resol_practica3.R

gemamaria

2024-09-30

```
## REGRESIÓN

## Ejercicio 1. Cargo datos y los "veo".

#library(MASS) #recomendada por el libro
library(ISLR2)
library(scatterplot3d)

## La siguiente linea de comando no os servirá.
## Para cargar la base de datos, File->Import Dataset, y elegid el directorio donde tengáis el fichero.

Advertising <- read.csv("~/matematicas/aprendizaje estadistico/GEMA 24_25/segunda semana/Advertising.csv
#View(Advertising)
head(Advertising)
```

```
##   X    TV radio newspaper sales
## 1 1 230.1  37.8      69.2  22.1
## 2 2  44.5  39.3      45.1  10.4
## 3 3  17.2  45.9      69.3   9.3
## 4 4 151.5  41.3      58.5  18.5
## 5 5 180.8  10.8      58.4  12.9
## 6 6   8.7  48.9      75.0   7.2
```

```
names(Advertising)
```

```
## [1] "X"         "TV"        "radio"     "newspaper" "sales"
```

```
dim(Advertising)
```

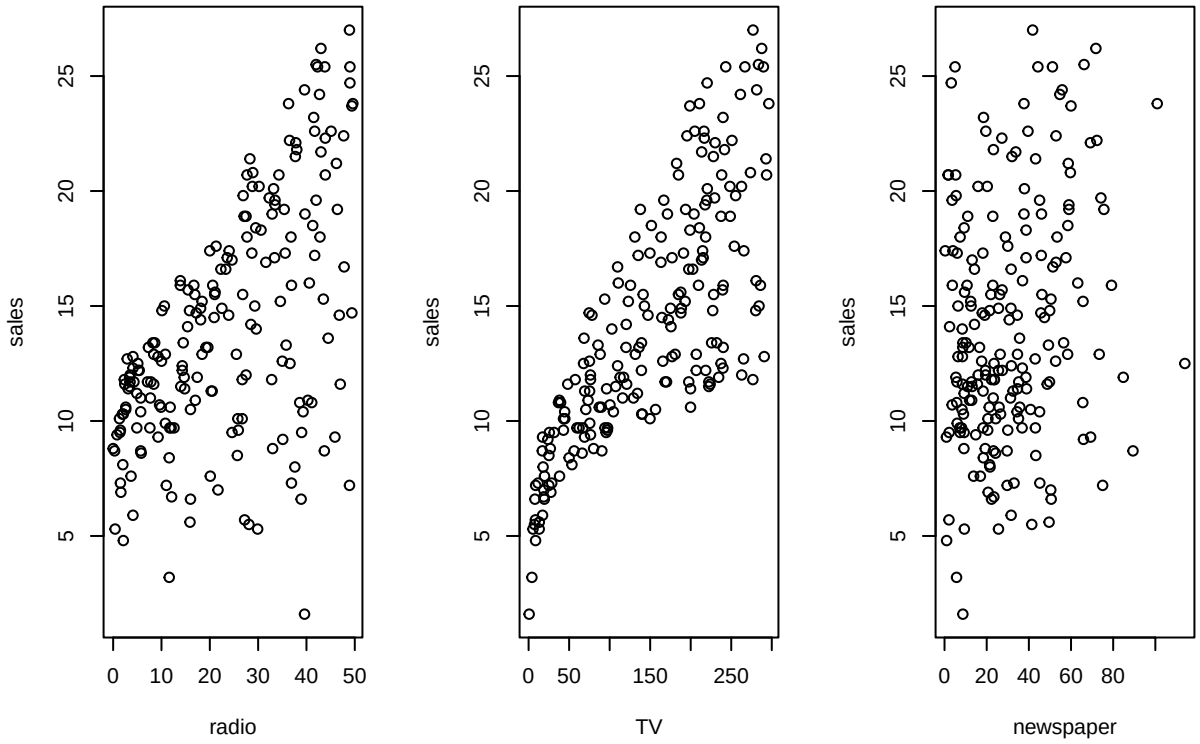
```
## [1] 200   5
```

```
str(Advertising)
```

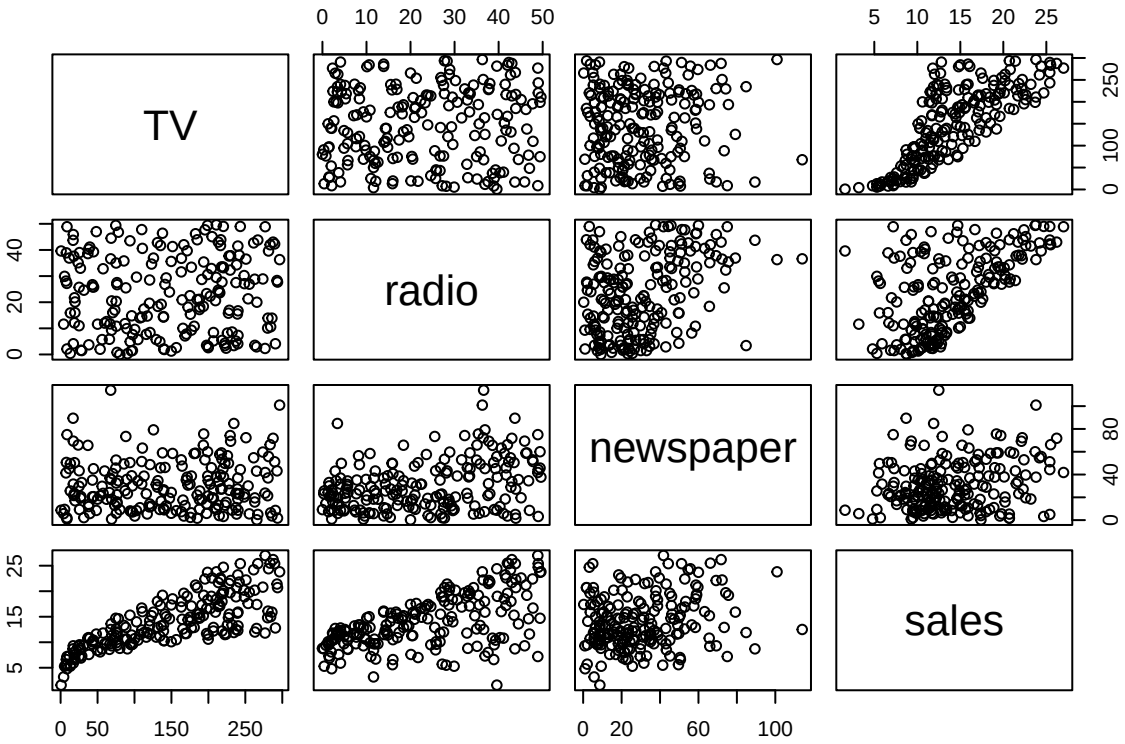
```
## 'data.frame':    200 obs. of  5 variables:
##  $ X        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ TV       : num  230.1 44.5 17.2 151.5 180.8 ...
##  $ radio    : num  37.8 39.3 45.9 41.3 10.8 48.9 32.8 19.6 2.1 2.6 ...
##  $ newspaper: num  69.2 45.1 69.3 58.5 58.4 75 23.5 11.6 1 21.2 ...
##  $ sales    : num  22.1 10.4 9.3 18.5 12.9 7.2 11.8 13.2 4.8 10.6 ...
```

```
  ## Visualizo los datos

## Veamos si la variable Sales depende del dinero invertido en publicidad en radio, TV y periódico.
## Los siguientes plots predicen que nos vamos a encontrar.
## Fijemos antes la base de datos:
```

```r
attach(Advertising)

## El siguiente comando junta 3 figuras en una sola fila.

par(mfcol=c(1,3))
plot( radio, sales)
plot(  TV, sales)
plot ( newspaper, sales)
```




```r
## ¿CONCLUSIONES?

## otra manera de visualizar:

plot(Advertising[,-1])
```




```
## Ejercicio 2. Regresión lineal - Lineal Simple


## lm:  lineal model

rectaTV<-lm(sales~TV,data = Advertising)
rectaRadio<-lm(sales~radio,data = Advertising)
rectaNews<-lm(sales~newspaper,data = Advertising)

## Analicemos rectaTV. Ploteemos la recta:
plot(TV,sales)
abline(rectaTV)

## Datos numéricos:

rectaTV
```

```
##
## Call:
## lm(formula = sales ~ TV, data = Advertising)
##
## Coefficients:
## (Intercept)           TV
##     7.03259      0.04754
```

```
## Es decir, el modelo lineal es: sales = 7.03+0.047*TV.
## Por cada 1000 dólares (TV=1) invertido, habría un aumento en ventas en
## promedio de 47 unidades del producto.

## La siguiente orden, summary, da datos que me permiten
## analizar la efectividad del modelo.
```

```r
summary(rectaTV)
```

```
##
## Call:
## lm(formula = sales ~ TV, data = Advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.032594   0.457843   15.36   <2e-16 ***
## TV          0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

```r
## CONCLUSIONES:

## ESTUDIO DE LA PRECISIÓN DE LAS ESTIMACIONES DE LOS COEFICIENTES

## Observemos los errores standar de ambos coeficientes (Std. Error).
## Con ellos, podemos definir (estimar) los intervalos de confianza, de tal manera que
## hay un 95% de probabiliad de que el intervalo contenga la "verdadera" pendiente,\beta_1

0.047537-2*0.002691
```

```
## [1] 0.042155
```

```r
0.047537+2*0.002691
```

```
## [1] 0.052919
```

```r
## idem para el término independiente,

7.032594-2*0.457843
```

```
## [1] 6.116908
```

```r
7.032594+2*0.457843
```

```
## [1] 7.94828
```

```r
## también podemos utilizar la función confint
confint(rectaTV,level=0.95)
```

```
##                 2.5 %     97.5 %
## (Intercept) 6.12971927 7.93546783
## TV          0.04223072 0.05284256
```

```r
## De estos intervalos de confianza, deducimos que, en ausencia de publicidad,
## las ventas, en promedio, caerán entre 6130 y 7935 unidades.
## Además, por cada $1000 de aumento en la publicidad televisiva,
## habrá un aumento promedio en las ventas de entre 42 y 53 unidades.
```

```
## Respecto al contraste de hipótesis, un p-value pequeño en ambos coeficiente indica que
## podemos rechazar la hipótesis de que los "verdaderos" coeficientes sean nulos.
## Recordad que el t-statistic es el valor estimado para el parámetro dividido por el
## error estandar, y el p-value es la probabilidad de obtener dicho estadístico o mayor,
## sin que haya relación, o sea, H_0 cierta.

## ESTUDIO DE LA PRECISIÓN DEL MODELO (bondad de ajuste: cómo se ajusta el modelo a los datos)

## Una vez que hemos rechazado que los coeficientes del modelo sean nulos,
## veamos cómo se ajustan a los datos.

## Observamos que el summary proporciona el RSE,
## RSE=ERROR ESTÁNDAR RESIDUAL=DESVIACIÓN TÍPICA DE la VARIABLE RESIDUAL = 3.26
## Recordemos que a medida de que el RSE se aleja de 0, peor será el ajuste.
## RSE=3.26 indica que si seguimos el modelo, nos podríamos desviar 3260 unidades
## por 1000 dólares invertidos, es decir,
## como la media de las ventas es de 14000 unidades (redondeando),

mean(sales)
```

```
## [1] 14.0225
```

```
##entonces el error sería de 23% :
3.26/14
```

```
## [1] 0.2328571
```

```
## El cálculo del RSE se considera la variable sales y el modelo lineal,
## residuos: sales-fitted:

r1<-residuals(rectaTV)
sqrt(sum(r1^2)/(200-2)) #así recordamos la definición
```

```
## [1] 3.258656
```

```
##sales con el modelo lineal:
fitted(rectaTV)
```

```
##         1         2         3         4         5         6         7         8
## 17.970775  9.147974  7.850224 14.234395 15.627218  7.446162  9.765950 12.746498
##         9        10        11        12        13        14        15        16
##  7.441409 16.530414 10.174765 17.238710  8.163966 11.667416 16.734822 16.321253
##        17        18        19        20        21        22        23        24
## 10.255578 20.409404 10.322129 14.034741 17.414596 18.317792  7.660077 17.885209
##        25        26        27        28        29        30        31        32
##  9.994126 19.529976 13.825579 18.446141 18.859710 10.388680 20.956076 12.399480
##        33        34        35        36        37        38        39        40
## 11.653155 19.658325 11.581850 20.851495 19.720123 10.583581  9.081423 17.870948
##        41        42        43        44        45        46        47        48
## 16.658763 15.446579 20.989351 16.867924  8.225763 15.356259 11.296630 18.436634
##        49        50        51        52        53        54        55        56
## 17.832918 10.212795 16.530414 11.805272 17.319523 15.712784 19.520469 16.487631
##        57        58        59        60        61        62        63        64
##  7.379611 13.507084 17.053317 17.048564  9.575804 19.453918 18.408112 11.914607
##        65        66        67        68        69        70        71        72
## 13.264647 10.312622  8.529998 13.654448 18.317792 17.338537 16.497139 12.252117
```

```
##        73        74        75        76        77        78        79        80
##  8.306576 13.183835 17.176913  7.835963  8.339851 12.760759  7.289291 12.546844
##        81        82        83        84        85        86        87        88
## 10.664393 18.431880 10.612103 10.284100 17.181666 16.216672 10.659639 12.294900
##        89        90        91        92        93        94        95        96
## 11.230079 12.252117 13.416764  8.392141 17.381320 18.959537 12.138029 14.795327
##        97        98        99       100       101       102       103       104
## 16.425834 15.822118 20.803958 13.459547 17.604742 21.122454 20.352360 15.964728
##       105       106       107       108       109       110       111       112
## 18.355821 13.587896  8.221010 11.329906  7.655324 19.173452 17.766367 18.522200
##       113       114       115       116       117       118       119       120
## 15.384781 16.996273 10.749959 10.602595 13.649694 10.664393 13.007949  7.954804
##       121       122       123       124       125       126       127       128
## 13.749521  7.926282 17.680801 12.884354 17.942253 11.177789  7.403379 10.845032
##       129       130       131       132       133       134       135       136
## 17.504915  9.865777  7.065869 19.639311  7.431901 17.481147  8.786696  9.328613
##       137       138       139       140       141       142       143       144
##  8.249532 20.043372  9.076669 15.822118 10.521783 16.240441 17.514423 12.004926
##       145       146       147       148       149       150       151       152
## 11.605618 13.701984 18.446141 18.593505  8.838986  9.157481 20.376129 12.784527
##       153       154       155       156       157       158       159       160
## 16.425834 15.175620 15.959975  7.227494 11.496284 14.153582  7.588772 13.293169
##       161       162       163       164       165       166       167       168
## 15.232664 11.106484 15.988497 14.804834 12.603888 18.179936  7.883499 16.863171
##       169       170       171       172       173       174       175       176
## 17.271986 20.547260  9.409426 14.852371  7.964312 15.037764 17.604742 20.195489
##       177       178       179       180       181       182       183       184
## 18.840695 15.123330 20.185982 14.904661 14.476831 17.419349  9.704153 20.704131
##       185       186       187       188       189       190       191       192
## 19.097393 16.777605 13.663955 16.116846 20.628073  7.921529  8.910291 10.621610
##       193       194       195       196       197       198       199       200
##  7.850224 14.961705 14.148829  8.848493 11.510545 15.446579 20.513985 18.065848
```

```
## en cambio, el cálculo del estadístico R^2 requiere sales y presupuesto.
## Tenemos R^2= 0.61, se puede entender como r^2,
## con lo cual es un buen indicador de la relación lineal entre ambas variables;
## y también como una proporción: 61% de la variabilidad de la variable SALES
## se explica con el gasto en TV.

## Recordad que:
## R^2=VE/TSS= (TSS-RSS)/TSS = 1 - RSS/TSS

f1<-fitted(rectaTV) ##variable sales ajustada con el modelo lineal
sum((f1-mean(sales))^2)/sum((sales-mean(sales))^2)
```

```
## [1] 0.6118751
```

```
## MÁS COSAS
## Pensemos en la idea de "estimador".
## Podríamos considerar la población como todo el conjunto;
## veamos qué ocurre si consideramos cuatro conjuntos de entrenamiento.

set.seed(1)
train1 <- sample(1:nrow(Advertising), nrow(Advertising) / 2)
train1
```

```
##   [1]  68 167 129 162  43  14 187  51  85  21 106 182  74   7  73  79  37 105
##  [19] 110 165  34 190 126  89 172  33  84 163  70 188  42 166 111 148 156  20
##  [37]  44 121  87 176 173  40  25 119 122  39 170 134  24 195 130  45 146  22
##  [55] 115 104 161 144 145 103  75  13 159 177  23 189 174 141  29 108  48 175
##  [73] 149 191  31 102  17 186 133 197  83 118 114  90 150 107  64  94 179  96
##  [91] 169  60 193  93 180  10   1 196  59  26
```

```
train2 <- sample(1:nrow(Advertising), nrow(Advertising) / 2)
train2
```

```
##   [1] 143 186  29 152 170  48  39  24 181  40  83  90 163  43   1 198  78 150
##  [19]  70  28 116  37  61 174 113  86  71 110  99  51  44  49 105  60 169  50
##  [37] 135 111  20 121 197 171  53 144 100 130 195  65 196 103 168 124  77  98
##  [55]  19  17  31 172  75  16 191   9 165  92 122 126  14 180 141  15 155 153
##  [73] 128  67 120  73 102 145  84   5  41  91 108 157  72  36 166 164 185  85
##  [91]  64 199 158 136 106  57  88 118 175 138
```

```
train3 <- sample(1:nrow(Advertising), nrow(Advertising) / 2)
train3
```

```
##   [1]  80  30  93 130  72 126  78 164 168 165 196 184 116 100 113 121  73  27
##  [19]  41  15  38  62 134 132  35 125  99  77 105  71 153  31  37  28 179 148
##  [37]  29 127  42  60 167 194 169  12 198 200  44  98  26  33 177 117  86  24
##  [55] 192 108 181  14  82  97 190  53  56 170 101 156  75   2 188 131 111 144
##  [73] 160  55  69  92 159 146  43 107  81   1  90  45 141  83 118  23  68 155
##  [91]  91  57 103 102  52 197  34 135 176  74
```

```
train4 <- sample(1:nrow(Advertising), nrow(Advertising) / 2)
train4
```

```
##   [1] 198 188 167  49 190 161  35 139 141 178  60  90 189 112 168 162  68  56
##  [19]  25 166  81  73 173 179 150 104   3 147 195 128 127   4 196  65 133 106
##  [37] 194 121  23 157 117 164 119  99 148 176 163  86 108  18  44  29  91  32
##  [55]  51 158 146 125  89  21  79 115 101  78  71 171  59 199  58  15 114  64
##  [73]   6  34 192   1  17  77  62 160  45 181  40  66  41   8 182  95  20 110
##  [91]  84   5  74 187  55 183 140 149   2 143
```

```
dim(Advertising[train1, ])
```

```
## [1] 100   5
```

```
lm1=lm(Advertising[train1, ]$sales~Advertising[train1, ]$TV )
lm1$coefficients
```

```
##              (Intercept) Advertising[train1, ]$TV
##               7.11989658               0.04691396
```

```
lm2=lm(Advertising[train2, ]$sales~Advertising[train2, ]$TV )
lm2$coefficients
```

```
##              (Intercept) Advertising[train2, ]$TV
##               7.27719228               0.04699388
```

```
lm3=lm(Advertising[train3, ]$sales~Advertising[train3, ]$TV )
lm3$coefficients
```

```
##              (Intercept) Advertising[train3, ]$TV
##               6.28910846               0.05048601
```

```
lm4=lm(Advertising[train4, ]$sales~Advertising[train4, ]$TV )
lm4$coefficients
```

```
##              (Intercept) Advertising[train4, ]$TV
##                6.98358581               0.04967699
```

```
## todos los estimadores están en el intervalo de confianza anterior:

confint(rectaTV)
```

```
##                   2.5 %     97.5 %
## (Intercept) 6.12971927 7.93546783
## TV          0.04223072 0.05284256
```

```
## Al ser estimadores insesgados, a medida que calculo más, me voy acercando
## al valor correcto.

( lm1$coefficients[1]+lm2$coefficients[1]+lm3$coefficients[1]+lm4$coefficients[1])/4
```

```
## (Intercept)
##    6.917446
```

```
( lm1$coefficients[2]+ lm2$coefficients[2]+ lm3$coefficients[2]+ lm4$coefficients[2])/4
```

```
## Advertising[train1, ]$TV
##                0.04851771
```

```
rectaTV[1]
```

```
## $coefficients
## (Intercept)          TV
##  7.03259355  0.04753664
```

```
## Observad que a menos datos, más longitud tiene el intervalo de confianza.
confint(lm1)
```

```
##                               2.5 %     97.5 %
## (Intercept)              5.81283581 8.42695736
## Advertising[train1, ]$TV 0.03938022 0.05444769
```

```
## sales = 7.03+0.047*TV

## Por otro lado, si queremos predecir las ventas dependiendo de lo invertido en TV
## en el complementario de train1, haríamos:
predict(lm1, Advertising[-train1, ] )
```

```
##         2         3         4         5         6         8         9        11
## 13.655011  7.959656 17.455042 11.140423 20.893835 11.694008 13.664394 16.493305
##        12        15        16        18        19        27        28        30
## 17.136027 17.365905 13.589331 17.370596 13.190563  9.817449  8.377191  7.373232
##        32        35        36        38        41        46        47        49
## 19.641232 18.294801 19.101722 12.618212 19.580244  7.997188 11.210794 11.262399
##        50        52        53        54        55        56        57        58
## 14.837243 11.679933 10.328811 15.958486 17.290843 16.085154 15.423667 18.121220
##        61        62        63        65        66        67        69        71
## 17.713068 18.529371  7.312244 14.030323 16.826395 13.748839 10.699432 20.110372
##        72        76        77        78        80        81        82        86
##  8.039410 17.816279 10.042636 13.016981  8.001879  9.141888 20.457535 17.431585
##        88        91        92        95        97        98        99       100
```

```
## 17.830353 14.142916  9.915969  8.297437 13.701925 18.257270 10.788568 15.935029
##       101       109       112       113       116       117       120       123
## 15.212554 12.027097 11.633019 20.265188 17.131335  8.236449  7.668790 18.773324
##       124       125       127       128       131       132       135       136
##  7.739161 20.537289 15.020207 10.563381 18.792089 11.360918 18.374555 17.553561
##       137       138       139       140       142       143       147       151
##  8.902627  8.972998 20.860995 21.025194 10.300663 16.737258  7.513974 11.539191
##       152       153       154       155       157       158       160       164
## 10.652518 10.704123 16.953062 12.271049  9.216951  8.292746 11.937960 18.890609
##       168       171       178       181       183       184       185       192
## 20.100989 14.780946 17.225163 17.004668  7.926817 17.333065 14.888848 16.493305
##       194       198       199       200
## 17.914798  8.912010 17.009359 19.453576
```

```
## EJERCICIO: Analizad vosotros las otras dos rectas de regresión, rectaRadio y rectaNews.
## Deberíais observar que el modelo lineal no es lo ideal para estudiar la relación entre
## el presupuesto de publicidad en periódicos y ventas.

## Ejercicio 3. Regresión lineal múltiple

## Estudiemos el modelo lineal con los tres predictores, TV, Radio y Newspaper.

rectaMul=lm(sales~TV+radio+newspaper)
rectaMul
```

```
##
## Call:
## lm(formula = sales ~ TV + radio + newspaper)
##
## Coefficients:
## (Intercept)           TV        radio    newspaper
##    2.938889     0.045765     0.188530    -0.001037
```

```
## Obtenemos sales= 2.9389 +  0.045765 TV + 0.188530 radio  - 0.001037 newspaper

## Este modelo se podría interpretar como sigue:
## para un presupuesto fijo en TV y Newsp., gastar $1000 en publicidad radiofónica
## implica 189 unidades más en ventas de promedio.

##Hacemos summary.
summary(rectaMul)
```

```
##
## Call:
## lm(formula = sales ~ TV + radio + newspaper)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## TV           0.045765   0.001395  32.809   <2e-16 ***
## radio        0.188530   0.008611  21.893   <2e-16 ***
## newspaper   -0.001037   0.005871  -0.177     0.86
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

```
## ESTUDIO DE LA PRECISIÓN DEL MODELO (bondad de ajuste: cómo se ajusta el modelo a los datos)

## tanto el  RSE=sqrt(RSS/n-p-1), con p=3=nº de predictores=nºde vbles independientes,
## como R^2 son óptimos, pero el valor de beta_3 indica que algo no va bien.

## ESTUDIO DE LA PRECISIÓN DE LAS ESTIMACIONES DE LOS COEFICIENTES

## Observamos que los coeficientes de TV y radio son similares a los obtenidos en el modelo simple,
## pero el de newspaper es negativo, cercano a 0, y su p-value es muy elevado.

## No aparecía negativo en rectaNews, aunque su p value no era óptimo y 0 pertenecía
## al intervalo de confianza.

## Esto nos hace pensar que podría haber una relación entre los predictores, lo que, a priori,
## resta precisión al modelo.

## Recordemos que la correlación entre dos variables (el estimador r) mide su relación lineal.
cor(data.frame(TV,radio,newspaper,sales))
```

```
##                    TV      radio  newspaper     sales
## TV         1.00000000 0.05480866 0.05664787 0.7822244
## radio      0.05480866 1.00000000 0.35410375 0.5762226
## newspaper  0.05664787 0.35410375 1.00000000 0.2282990
## sales      0.78222442 0.57622257 0.22829903 1.0000000
```

```
cov(data.frame(TV,radio,newspaper,sales))
```

```
##                    TV     radio newspaper     sales
## TV         7370.94989  69.86249 105.91945 350.39019
## radio        69.86249 220.42774 114.49698  44.63569
## newspaper   105.91945 114.49698 474.30833  25.94139
## sales       350.39019  44.63569  25.94139  27.22185
```

```
## Observamos que hay una correlación entre radio-newspaper mayor que newspaper-sales.
## Mirad la recta de regresión
plot(radio,newspaper)
abline(lm(newspaper ~ radio))
## en mercados donde invierten en publicidad en radio, también invierten en newspaper.
lm(newspaper ~ radio)
```

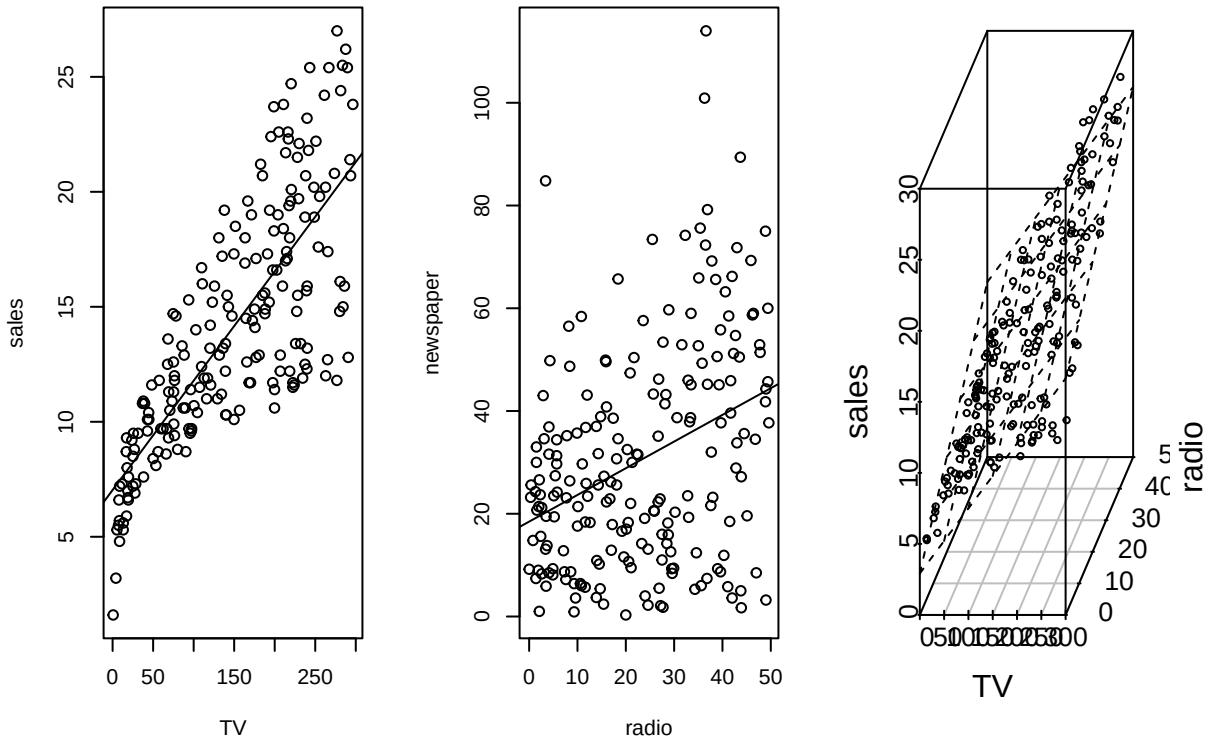
```
##
## Call:
## lm(formula = newspaper ~ radio)
##
## Coefficients:
## (Intercept)        radio
##     18.4700       0.5194
```

```
##Veamos el modelo eliminando newspaper
rectaMul2=lm(sales ~ TV+radio)
```

```
rectaMul2
```

```
##
## Call:
## lm(formula = sales ~ TV + radio)
##
## Coefficients:
## (Intercept)           TV        radio
##     2.92110      0.04575      0.18799
```

```
## sales=2.92110  +  0.04575*TV +  0.18799*radio
## los coeficientes no son tan diferentes en TV y radio que en el modelo considerando los
## tres predictores, y si realmente hay colinealidad, los residuos deben ser iguales.
## El álgebra lineal nos da el motivo.

## Ploteemos el modelo lineal con TV y radio como predictores, y saquemos conclusiones.

scatterplot3d(TV,radio,sales)$plane3d(rectaMul2)
```




```
## Discutamos las preguntas de la sección 3.2.2.

## ¿Al menos un predictor es útil para predecir las ventas?

## si algún coeficiente es diferente a 0, los F-statistic son mayores que 1.
## Observamos que los F-statistic de rectaMul y rectaMul2 se alejan bastante de 1,
## y los p'values son nulos, por lo que ambos estadísticos indican que al menos
## una variable predictora es útil para predecir las ventas (Sales)
summary(rectaMul2)
```

```
##
## Call:
```

```
## lm(formula = sales ~ TV + radio)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7977 -0.8752  0.2422  1.1708  2.8328
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.92110    0.29449   9.919   <2e-16 ***
## TV           0.04575    0.00139  32.909   <2e-16 ***
## radio        0.18799    0.00804  23.382   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.681 on 197 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8962
## F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

```
summary(rectaMul)
```

```
##
## Call:
## lm(formula = sales ~ TV + radio + newspaper)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## TV           0.045765   0.001395  32.809   <2e-16 ***
## radio        0.188530   0.008611  21.893   <2e-16 ***
## newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

```
## ¿Hay que considerar todos los predictores o sólo un subconjunto?

## Tras lo visto, y observando los p-values individuales en ambos modelos,
## se puede prescindir de la variable Newspaper.

## ¿Cómo se ajusta el modelo a los datos?

## Para contestar a esta pregunta, nos debemos fijar en los estadísticos R^2 y
## RSE=ERROR ESTÁNDAR RESIDUAL=DESVIACIÓN TÍPICA DEL ERROR

## Tanto el modelo simple como en el múltiple,
## R^2 nos da la proporción de la varianza de la variable Y (Sales en este caso),
## explicada por el modelo. A priori, cuanto más se acerque a 1, mejor se ajusta el modelo.
## Observamos que los R^2 en ambos modelos múltiples son iguales, 0.8972, muy cercanos a 1.
```

```r
## Observamos también que el R^2 "ajustado" es ligeramente mayor eliminando Newspaper.

## R^2 ajustado aparece en el capítulo 6, página 234:

## R^2 ajustado = 1-[RSS/(n-p-1)]/[TSS/(n-1)]


## Todo ello indica que Newspaper se debe eliminar del modelo.
## En general, el estadística R^2 aumenta al aumentar el número de variables,
## aunque éstas estén debilmente relacionadas con la variable que se quiere predecir.

## Observad que en cambio, R^2 de modelo simple utilizando sólo TV o sólo radio,
## se aleja de 0.89

summary(rectaRadio)
```

```
##
## Call:
## lm(formula = sales ~ radio, data = Advertising)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7305  -2.1324   0.7707   2.7775   8.1810
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.31164    0.56290  16.542   <2e-16 ***
## radio        0.20250    0.02041   9.921   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.275 on 198 degrees of freedom
## Multiple R-squared:  0.332,  Adjusted R-squared:  0.3287
## F-statistic: 98.42 on 1 and 198 DF,  p-value: < 2.2e-16
```

```r
summary(rectaTV)
```

```
##
## Call:
## lm(formula = sales ~ TV, data = Advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.032594   0.457843   15.36   <2e-16 ***
## TV          0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
```

```
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

```r
## Dado valores concretos de los predictores, ¿qué respuesta obtenemos?

## Para utilizar los modelos lineales para predicción, por ejemplo,
## si TV=220, radio=30,
## con predict se puede mostrar el intervalo de confianza para la venta media
## con TV=220, radio=30,

predict(rectaMul2,data.frame(TV=220,radio=30) ,interval="confidence")
```

```
##        fit      lwr      upr
## 1 18.62699 18.30441 18.94956
```