obtiene los datos mencionados anteriormente.

A partir de los datos suministrados se pretende generar un modelo predictivo para estimar el potencial que tenemos para aumentar las ventas que el laboratorio realiza a las distintas farmacias. La idea sobre la que gira el modelo es la siguiente: usaremos la muestra de 438 clientes, de los cuales conocemos sus ventas totales y las compras que hacen al laboratorio, para estimar las ventas totales del resto de farmacias, las cuales son desconocidas.

Con estas ventas totales podremos averiguar qué cantidad de compras están realizando a otros laboratorios, calculada como la diferencia entre sus ventas a clientes finales y el total de las compras que realizan al laboratorio menos el margen comercial medio. Esta diferencia determinará el potencial de venta o recorrido que tiene cada farmacia para el laboratorio, de manera que se pueda elaborar una estrategia específica para aumentar las ventas. En particular, nos interesa el potencial de ventas de nuestros tres mejores productos, de ahí la necesidad de incluirlos en el modelo.

2. Integración y selección de los datos de interés a analizar.

Los ficheros obtenidos se encuentran en formato "csv" por lo que pueden ser manejados con cualquier herramienta de análisis de datos. Lo primero que tenemos que tener en cuenta es que, para generar nuestro modelo, es necesario conocer las ventas que el laboratorio realiza a cada farmacia que aparece en la muestra, ya que en el fichero muestral no tenemos este dato. Este dato, en cambio, si lo tenemos en el fichero suministrado por el laboratorio.

Por otro lado, hemos de tener en cuenta que los clientes aparecen con sus datos, los cuales es necesario anonimizar previamente para cumplir con el RGPD. El proceso de anonimización consistiría en hacer corresponder cada registro de cliente con un código identificativo único, de manera que dicha empresa no pueda ser identificable por terceros. Solo quien tiene autorización para ver datos de cliente debe conocer este dato. Este proceso debe ser realizado con ambos ficheros. Por razones obvias, al ser datos reales, no muestro este proceso y para esta actividad parto de los ficheros anonimizados. Los ficheros leídos tienen la siguiente estructura y contenido:

Fichero **SELLIN**

```
# Cargamos los ficheros

SELLIN <- read.csv2("sell-in.csv")
head(SELLIN)
```

```
##   Cod_Cli PROVINCIA Uds_AH Uds_DNT Uds_VH Resto_Farmacos
## 1 anon_S0     ALAVA  90300     162    117     21317197.4
## 2 anon_S1     ALAVA 127200     298     42       321079.2
## 3 anon_S2     ALAVA  45825     286     84      1749363.5
## 4 anon_S3     ALAVA 258600     152     47      3960844.2
## 5 anon_S4     ALAVA 210100     327    164      3079924.3
## 6 anon_S5     ALAVA   7800      35      8       327995.5
```

Fichero **SELLOUT**

```
SELLOUT <- read.csv2("sell-out.csv")
head(SELLOUT)
```

```
##     Cod_Cli TOTAL_VOL
## 1  anon_S69  29506650
## 2 anon_S103 323450171
## 3 anon_S107 126806078
## 4 anon_S160   9523747
## 5 anon_S170 150234420
## 6 anon_S238 273268756
```

En esta fase integramos los dos ficheros en uno, uniéndolos por el código de cliente anonimizado y seleccionando los campos de nuestro interés para la integración. Por último, dado que el fichero final integrará los registros de ambos, podemos añadir una variable dicotómica con el objetivo de saber si un registro concreto pertenece a la muestra o no.

El fichero integrado tendría la siguiente forma:

```
# Añadimos a sellout los campos conocidos presentes en sellin
DATOS_UNIDOS <- merge(SELLIN,SELLOUT)
head(DATOS_UNIDOS)
```

```
##       Cod_Cli  PROVINCIA Uds_AH Uds_DNT Uds_VH Resto_Farmacos
TOTAL_VOL
## 1 anon_S10069   ZARAGOZA      0       6     33        1811288
14985571
## 2 anon_S10181     LLEIDA  40200      28     33        3116607
65729190
## 3 anon_S10199 PONTEVEDRA  88600      36     40        6335775
77156074
## 4 anon_S10217    CACERES  38400      19     16        3180451
41015251
## 5   anon_S103   ALBACETE   5700     140    520       17555795
323450171
## 6 anon_S10361     HUELVA  72300      97    199       20349010
288057061
```

En resumen:

- Anonimizar datos
- Seleccionar campos necesarios
- Integrar los ficheros en uno
- Añadir una variable dicotómica

Nota: la variable dicotómica no es necesario añadirla durante este análisis, sino que se podrá añadir en una fase posterior de este proceso iterativo con el objetivo de dejar indicado en el fichero final los registros pertenecientes a la muestra.

3. Limpieza de los datos.

3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

En este caso, al unir los ficheros, aparece un campo ?TOTAL_VOL? que no se encuentra en ?sell-in? por lo que este valor queda indeterminado. Esto lo manejamos utilizando la función ?merge? de manera que se genere un tercer fichero solo para los códigos de clientes que aparecen en la muestra y que son los únicos códigos comunes a ambos ficheros.

Una vez realizada la fusión, para comprobar si los datos están completos, ejecutaríamos el siguiente ?test de valores ausentes?.

```
# Ejecutamos el test de valores ausentes para verificar que los
# datos están completos

sapply(DATOS_UNIDOS, function(x)(sum(is.na(x))))
```

```
##        Cod_Cli      PROVINCIA         Uds_AH        Uds_DNT
Uds_VH
##              0              0              0              0
0
## Resto_Farmacos      TOTAL_VOL
##              0              0
```

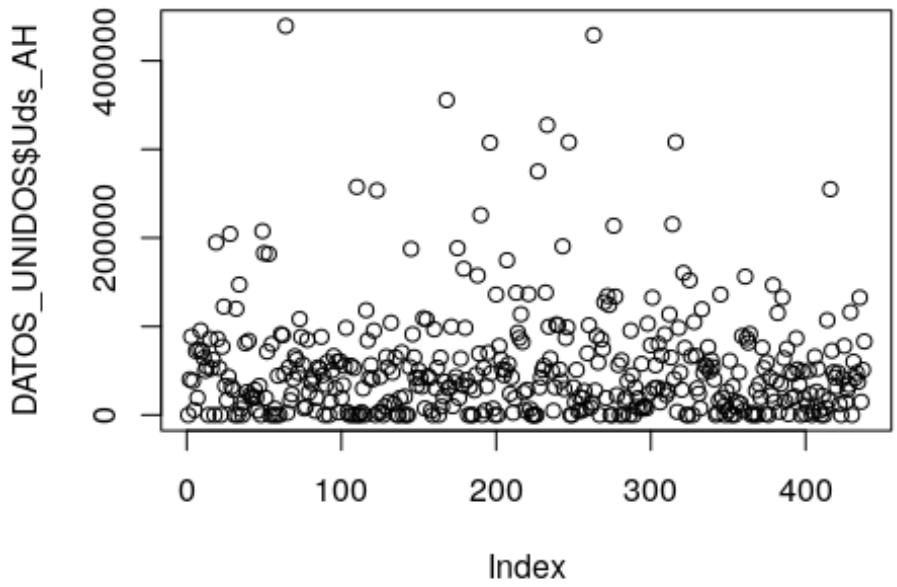
Podemos observar que los datos están completos y no hay valores ausentes.

3.2. Identificación y tratamiento de valores extremos.

Un primer paso para averiguar la distribución de los datos consiste en una simple visualización de los mismos. Así, de manera rápida, se puede ver si hay valores extremos e investigar posibles errores en los datos que puedan llevar a un sesgo en la interpretación de resultados.

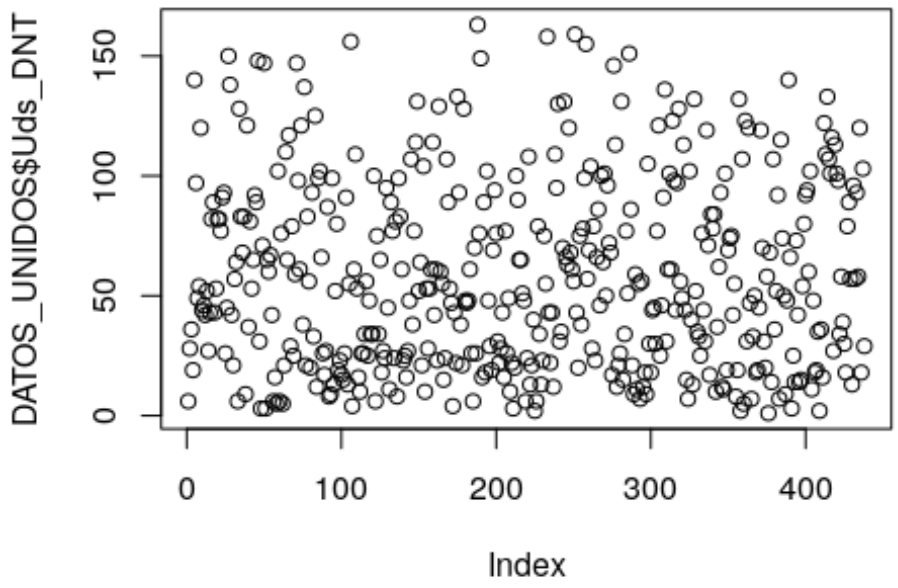
Variable **AH**

```
plot(DATOS_UNIDOS$Uds_AH)
```



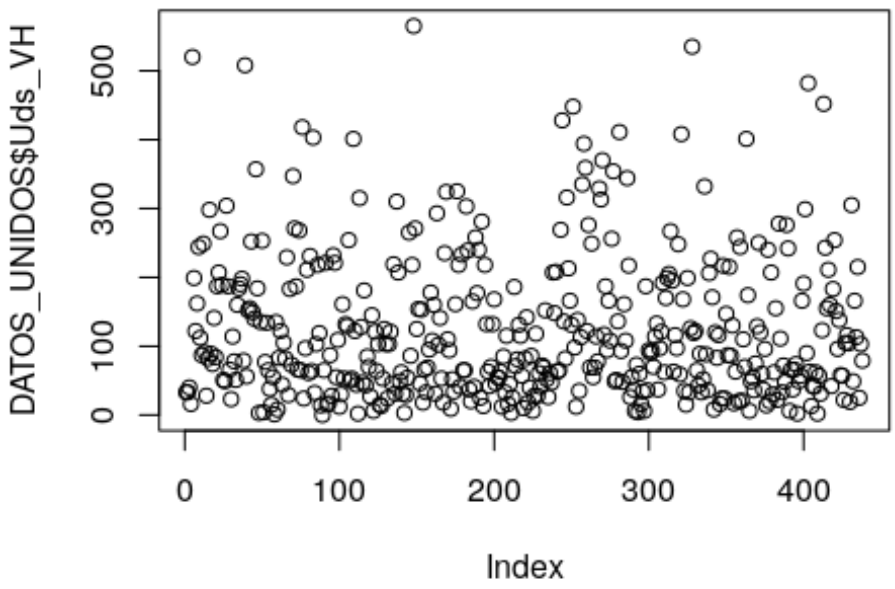
Variable **DNT**

```
plot(DATOS_UNIDOS$Uds_DNT)
```



Variable **VH**

```
plot(DATOS_UNIDOS$Uds_VH)
```



Si bien existen valores po encima de lo normal en algunas de las variables, no se puede tratar como un caso fuera de lo común pues son datos que tienen sentido y su proporción es lo suficientemente pequeña como para no desvirtuar los valores medios.

4. Análisis de los datos.

4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

En el caso que nos ocupa lo que queremos es estimar las ventas del resto de clientes fuera de la muestra. Por ello, para generar el modelo, no vamos a necesitar los nombres de los clientes ni la población, tan solo seleccionaremos las ventas de los distintos medicamentos así como las ventas totales. No será necesario preseleccionar nada pues a la correspondiente función de R podemos pasarle como parámetros sólo las variables que necesitemos justo en el momento de ejecutar el modelo.

Una vez construyamos nuestro modelo, introduciremos en él los valores de las variables correspondientes al resto de clientes que no forman parte de la muestra, con el fin de estimar sus ventas. De esta manera, también sería posible estimar la cuota de mercado que tiene nuestro laboratorio como diferencia de las ventas totales menos las ventas que el laboratorio realiza dividido entre el total del mercado.

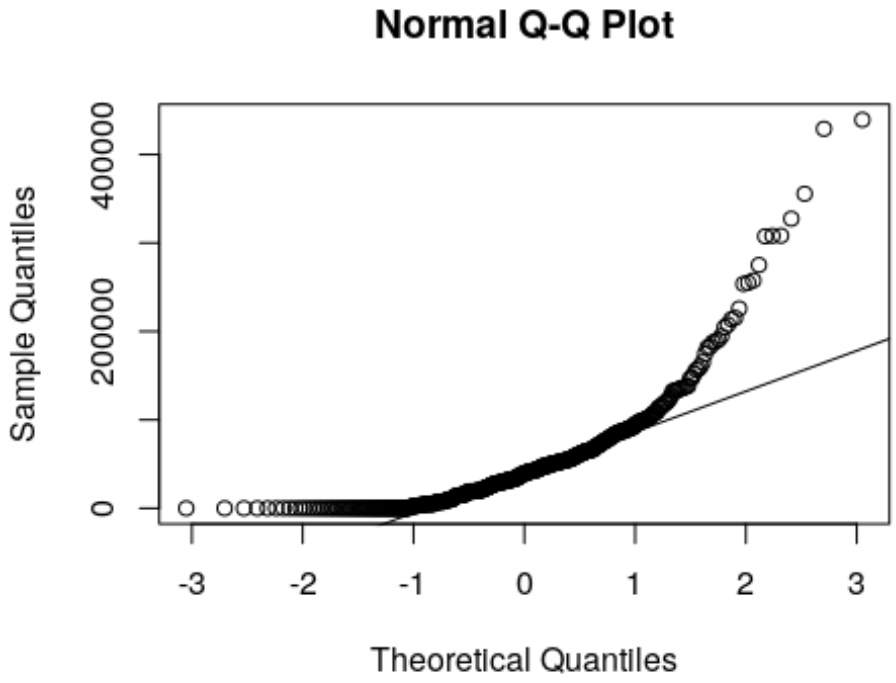
También es posible calcular el potencial o recorrido de cada cliente, de manera que si un cliente nos compra una proporción relativamente pequeña con respecto al total de sus ventas, sabemos que puede comprarnos muchos más productos. Con estos datos podemos segmentar el mercado con el objetivo de centrar nuestra estrategia de marketing y nuestros esfuerzos en aquellos clientes que nos sean más rentables y ganemos mayor cuota de mercado.

4.2. Comprobación de la normalidad y homogeneidad de la varianza.

Podemos ejectuar distintas pruebas para comprobar la normalidad de las variables. Por ejemplo, gráficamente tenemos:

Variable **AH**

```
# Comprobación de normalidad
qqnorm(DATOS_UNIDOS$Uds_AH)
qqline(DATOS_UNIDOS$Uds_AH)
```



Al parecer, esta variable no está normalmente distribuida, pues los puntos no quedan repartidos de forma, más o menos uniforme a cada lado de la línea. Al haber más de treina observaciones, podemos construir el estadístico de Shapiro-Wilk planteando las siguientes hipótesis:

Ho -> La variable sigue una distribución normal

H1 -> La variable no sigue una distribución normal

```
shapiro.test(DATOS_UNIDOS$Uds_AH)
```

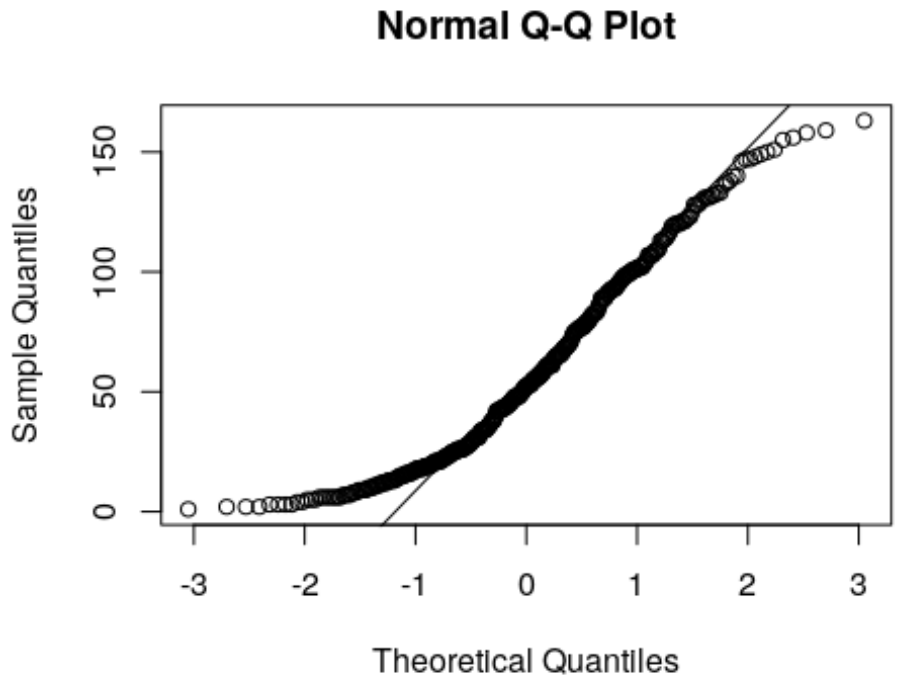
```
##
##  Shapiro-Wilk normality test
##
## data:  DATOS_UNIDOS$Uds_AH
## W = 0.75247, p-value < 2.2e-16
```

Al ser el p-value menor que 0,05 rechazo la hipótesis nula, por tanto, la variable no está distribuida normalmente tal y como se apreciaba de manera gráfica.

Ejecutando ahora estos test para el resto de variables tenemos:

Variable **DNT**

```
# Comprobación de normalidad
qqnorm(DATOS_UNIDOS$Uds_DNT)
qqline(DATOS_UNIDOS$Uds_DNT)
```



Ho -> La variable sigue una distribución normal

H1 -> La variable no sigue una distribución normal

```
shapiro.test(DATOS_UNIDOS$Uds_DNT)
```

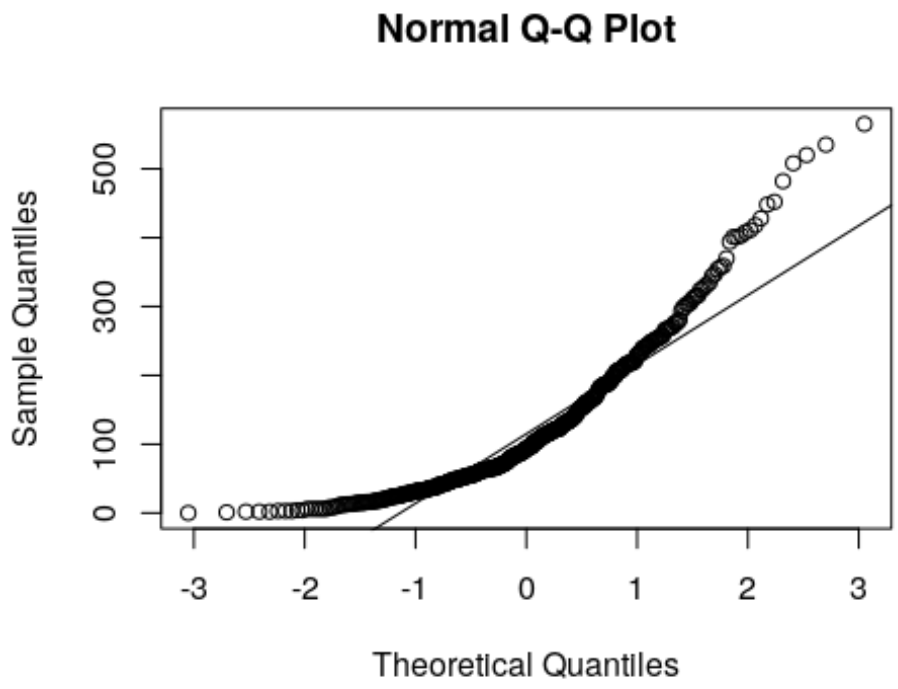
```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  DATOS_UNIDOS$Uds_DNT
## W = 0.9466, p-value = 1.834e-11
```

Rechazamos la hipótesis nula.

Variable **VH**

```
# Comprobación de normalidad
qqnorm(DATOS_UNIDOS$Uds_VH)
qqline(DATOS_UNIDOS$Uds_VH)
```



Ho -> La variable sigue una distribución normal

H1 -> La variable no sigue una distribución normal

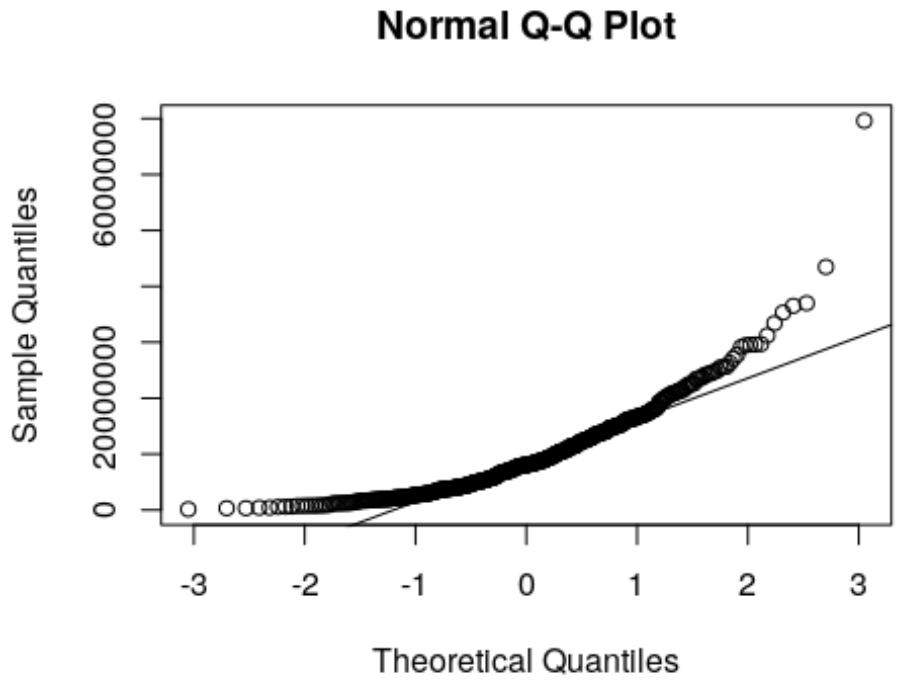
```
shapiro.test(DATOS_UNIDOS$Uds_VH)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  DATOS_UNIDOS$Uds_VH
## W = 0.87707, p-value < 2.2e-16
```

Rechazamos la hipótesis nula.

Variable **Resto_farmacos**

```r
# Comprobación de normalidad
qqnorm(DATOS_UNIDOS$Resto_Farmacos)
qqline(DATOS_UNIDOS$Resto_Farmacos)
```



Ho -> La variable sigue una distribución normal

H1 -> La variable no sigue una distribución normal

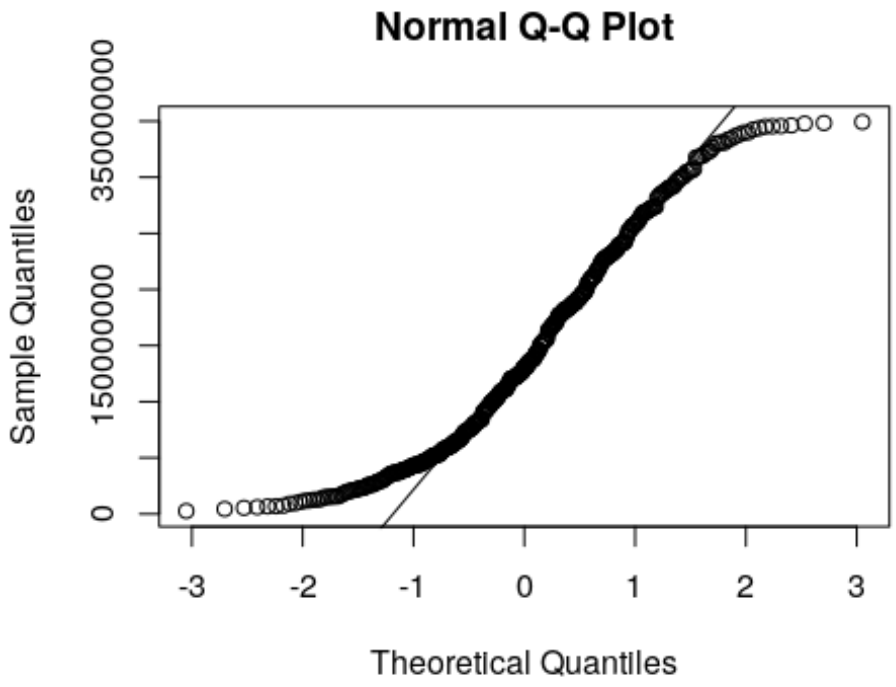
```r
shapiro.test(DATOS_UNIDOS$Resto_Farmacos)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  DATOS_UNIDOS$Resto_Farmacos
## W = 0.86123, p-value < 2.2e-16
```

Rechazamos la hipótesis nula.

Variable **TOTAL_VOL**

```r
# Comprobación de normalidad
qqnorm(DATOS_UNIDOS$TOTAL_VOL)
qqline(DATOS_UNIDOS$TOTAL_VOL)
```



Ho -> La variable sigue una distribución normal

H1 -> La variable no sigue una distribución normal

```
shapiro.test(DATOS_UNIDOS$TOTAL_VOL)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  DATOS_UNIDOS$TOTAL_VOL
## W = 0.94627, p-value = 1.661e-11
```

Rechazamos la hipótesis nula.

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc.

Podemos aplicar distintas pruebas sobre aspectos concretos de los datos. Por ejemplo, un resumen estadístico descriptivo como el siguiente:

```
# Ejecutamos algunos test básicos de estadística descriptiva
summary(DATOS_UNIDOS)
```

```
##         Cod_Cli       PROVINCIA       Uds_AH          Uds_DNT

##  anon_S10069:  1   TOLEDO  : 28   Min.   :     0   Min.   :
```

```
1.00
##  anon_S10181:  1   BADAJOZ : 23   1st Qu.:  9638   1st Qu.:
24.25
##  anon_S10199:  1   A CORUNA: 21   Median : 39300   Median :
52.00
##  anon_S10217:  1   JAEN    : 21   Mean   : 53888   Mean   :
58.09
##  anon_S103  :  1   MURCIA  : 21   3rd Qu.: 71538   3rd Qu.:
88.50
##  anon_S10361:  1   VALENCIA: 17   Max.   :439250
Max.   :163.00
##  (Other)    :432   (Other) :307

##      Uds_VH      Resto_Farmacos       TOTAL_VOL
##  Min.   :  0.0   Min.   :  123377   Min.   :  2541182
##  1st Qu.: 47.0   1st Qu.: 3875183   1st Qu.: 60015040
##  Median : 93.5   Median : 8050155   Median :130421221
##  Mean   :124.6   Mean   : 9799882   Mean   :145382223
##  3rd Qu.:183.0   3rd Qu.:13820033   3rd Qu.:218810932
##  Max.   :565.0   Max.   :69609887   Max.   :349199044
##
```

Si agrupamos por provincia tenemos:

```
# Ejecutamos un resumen agrupado por provincias

library(abind, pos=17)
library(e1071, pos=18)

numSummary(DATOS_UNIDOS[,c("Resto_Farmacos", "TOTAL_VOL",
                           "Uds_AH", "Uds_DNT", "Uds_VH"),
drop=FALSE],
           groups=DATOS_UNIDOS$PROVINCIA, statistics=c("mean",
"sd", "IQR",

"quantiles"), quantiles=c(0,.25,.5,.75,1))

##
## Variable: Resto_Farmacos
##                 mean       sd      IQR         0%      25%
50%
## A CORUNA     8358831  5462348  7614900   329565.1  3795692
7216498
## ALAVA        1461213       NA        0  1461212.6  1461213
1461213
## ALBACETE    11403746  9174223 12329800  1663768.9  4425171
8285430
## ALICANTE    10572745  6333000  7279962   771271.2  7136175
9571778
## ALMERIA      8384113  6173402  6278723   816907.5  4196250
```

```
8915250
## ASTURIAS     8466419  5294232  8208077  1716101.4  3801331
8195210
## AVILA       13670413 12789977 21471908   554814.0  1961004
12941556
## BADAJOZ      7443849  7583257  5421908   320548.6  3124717
6721314
## BARCELONA   12268504       NA        0 12268504.5 12268504
12268504
## BURGOS      11552856  7452299 10201430  2480426.7  5593463
12708138
## CACERES      8330926  5522619  8471455  2769335.4  3564202
7117253
## CADIZ       12535242 10292732  6293106  1922967.8  6303334
8794298
## CANTABRIA   11804999  8087826  8978405  3405805.2  5718904
7730863
## CASTELLON   12974643 12737972  9007106  3967536.4  8471089
12974643
## CIUDAD REAL  8397712  9241360  6212140  1525848.2  2483474
4688182
## CORDOBA      8727231  5632645  8854057  2352035.6  3846824
8218118
## CUENCA       6402422  4496022  7713442   730166.2  2685138
4598245
## GIRONA       4661383       NA        0  4661383.0  4661383
4661383
## GRANADA      9970194 10751826  6288006   903206.9  3918987
7536280
## GUADALAJARA 11968453 10931413 15677172   880954.7  3854572
11417863
## HUELVA      13430838  6688935  9310410  2871939.9  8107029
15600558
## HUESCA       9322764 10830066  4608316  1910571.5  3933815
6298348
## JAEN        13303864  7992552 11492463  1826696.6  7794377
12277097
## LA RIOJA     7958801  3528676  1448541  2628128.0  7450461
8377375
## LEON         7090331  6268295  7913429  1239684.5  2337485
5938384
## LLEIDA       7332586  5966426  5842159   770621.4  3116607
7605144
## LUGO        10684817 11356245 10451515  2797811.7  4176804
5555796
## MADRID      16896239  3203142  3178029 13487090.0 15422784
17358479
## MALAGA      11360261  6824799  4188923  5221498.3  7379164
9989801
## MURCIA      13107368 10349400 12376062  1492744.8  4963487
```

```
10137558
## NAVARRA     10600613  5861290  9416893  4448841.0  6159540
11135361
## OURENSE      5425178  3598683  1371826  1926342.0  3871309
4818729
## PALENCIA    16130298 20713994  4198216  1199714.6  8061416
10118451
## PONTEVEDRA   5031055  3949346  4076889   417818.2  2258886
4271384
## SALAMANCA    6679609  6140744  7273200   407935.4  2289379
5077668
## SEGOVIA      9586016  8764087  6161877  4468871.8  4520800
5617460
## SEVILLA      9475801  4435539  5106900  1727741.3  7888809
8248054
## SORIA        7615785  5398269  3817152  3798632.2  5707208
7615785
## TARRAGONA    6162264  4452483  4158544  2922255.3  3623725
4325194
## TERUEL       9056090  2369189  1514624  5804708.6  8503027
9464589
## TOLEDO      12559899  4990648  6718398  3538271.7  8395276
13498242
## VALENCIA    14152865  8750734 15818158  2265520.8  5092979
15082115
## VALLADOLID   8302254  6741057  7992168  1149542.2  3295783
7433513
## VIZCAYA      5904338       NA        0  5904337.6  5904338
5904338
## ZAMORA       3812069  4042538  2818692   123377.5  1295981
2809948
## ZARAGOZA     7671218  6972664  8516282   618628.2  2059518
6589180
##                  75%     100%  n
## A CORUNA    11410591 19928095 21
## ALAVA        1461213  1461213  1
## ALBACETE    16754971 29134496 11
## ALICANTE    14416137 21944269 11
## ALMERIA     10474973 18039142  6
## ASTURIAS    12009407 17212199 12
## AVILA       23432912 29461780  5
## BADAJOZ      8546625 36989175 23
## BARCELONA   12268504 12268504  1
## BURGOS      15794893 21374717  6
## CACERES     12035658 20580348 16
## CADIZ       12596440 33393734  9
## CANTABRIA   14697309 25502762 14
## CASTELLON   17478196 21981749  2
## CIUDAD REAL  8695614 29570778 11
## CORDOBA     12700881 19296481 13
```

```
## CUENCA      10398579 13202539 15
## GIRONA       4661383  4661383  1
## GRANADA     10206993 43474994 14
## GUADALAJARA 19531744 24157130  4
## HUELVA      17417439 21441741  8
## HUESCA       8542131 35320123  8
## JAEN        19286840 29671404 21
## LA RIOJA     8899002 12439040  5
## LEON        10250914 24572439 15
## LLEIDA       8958766 16211794  5
## LUGO        14628319 23700842  3
## MADRID      18600813 19843147  3
## MALAGA      11568087 24357333  6
## MURCIA      17339549 36513668 21
## NAVARRA     15576434 15682890  4
## OURENSE      5243134 14570743  9
## PALENCIA    12259632 69609887  9
## PONTEVEDRA   6335775 13595272  9
## SALAMANCA    9562579 22457535 13
## SEGOVIA     10682677 22640273  4
## SEVILLA     12995709 16157660 16
## SORIA        9524361 11432937  2
## TARRAGONA    7782269 11239343  3
## TERUEL      10017652 11490475  4
## TOLEDO      15113674 25623745 28
## VALENCIA    20911138 31214473 17
## VALLADOLID  11287951 19307655  6
## VIZCAYA      5904338  5904338  1
## ZAMORA       4114673 14893071 12
## ZARAGOZA    10575801 21012939 10
##
## Variable: TOTAL_VOL
##                  mean        sd       IQR        0%       25%
50%
## A CORUNA    132491031  75119010 116952313   4348833  67282627
143813624
## ALAVA        29575163        NA         0  29575163  29575163
29575163
## ALBACETE    166678399 114201212 202066612  15394158  73058665
184520082
## ALICANTE    163744462  94351375 141590466   9523747 102040824
156382834
## ALMERIA     108084912  69077380  76930439  12623842  64157713
117474574
## ASTURIAS    132463698  84499087 111169882  22967047  64312764
121294804
## AVILA       195611944 164921276 311595039   9085677  33502476
242517058
## BADAJOZ     117561166  80246968 135484179   5383498  45175996
117872396
```

```
## BARCELONA   111325237        NA         0 111325237 111325237
111325237
## BURGOS      178652557 122976900 165850523  22633714  86273833
193666037
## CACERES     134343068  90261570 122350999  39200244  60762892
120160377
## CADIZ       171400902  96703687 131405761  38523598  96340422
153725125
## CANTABRIA   173174771  92714980 120888016  61562939 105668287
130632642
## CASTELLON   181426918 204917568 144898602  36528316 108977617
181426918
## CIUDAD REAL 109534733  99651925  82716675  29099655  44575583
71945514
## CORDOBA     142962711  85571053 118905378  26727303  81680613
125636383
## CUENCA      106787524  88369938 140460509   6649916  36113576
58356462
## GIRONA       76849066        NA         0  76849066  76849066
76849066
## GRANADA     129147131  87396649 104881857  14829808  67183036
113422769
## GUADALAJARA 159377378 122692044 160342657  12721802  84356807
169678894
## HUELVA      219729553  85619643 130192698  64164005 162005529
237791755
## HUESCA      133249299  83720121 123632007  28622474  71505339
122837295
## JAEN        175903280  95446932 174436496  39249359 110787540
166524555
## LA RIOJA    127312871  80039001  54169972  21109552  91744653
138156753
## LEON         96829893  85819372  79083010  15637635  39381911
58665819
## LLEIDA      119487082  93355209 103230579  11701434  65729190
99280664
## LUGO        135390452 117999648 111137757  47146041  68374901
89603761
## MADRID      280583211  29745972  26458644 246276786 271277780
296278773
## MALAGA      181715330  86764567 130253089  92386080 111517626
170278858
## MURCIA      158540615 104604103 174304884  25964498  66400015
109734498
## NAVARRA     159045874  70767272  41905663  56077944 149180272
181220334
## OURENSE      79162233  48983960  50571701  35893417  40918081
81057595
## PALENCIA    162873695  81572762  81047301  13096669 130599102
167863499
```

```
## PONTEVEDRA   97651883  72904374  62068679   7291732  58900518
77156074
## SALAMANCA   104275308  98364309 107330389   5956947  35489371
73796224
## SEGOVIA     140461984  87948928 105089079  73028556  74905471
114438039
## SEVILLA     168778621  96719676  79180454  22918902 113976549
157411197
## SORIA       104777694  64125961  45343902  59433793  82105744
104777694
## TARRAGONA    82733593  67269950  58313657  43782819  43895324
44007829
## TERUEL      159447791  40568570  29772619  99329009 153011049
176346926
## TOLEDO      202537122  69727927  95571575  52722420 151884634
208906454
## VALENCIA    219880312 115830159 255799362  36344748  74891013
257758452
## VALLADOLID  106908116  84086035 105486497  23319659  44510554
89653107
## VIZCAYA     137496238        NA         0 137496238 137496238
137496238
## ZAMORA       64323158  69625793  44587189   2541182  26054112
37240910
## ZARAGOZA    122212057 118211753 111178344   6844511  22431821
114485024
##                    75%      100%  n
## A CORUNA    184234941 266000762 21
## ALAVA        29575163  29575163  1
## ALBACETE    275125277 323450171 11
## ALICANTE    243631290 286844499 11
## ALMERIA     141088152 205591711  6
## ASTURIAS    175482646 301554456 12
## AVILA       345097515 347856995  5
## BADAJOZ     180660176 271394998 23
## BARCELONA   111325237 111325237  1
## BURGOS      252124357 339862659  6
## CACERES     183113891 345336732 16
## CADIZ       227746183 330809260  9
## CANTABRIA   226556303 339725648 14
## CASTELLON   253876219 326325521  2
## CIUDAD REAL 127292259 322614100 11
## CORDOBA     200585991 317688294 13
## CUENCA      176574085 257034407 15
## GIRONA       76849066  76849066  1
## GRANADA     172064893 317820340 14
## GUADALAJARA 244699464 285429919  4
## HUELVA      292198227 305571614  8
## HUESCA      195137346 253242966  8
## JAEN        285224036 346208196 21
```

```
## LA RIOJA    145914625 239638774  5
## LEON        118464920 321077496 15
## LLEIDA      168959769 251764355  5
## LUGO        179512658 269421555  3
## MADRID      297736424 299194075  3
## MALAGA      241770714 299816305  6
## MURCIA      240704900 343627200 21
## NAVARRA     191085936 217664886  4
## OURENSE      91489783 193450314  9
## PALENCIA    211646402 292855269  9
## PONTEVEDRA  120969197 240279632  9
## SALAMANCA   142819759 336694202 13
## SEGOVIA     179994551 259943300  4
## SEVILLA     193157004 349199044 16
## SORIA       127449645 150121596  2
## TARRAGONA   102208980 160410132  3
## TERUEL      182783668 185768306  4
## TOLEDO      247456209 330737286 28
## VALENCIA    330690375 348340966 17
## VALLADOLID  149997050 239248088  6
## VIZCAYA     137496238 137496238  1
## ZAMORA       70641300 259276634 12
## ZARAGOZA    133610165 342601893 10
##
## Variable: Uds_AH
##                  mean        sd       IQR    0%       25%
50%
## A CORUNA     57851.19  62810.96  90500.00     0   4800.00
40200.0
## ALAVA        22500.00        NA      0.00 22500  22500.00
22500.0
## ALBACETE     51372.73  61344.81  71050.00     0      0.00
50100.0
## ALICANTE     24859.09  21992.58  43000.00     0      0.00
25950.0
## ALMERIA      27450.00  25902.70  39637.50     0   4800.00
26475.0
## ASTURIAS     59350.00  44260.77  42637.50  4800  29175.00
58400.0
## AVILA        41004.00  41639.71  57600.00  2500   4800.00
34020.0
## BADAJOZ      31565.22  49706.99  44700.00     0      0.00
7500.0
## BARCELONA    86250.00        NA      0.00 86250  86250.00
86250.0
## BURGOS       87745.83  74190.88  94906.25 15000  36750.00
61862.5
## CACERES      59159.38  59793.78  37650.00     0  27525.00
47500.0
## CADIZ        55583.33  51443.39  62850.00     0  12000.00
```

```
55200.0
## CANTABRIA   137189.29 129911.35 120168.75 28500  51525.00
77050.0
## CASTELLON    53550.00  75731.14  53550.00     0  26775.00
53550.0
## CIUDAD REAL  31859.09  75950.14  19875.00     0      0.00
4800.0
## CORDOBA      49305.77  43854.89  61350.00     0  14400.00
41700.0
## CUENCA       32373.33  28432.97  36500.00     0  11250.00
25500.0
## GIRONA       52125.00        NA      0.00 52125  52125.00
52125.0
## GRANADA      26057.14  33303.41  29381.25     0      0.00
19950.0
## GUADALAJARA  85987.50  77794.28  66337.50     0  48937.50
78225.0
## HUELVA       38265.62  18430.47  14193.75 10600  29981.25
37600.0
## HUESCA       44393.75  43636.54  57412.50     0   6825.00
38000.0
## JAEN         59142.86  76959.40  45000.00     0  19800.00
47700.0
## LA RIOJA     31405.00  29252.54  50225.00     0   2500.00
37000.0
## LEON         37640.00  34204.18  47000.00     0  12150.00
31500.0
## LLEIDA       41470.00  34358.40  11850.00 10000  28350.00
29000.0
## LUGO         73883.33  28095.66  27575.00 43200  61650.00
80100.0
## MADRID      131450.00  69611.37  67275.00 53850 102975.00
152100.0
## MALAGA       66175.00  61740.06  74612.50  1250  17062.50
64250.0
## MURCIA       59347.62  79536.62  54900.00     0  14400.00
32100.0
## NAVARRA      47662.50  62506.92  64162.50     0   5400.00
27300.0
## OURENSE      46696.67  27191.87  36450.00     0  26700.00
50850.0
## PALENCIA     68547.22  45487.42  67575.00  4800  25425.00
82350.0
## PONTEVEDRA   34316.67  47519.96  32400.00     0      0.00
16200.0
## SALAMANCA    50075.38  70182.21  32620.00     0  16380.00
40900.0
## SEGOVIA      77025.00  37944.47  57675.00 38400  48075.00
76800.0
## SEVILLA      68454.69  79990.67  56850.00     0  25650.00
```

```
44725.0
## SORIA        97350.00 137673.69  97350.00     0  48675.00
97350.0
## TARRAGONA    47200.00  28506.84  25800.00 14400  37800.00
61200.0
## TERUEL       11000.00  10152.26  15850.00     0   3562.50
11975.0
## TOLEDO       72712.32  83978.21  74306.25     0  19350.00
53800.0
## VALENCIA     61545.59  51350.61  46750.00     0  34950.00
52100.0
## VALLADOLID   36720.83  28962.21  38400.00     0  18000.00
33300.0
## VIZCAYA      31000.00        NA      0.00 31000  31000.00
31000.0
## ZAMORA       27404.17  30295.65  30187.50     0   7312.50
17475.0
## ZARAGOZA     66855.00  66782.71  58087.50     0  21900.00
56100.0
##                   75%   100%  n
## A CORUNA     95300.00 253700 21
## ALAVA        22500.00  22500  1
## ALBACETE     71050.00 181800 11
## ALICANTE     43000.00  54100 11
## ALMERIA      44437.50  63750  6
## ASTURIAS     71812.50 174800 12
## AVILA        62400.00 101300  5
## BADAJOZ      44700.00 207650 23
## BARCELONA    86250.00  86250  1
## BURGOS      131656.25 204450  6
## CACERES      65175.00 255000 16
## CADIZ        74850.00 160500  9
## CANTABRIA   171693.75 439250 14
## CASTELLON    80325.00 107100  2
## CIUDAD REAL  19875.00 257600 11
## CORDOBA      75750.00 132700 13
## CUENCA       47750.00 104250 15
## GIRONA       52125.00  52125  1
## GRANADA      29381.25 113700 14
## GUADALAJARA 115275.00 187500  4
## HUELVA       44175.00  72300  8
## HUESCA       64237.50 109200  8
## JAEN         64800.00 355600 21
## LA RIOJA     52725.00  64800  5
## LEON         59150.00 115200 15
## LLEIDA       40200.00  99800  5
## LUGO         89225.00  98350  3
## MADRID      170250.00 188400  3
## MALAGA       91675.00 165150  6
## MURCIA       69300.00 307500 21
```

```
## NAVARRA      69562.50 136050  4
## OURENSE      63150.00  83820  9
## PALENCIA     93000.00 138000  9
## PONTEVEDRA   32400.00 136550  9
## SALAMANCA    49000.00 275100 13
## SEGOVIA     105750.00 116100  4
## SEVILLA      82500.00 308100 16
## SORIA       146025.00 194700  2
## TARRAGONA    63600.00  66000  3
## TERUEL       19412.50  20050  4
## TOLEDO       93656.25 429050 28
## VALENCIA     81700.00 213700 17
## VALLADOLID   56400.00  76725  6
## VIZCAYA      31000.00  31000  1
## ZAMORA       37500.00 103100 12
## ZARAGOZA     79987.50 215400 10
##
## Variable: Uds_DNT
##                  mean       sd    IQR 0%    25%   50%    75%
100%  n
## A CORUNA     51.95238 30.34547  49.00  2  26.00  61.0  75.00
100 21
## ALAVA        14.00000       NA   0.00 14  14.00  14.0  14.00
14  1
## ALBACETE     68.90909 49.93887  91.00  6  27.50  60.0 118.50
140 11
## ALICANTE     65.72727 38.34864  58.00  4  42.00  61.0 100.00
121 11
## ALMERIA      43.66667 30.23023  28.75  6  25.25  43.0  54.00
93  6
## ASTURIAS     54.91667 33.91824  46.00 10  25.00  53.5  71.00
116 12
## AVILA        76.20000 66.23972 118.00  3  12.00  89.0 130.00
147  5
## BADAJOZ      46.08696 32.75933  48.00  2  18.50  47.0  66.50
122 23
## BARCELONA    43.00000       NA   0.00 43  43.00  43.0  43.00
43  1
## BURGOS       78.50000 53.80242  83.50  9  35.75  88.0 119.25
138  6
## CACERES      51.00000 34.72175  47.50 15  22.00  44.0  69.50
132 16
## CADIZ        68.88889 46.18020  57.00 13  34.00  55.0  91.00
156  9
## CANTABRIA    73.42857 43.52137  63.00 22  43.25  52.0 106.25
158 14
## CASTELLON    73.00000 84.85281  60.00 13  43.00  73.0 103.00
133  2
## CIUDAD REAL  43.45455 39.65189  38.00 10  17.50  26.0  55.50
125 11
```

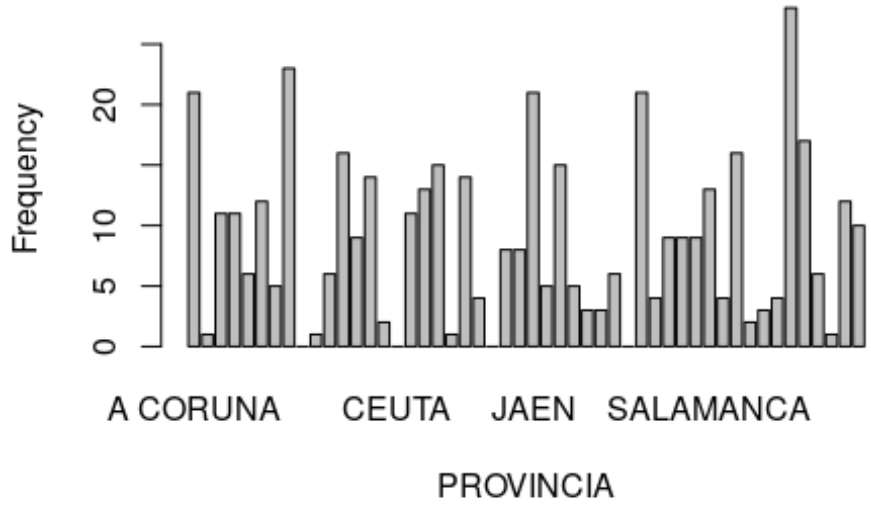
```
## CORDOBA      56.92308 35.68020  49.00 12  34.00  48.0  83.00
137 13
## CUENCA       41.00000 32.15365  53.50  3  13.50  27.0  67.00
93 15
## GIRONA       35.00000       NA   0.00 35  35.00  35.0  35.00
35  1
## GRANADA      51.71429 38.26714  35.50  6  25.50  41.5  61.00
148 14
## GUADALAJARA  56.75000 44.58980  54.75  5  29.75  57.5  84.50
107  4
## HUELVA       84.37500 36.04338  44.00 21  61.00  95.5 105.00
131  8
## HUESCA       53.62500 38.31053  46.50 10  25.25  47.5  71.75
114  8
## JAEN         66.90476 36.41141  57.00 13  42.00  61.0  99.00
129 21
## LA RIOJA     52.20000 27.36238  22.00 10  43.00  61.0  65.00
82  5
## LEON         43.13333 40.10498  37.50  7  15.00  25.0  52.50
150 15
## LLEIDA       49.80000 35.81480  51.00  4  28.00  47.0  79.00
91  5
## LUGO         64.00000 56.32051  53.00 22  32.00  42.0  85.00
128  3
## MADRID      111.33333 18.82374  17.00 99 100.50 102.0 117.50
133  3
## MALAGA       74.50000 36.44036  50.75 38  45.00  70.0  95.75
128  6
## MURCIA       63.38095 44.57743  63.00 11  26.00  48.0  89.00
163 21
## NAVARRA      64.50000 32.10919  24.00 19  56.50  72.5  80.50
94  4
## OURENSE      31.88889 19.48361  18.00 14  19.00  28.0  37.00
77  9
## PALENCIA     60.55556 30.35256  41.00  6  48.00  65.0  89.00
100  9
## PONTEVEDRA   38.77778 31.28409  21.00  3  21.00  36.0  42.00
108  9
## SALAMANCA    39.53846 37.17216  41.00  2  13.00  29.0  54.00
132 13
## SEGOVIA      54.50000 29.32007  32.50 31  34.00  46.0  66.50
95  4
## SEVILLA      66.81250 40.80885  34.75  9  40.50  62.0  75.25
159 16
## SORIA        40.00000 18.38478  13.00 27  33.50  40.0  46.50
53  2
## TARRAGONA    37.00000 32.04684  28.00 18  18.50  19.0  46.50
74  3
## TERUEL       64.50000 18.26655  16.00 38  59.75  71.0  75.75
78  4
```

```
## TOLEDO       81.35714 30.62463  37.50 23  62.75  80.0 100.25
155 28
## VALENCIA     89.00000 47.60121  89.00 12  31.00 102.0 120.00
151 17
## VALLADOLID   43.00000 34.54273  47.25  9  17.25  33.0  64.50
96  6
## VIZCAYA      59.00000       NA   0.00 59  59.00  59.0  59.00
59  1
## ZAMORA       26.16667 28.53653  16.00  1  11.75  16.5  27.75
105 12
## ZARAGOZA     47.20000 44.06762  44.50  3   9.00  44.5  53.50
123 10
##
## Variable: Uds_VH
##                  mean        sd    IQR  0%    25%   50%    75%
100%  n
## A CORUNA     84.52381  62.78266  94.00   2  27.00  78.0 121.00
215 21
## ALAVA        20.00000        NA   0.00  20  20.00  20.0  20.00
20  1
## ALBACETE    183.00000 143.80125 148.50  15  90.50 187.0 239.00
520 11
## ALICANTE    168.63636 145.24894 183.00  49  57.50 135.0 240.50
508 11
## ALMERIA     192.50000 131.30537 204.00  20  85.25 233.0 289.25
325  6
## ASTURIAS     92.08333  68.38588  90.00  16  33.00  86.5 123.00
244 12
## AVILA       129.00000 110.87606 161.00   3  47.00 116.0 208.00
271  5
## BADAJOZ     103.82609  67.34015  76.50   6  59.00 108.0 135.50
252 23
## BARCELONA    82.00000        NA   0.00  82  82.00  82.0  82.00
82  1
## BURGOS      125.50000  81.05245 110.25  17  70.00 128.5 180.25
231  6
## CACERES     133.37500 133.37060 164.75  12  48.75  99.5 213.50
535 16
## CADIZ       161.33333 111.95981  75.00  48  96.00 132.0 171.00
408  9
## CANTABRIA   104.07143  69.70318  80.75  26  59.50  67.5 140.25
253 14
## CASTELLON   131.00000 158.39192 112.00  19  75.00 131.0 187.00
243  2
## CIUDAD REAL 151.81818 149.30694 178.00   2  43.50 113.0 221.50
403 11
## CORDOBA     131.38462 109.80023 134.00   0  47.00 111.0 181.00
418 13
## CUENCA       83.40000  68.94180  78.00   4  34.00  55.0 112.00
219 15
```

```
## GIRONA       62.00000        NA   0.00  62  62.00  62.0  62.00
62  1
## GRANADA     108.07143  96.46240 108.50   3  49.50  63.5 158.00
357 14
## GUADALAJARA 142.50000 120.99725 165.00   1  64.75 152.0 229.75
265  4
## HUELVA      267.62500 178.27101 198.00  46 146.75 235.0 344.75
565  8
## HUESCA       80.62500  57.98507  74.25  18  35.25  66.0 109.50
178  8
## JAEN        142.19048 102.55273 154.00  18  53.00 108.0 207.00
332 21
## LA RIOJA    129.00000 119.49477 153.00   8  52.00  82.0 205.00
298  5
## LEON         73.26667  78.65973  78.50   6  22.00  33.0 100.50
304 15
## LLEIDA       72.40000  48.63949  72.00   9  33.00  94.0 105.00
121  5
## LUGO         56.33333  35.27511  31.50  34  36.00  38.0  67.50
97  3
## MADRID      154.33333  48.34598  48.00 103 132.00 161.0 180.00
199  3
## MALAGA      130.33333  68.26322  95.50  65  76.75 114.0 172.25
233  6
## MURCIA      139.19048  87.64794 155.00  13  63.00 155.0 218.00
281 21
## NAVARRA     103.00000  57.37595  78.00  42  63.00 101.0 141.00
168  4
## OURENSE      48.00000  21.62175  19.00  12  37.00  54.0  56.00
85  9
## PALENCIA     94.55556  53.22854  71.00  10  60.00  81.0 131.00
186  9
## PONTEVEDRA   64.44444  42.90137  46.00   3  40.00  82.0  86.00
142  9
## SALAMANCA    66.84615  63.84075  43.00   6  29.00  57.0  72.00
258 13
## SEGOVIA      95.00000  40.05829  50.50  63  64.50  84.5 115.00
148  4
## SEVILLA     182.43750 129.37901 186.25  44  81.25 133.5 267.50
448 16
## SORIA        84.50000  79.90307  56.50  28  56.25  84.5 112.75
141  2
## TARRAGONA    64.00000  22.51666  22.50  42  52.50  63.0  75.00
87  3
## TERUEL       87.75000  46.41390  61.75  35  57.50  89.0 119.25
138  4
## TOLEDO      209.67857 115.75615 201.50  55 115.50 189.5 317.00
452 28
## VALENCIA    204.29412 125.79237 199.00  49  79.00 187.0 278.00
411 17
```

```
## VALLADOLID   92.33333 106.47003  44.25  26  35.50  52.0  79.75
305  6
## VIZCAYA      37.00000        NA   0.00  37  37.00  37.0  37.00
37  1
## ZAMORA       39.33333  49.77099  25.50   2  12.00  34.5  37.50
187 12
## ZARAGOZA     94.80000  78.85965  83.00   6  40.25  91.5 123.25
267 10
```

```
with(DATOS_UNIDOS, Barplot(PROVINCIA, xlab="PROVINCIA",
ylab="Frequency"))
```



Para ver la correlación existente entre las distintas variables podemos ejecutar un test de correlaciones.

```
cor(DATOS_UNIDOS[,c("Resto_Farmacos","TOTAL_VOL","Uds_AH","Uds_DN
T",
  "Uds_VH")], use="complete")
```

```
##                Resto_Farmacos TOTAL_VOL    Uds_AH   Uds_DNT
Uds_VH
## Resto_Farmacos      1.0000000 0.8234388 0.3023833 0.8211733
0.6679052
## TOTAL_VOL           0.8234388 1.0000000 0.3855773 0.9803340
0.7781667
## Uds_AH              0.3023833 0.3855773 1.0000000 0.3852610
```

```
0.2002995
## Uds_DNT             0.8211733 0.9803340 0.3852610 1.0000000
0.7705239
## Uds_VH              0.6679052 0.7781667 0.2002995 0.7705239
1.0000000
```

Las variables con coeficiente superior a 0,80 están fuertemente correlacionadas.

Dado que queremos generar un modelo predictivo para estimar las ventas, podemos realizar una regresión lineal. Para ello planteamos la siguiente regresión multivariante:

$$Ventas_i = \beta_1 + \beta_2 UdsAH + \beta_3 UdsDNT + \beta_4 UdsVH + \beta_5 Resto$$

Ejecutamos en modelo en R con el siguiente comando:

```
MODELO <- lm(TOTAL_VOL~Uds_AH + Uds_DNT + Uds_VH +
Resto_Farmacos, data = DATOS_UNIDOS)
```

La regresión lineal multivariante tiene 5 supuestos:

- Linealidad
- Independencia
- Homocedasticidad
- Normalidad
- No-Colinealidad

La independencia, homocedasticidad y normalidad están asociados al comportamiento de los residuos, que son los errores que comete el modelo. A efectos prácticos, la homocedasticidad y la normalidad son DESEABLES pero no OBLIGATORIAS, ya que podemos asumir que los errores que cometamos "se comportan de una determinada manera".

En este momento, ya tenemos el modelo de estimación creado. Ahora hay que comprobar si es bueno y además si se cumplen los supuestos que acabamos de enunciar.

5. Representación de los resultados a partir de tablas y gráficas.

Una vez generado el modelo vamos a ir desgranando los resultados y comprobando los supuestos.

- Coeficientes del modelo

```
coefficients(MODELO)
```

```
##     (Intercept)          Uds_AH          Uds_DNT
Uds_VH
```

```
## 6328724.2778273      24.8554394 2159021.5282567
50037.4370553
##  Resto_Farmacos
##       0.6180379
```

Por tanto el modelo quedaría como sigue:

$$Ventas_i = 6328724.277 + 24.8554394\,\beta_2 + 2159021.5282567\,\beta_3 + 50037.4370553\,\beta_4 + 0.6180379\,\beta_5$$

Con la siguiente instrucción podemos ver el ajuste del modelo y la validez de los coeficientes y la constante:

```
summary(MODELO)
```

```
##
## Call:
## lm(formula = TOTAL_VOL ~ Uds_AH + Uds_DNT + Uds_VH +
Resto_Farmacos,
##     data = DATOS_UNIDOS)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -54013255  -9501017  -2928854  10204964  65791858
##
## Coefficients:
##                    Estimate   Std. Error t value  Pr(>|t|)
## (Intercept)    6328724.2778 1591871.0857   3.976 0.0000822 ***
## Uds_AH              24.8554      15.2991   1.625  0.104969
## Uds_DNT        2159021.5283   48061.4009  44.922   < 2e-16 ***
## Uds_VH           50037.4371   13225.1621   3.784  0.000176 ***
## Resto_Farmacos       0.6180       0.1952   3.166  0.001657 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18330000 on 433 degrees of freedom
## Multiple R-squared:  0.9634, Adjusted R-squared:  0.9631
## F-statistic:  2849 on 4 and 433 DF,  p-value: < 2.2e-16
```

El valor de R cuadrado corregida nos indica si la recta de regresión se ha ajustado bien a la nube de puntos. Este valor siempre oscila entre 0 y 1, cuanto más cercano a 1 esté R cuadrado mejor será nuestro modelo ya que la nube de puntos se ha ajustado perfectamente. Por el contrario, un R cuadrado próximo a cero indicará que las variables independientes no están explicando a la variable dependiente y por tanto, nuestro modelo tendrá una capacidad predictiva nula. En la práctica, un modelo se considera útil con un R cuadrado a partir de 0.7, por debajo de esta cifra el modelo comienza a ser muy mediocre. En nuestro caso, tenemos un valor de 0.9631, lo cual nos indicaría que se trata de un buen ajuste.

Para determinar la bondad del modelo no basta con un buen R cuadrado sino que tenemos que fijarnos en los valores p, los cuales siempre deben ser menores que

0.05 para que los coeficientes y la constante sean significativos, es decir, que aseguren que el modelo es consiste y realiza buenas estimaciones.

Para ello, debemos analizar la significación de cada variable por separado y al final del proceso analizaremos la significación conjunta.

A continuación, realizamos el análisis de la significación individual de cada variable, estableciendo la hipótesis nula y la alternativa:

Significación individual

$$\left.\begin{array}{l} H_0 : \beta_2 = 0 \\ H_1 : \beta_2 \neq 0 \end{array}\right\} \qquad \alpha = 0{,}05$$

Comparamos Valor p con α
Si valor p < α rechazamos H_0

0,1049689766 > 0,05 aceptamos H_0, por lo tanto, la variable Uds_AH **NO** es una buena variable explicativa de las ventas o es no significativa.

No podríamos garantizar que nuestro modelo, a pesar de tener un R cuadrado muy alto, va a realizar estimaciones consistentes. En estos casos lo que hacemos es eliminar la variable que tiene un p-valor superior a 0.05 y volvemos a construir el modelo.

El nuevo modelo eliminando esta variable es:

$$Ventas_i = \beta_1 + \beta_2 UdsDNT + \beta_3 UdsVH + \beta_4 Resto$$

Ejecutando en R tenemos:

```
MODELO2 <- lm(TOTAL_VOL~Uds_DNT + Uds_VH + Resto_Farmacos, data =
DATOS_UNIDOS)
```

Los coeficientes obtenidos son:

```
coefficients(MODELO2)
```

```
##     (Intercept)         Uds_DNT          Uds_VH
Resto_Farmacos
## 6790675.4776084 2182195.3409091   46548.1348928
0.6145876
```

Analizando los valores estadísticos asociados al modelo tenemos:

```
summary(MODELO2)
```

```
##
## Call:
## lm(formula = TOTAL_VOL ~ Uds_DNT + Uds_VH + Resto_Farmacos,
data = DATOS_UNIDOS)
##
```

```
## Residuals:
##       Min        1Q    Median        3Q       Max
## -55272123  -9638341  -3087770  10043680  65632450
##
## Coefficients:
##                    Estimate   Std. Error t value  Pr(>|t|)
## (Intercept)    6790675.4776 1569225.8943   4.327 0.0000187 ***
## Uds_DNT        2182195.3409   45982.5194  47.457   < 2e-16 ***
## Uds_VH           46548.1349   13074.2247   3.560  0.000411 ***
## Resto_Farmacos       0.6146       0.1956   3.142  0.001792 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18370000 on 434 degrees of freedom
## Multiple R-squared:  0.9632, Adjusted R-squared:  0.9629
## F-statistic:  3783 on 3 and 434 DF,  p-value: < 2.2e-16
```

Pasamos a

Significación individual

$$\left.\begin{array}{l} H_0 : \beta_2 = 0 \\ H_1 : \beta_2 \neq 0 \end{array}\right\} \qquad \alpha = 0{,}05$$

2e-16 < 0,05 rechazamos H_0, por lo tanto, la variable Uds_DNT es una buena variable explicativa de las ventas o es significativa.

$$\left.\begin{array}{l} H_0 : \beta_3 = 0 \\ H_1 : \beta_3 \neq 0 \end{array}\right\} \qquad \alpha = 0{,}05$$

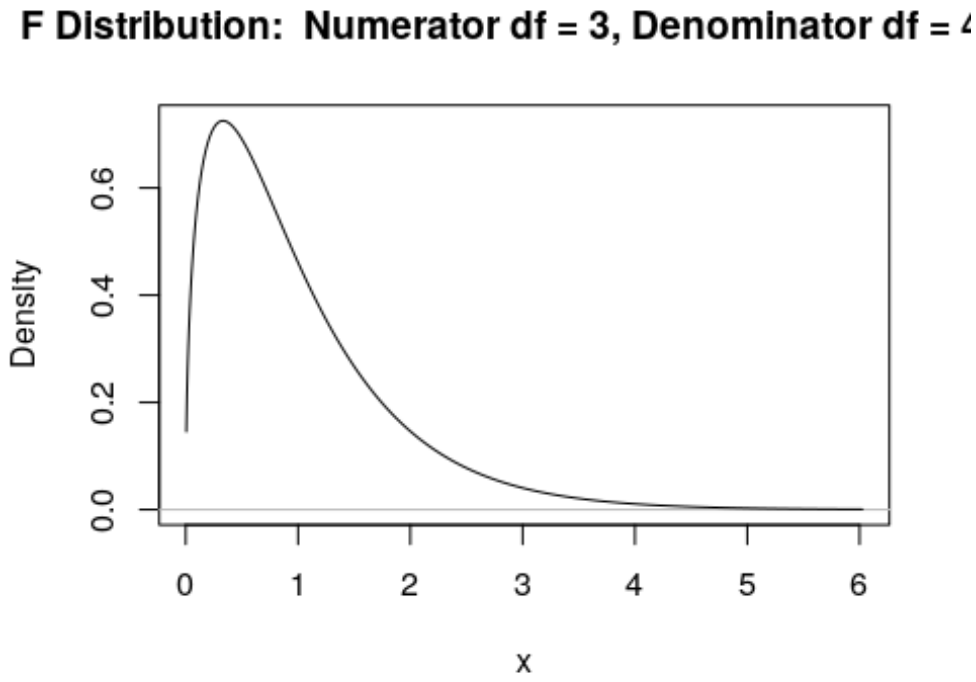
0,000411 < 0,05 rechazamos H_0, por lo tanto, la variable Uds_VH es una buena variable explicativa de las ventas o es significativa.

$$\left.\begin{array}{l} H_0 : \beta_4 = 0 \\ H_1 : \beta_4 \neq 0 \end{array}\right\} \qquad \alpha = 0{,}05$$

0,001792 < 0,05 rechazamos H_0, por lo tanto, la variable Resto_Farmacos es una buena variable explicativa de las ventas o es significativa.

Además de verificar la significación individual de cada variable, debemos verificar la significación conjunta. Para ello hemos de comparar el valor crítico del estadístico teórico F con el obtenido en la regresión. Construimos el estadístico (3,434) grados de libertad:

```
local({
  .x <- seq(0.005, 6.025, length.out=1000)
  plotDistr(.x, df(.x, df1=3, df2=434), cdf=FALSE, xlab="x",
ylab="Density",
   main=paste("F Distribution:  Numerator df = 3, Denominator df
= 434"))
})
```



Vamos a estudiar la significación conjunta de todas las variables:

Significación conjunta

$H_0 : \beta_2 = \beta_3 = \beta_4 = 0$
$H_1 :$ al menos un $\beta_j \neq 0$ } $\alpha = 0{,}05$

$F^* = F(3,434) = 3783{,}457393$ $F^* > F_{tco}$ rechazo H_0, es decir, el modelo es significativo en su conjunto

$F_{tco} = 2{,}62546$ (se saca de la tabla estadística)

Siguiendo con el análisis de la bondad del ajuste analizamos los coeficientes de regresión:

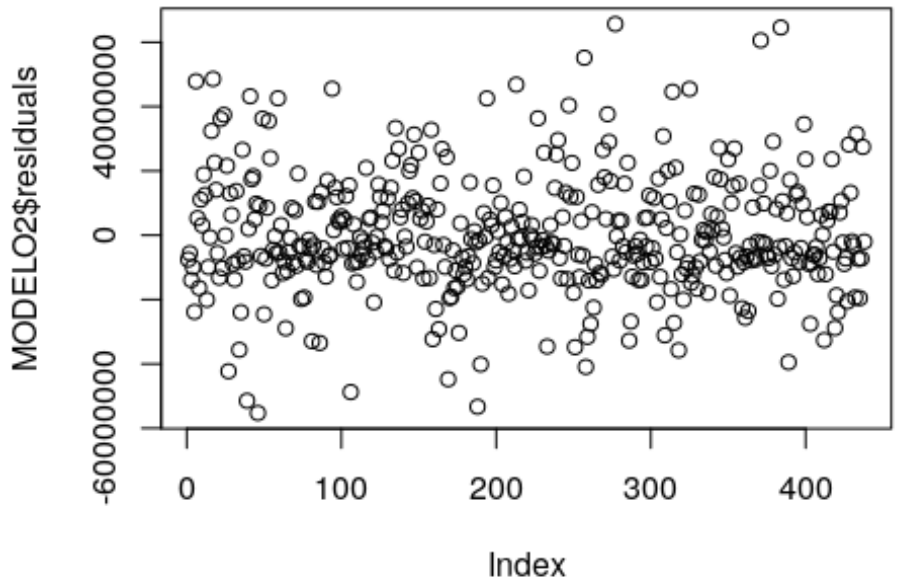
Coeficiente de determinación R^2	0,9632
R^2 ajustado	0,9629

Por tanto, nos encontramos ante un buen ajuste por ser datos de corte transversal y R2 > 50%.

El 96,31% de las variaciones de la variable endógena (Ventas) son explicadas a través de las variables del modelo.

Ahora analizamos la independencia de los residuos:

```
plot(MODELO2$residuals)
```

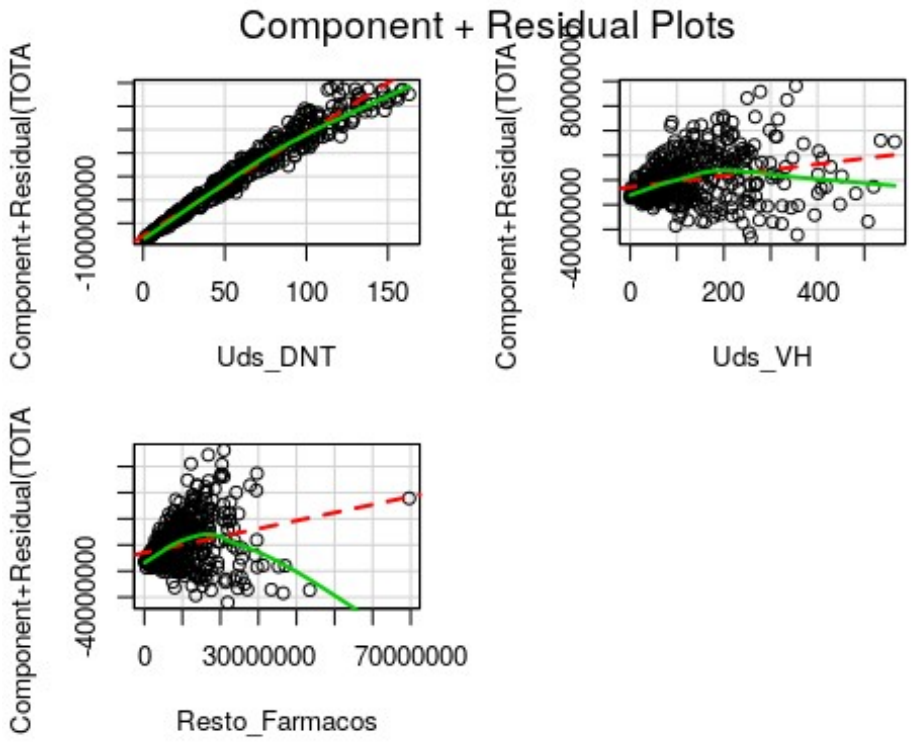


El gráfico de dispersión de los residuos no presenta un patrón definido, es decir, una recta, una parábola, tendencias de crecimiento o decrecimiento, sino que los residuos están distribuidos de forma aleatoria. Esto significa que cuando nuestro modelo falle, unas veces será al alza y otras a la baja. Esto es un requisito que deben cumplir los modelos estadísticos.

Se observa la presencia de homocedasticidad, ya que el comportamiento de los errores tiene varianza constante. Al parecer, el modelo es consistente.

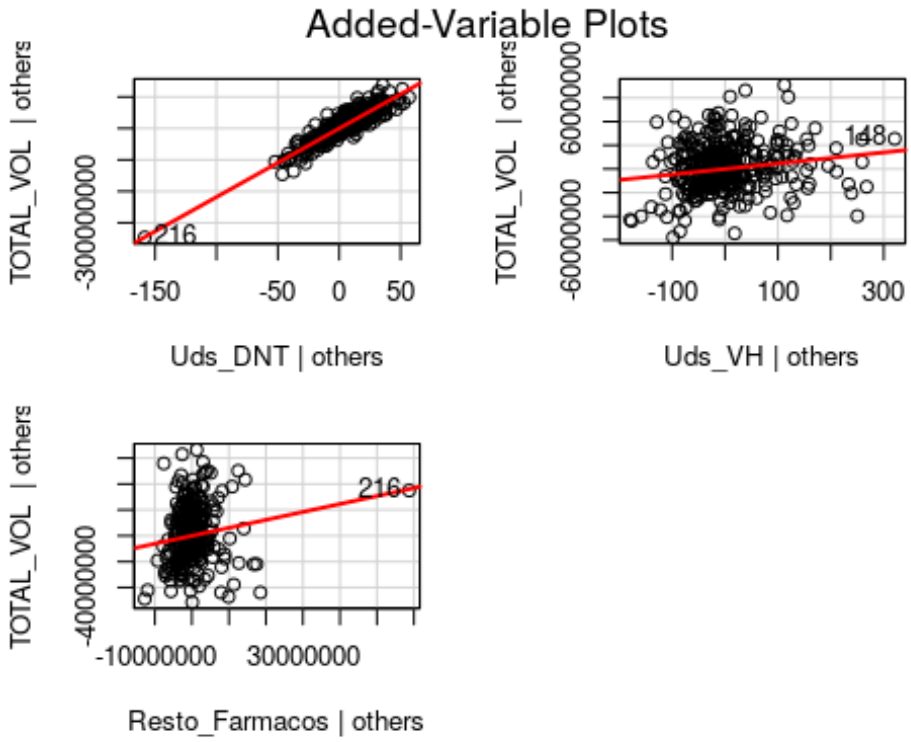
El gráfico de componentes y residuos por variable es:

```
crPlots(MODELO2, span=0.5)
```



Por último, representamos gráficamente el ajuste de cada variable en el modelo:

```
avPlots(MODELO2, id.method="mahal", id.n=1)
```



6.Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Tras el estudio realizado, hemos obtenido un buen modelo, por lo que podemos estimar las ventas totales con la siguiente ecuación:

$$Ventas_i = 6790675{,}48 + 2182195{,}34\,\beta_2 + 46548{,}1349\,\beta_3 + 0{,}61458762\,\beta_4$$

Los resultados permiten responder al problema planteado.

Para estimar el total del mercado usamos nuestra recta de regresión y vamos a sustituir las variables para cada uno de sus valores.

Coeficientes del modelo		
B1	**6790675,48**	**Constante**
B2	**2182195,34**	**Uds_DNT**
B3	**46548,1349**	**Uds_VH**
B4	**0,61458762**	**Resto_Farmacos**

A modo de ejemplo:

Uds_DNT	Uds_VH	Resto_Farmacos	Facturacion estimada
162	**117**	**21317197,37**	**378853738,10**
298	**42**	**321079,2096**	**659237240,04**
286	**84**	**1749363,513**	**635883723,47**
152	**47**	**3960844,178**	**343106415,44**

La suma de la facturación estimada de cada cliente del modelo dará la facturación total del mercado. En este caso:

6702743872712,94

Para calcular el potencial de cada cliente simplemente hemos de restarle a la facturación estimada la facturación actual:

FP = FE - FA

FA = Uds_AH + Uds_DNT + Uds_VH + Resto_Farmacos

Facturacion estimada	Facturacion actual	Diferencia
378853738,10	21407776,37	400080914,47
659237240,04	448619,2096	659431459,25
635883723,47	1795558,513	637587631,98
343106415,44	4219643,178	346808858,61
729895329,43	3290515,312	732765644,74

La cuota de mercado máxima para cada cliente la calculamos dividiendo la facturación estimada por el total:

CM = FE / FT

La cuota actual para cada cliente se calcula dividiendo la facturación actual entre la facturación estimada del mercado:

CA = FA / FT

Por último, la cuota potencial, es decir, la cuota de cada cliente que aún no hemos alcanzado, es decir, la diferencia entre lo que le facturamos y lo que le podríamos llegar a facturar o, lo que es lo mismo, el recorrido que tenemos con cada cliente lo calculamos como la diferencia entre las dos cuotas anteriores:

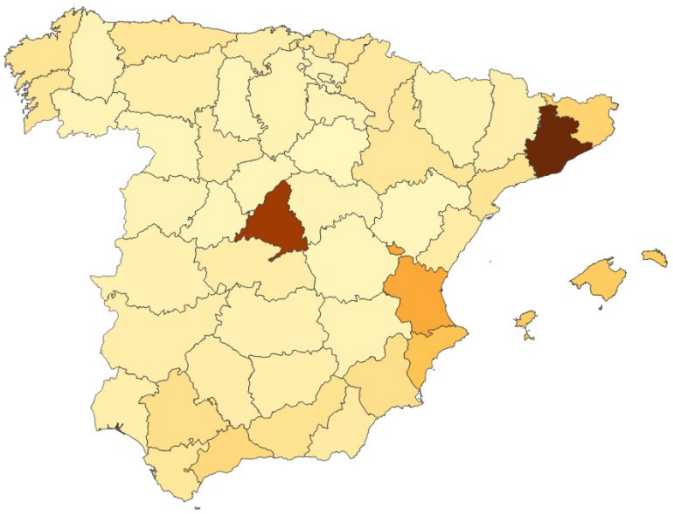
CP = CM – CA

Este último resultado es el que realmente nos interesa y es la base para seleccionar los clientes y proceder a establecer una estrategia.

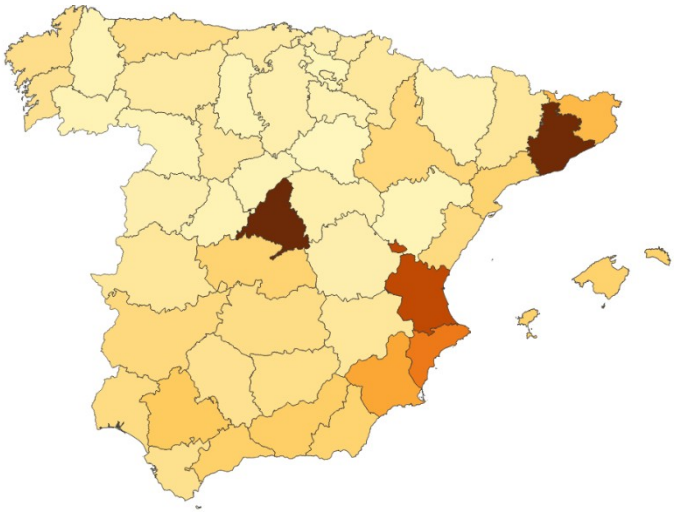
Facturacion estimada	Diferencia	Cuota total mercado	Cuota actual	Cuota potencial
378853738,10	400080914,47	0,005652%	0,000319%	0,005333%
659237240,04	659431459,25	0,009835%	0,000007%	0,009829%
635883723,47	637587631,98	0,009487%	0,000027%	0,009460%
343106415,44	346808858,61	0,005119%	0,000063%	0,005056%
729895329,43	732765644,74	0,010890%	0,000049%	0,010840%

Con una herramienta de visualización podemos crear unos mapas por cada variable donde aparezcan la distinta intensidad de las ventas con una escala de colores diferentes. En ese caso, realizo este trabajo con la herramienta QLIK SENSE y obtengo el siguiente resultado:

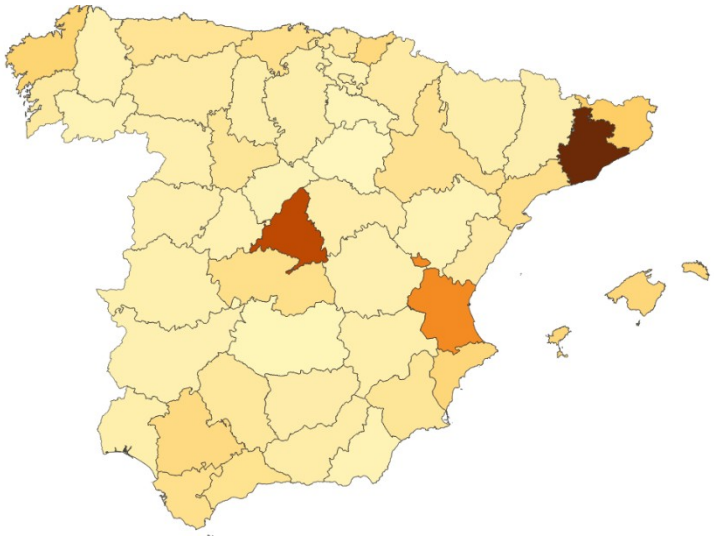
Facturación estimada



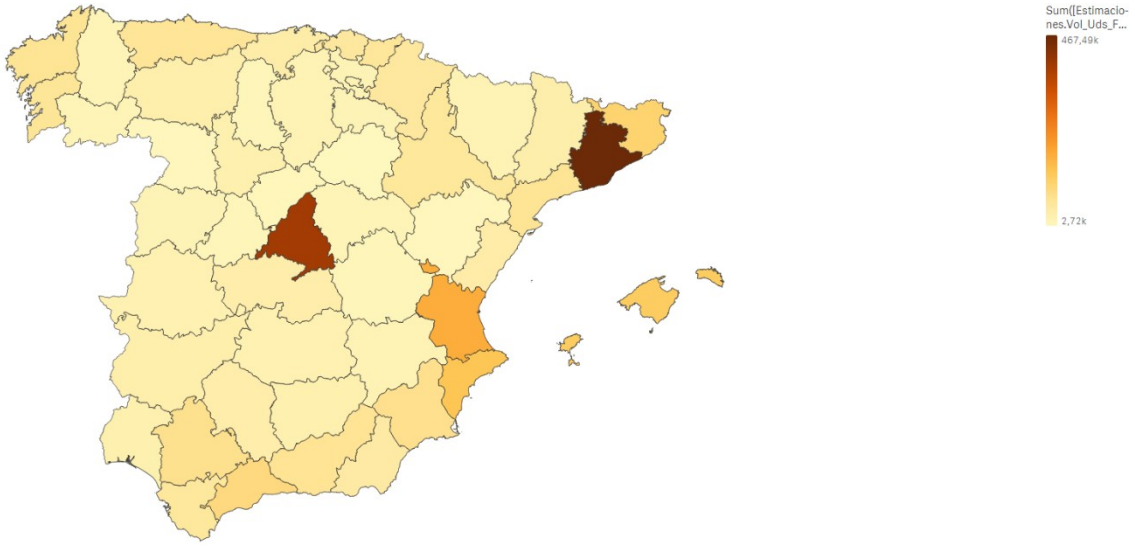
Volumen unidades VH



Volumen unidades AH



Volumen unidades DNT



Resto de productos