Análisis de datos para la toma de decisiones de comerciales en el entorno R

Jordi López Sintas

20 de enero de 2015

Presentación

Las tecnologías para el análisis de datos, tales como t-test, ANOVA, regresión, el análisis conjunto y análisis de los factores principales son ampliamente utilizadas en la toma de decisiones en marketing (análisis de las preferencias del consumidor, segmentación del mercado, decisiones del precio del producto, análisis los atributos del producto que influyen en las ventas, y las ventas etc.) Tradicionalmente se utilizan herramientas comerciales genéricas, como SPSS y SAS, o específicas, como ME|XL, sin embargo, el lenguaje de código abierto R está siendo utilizado cada vez más. En este artículo, vamos a presentar cómo utilizar R para llevar a cabo algunos análisis básicos para la toma de decisiones de comercialización basadas en evidencia.

Introducción

Una empresa, ABC store chain, vende un nuevo tipo de jugo de uva en algunas de sus tiendas piloto. El equipo de marketing de ABC quiere analizar:

1. ¿Qué tipo de anuncio en la tienda es más eficaz? Se han colocado dos tipos de anuncios en las tiendas para las pruebas, un tema es la producción natural del jugo, el otro tema es el cuidado de la salud de la familia;
2. La elasticidad-precio - las reacciones de los volúmenes de venta de zumo de uva a su cambio de precio;
3. La Elasticidad-cruzada del precio - las reacciones de los volúmenes de venta de zumo de uva a los cambios en los precios de otros productos, como el jugo de manzana y galletas en la misma tienda;
4. ¿Cómo encontrar el mejor precio unitario del jugo de uva que puede maximizar el beneficio¿ ¿Cuál sería la previsión de ventas con ese precio?

El equipo de marketing ha tomado muestras al azar de 30 observaciones y construido el siguiente conjunto de datos para el análisis. Hay 5 variables (columnas de datos) en el conjunto de datos.

Variable	Descripción
sales	Las ventas unitarias totales del zumo de uva en una semana en una tienda.
price	precio unitario medio del jugo de uva en la semana.
ad_type	El tipo de publicidad en las tiendas para promover el zumo de uva
ad_type = 0	el tema de la publicidad es la producción natural del zumo
ad_type= 1	el tema de la publicidad es el cuidado de la salud de la familia.
price_apple	precio unitario medio del zumo de manzana en la misma tienda en la semana
price_cookies	precio unitario medio de las cookies en la misma tienda en la semana

El conjunto de datos se pueden descargar desde este enlace: http://www.dataapple.net/wp-content/uploads/2013/04/grapeJuice.csv. Por favor, tenga en cuenta el conjunto de datos ha sido construido por el autor con fines ilustrativos, por lo que tal vez pueda parecer diferente de los datos del mundo real.

Lectura de Datos

Vamos a tener un poco de exploración básica para saber más sobre el conjunto de datos.

```
# Cargar las bibliotecas necesarias en el siguiente
#lectura de datos
#install.packages('s20x')
#install.packages('car')
library(s20x)
library(car)
```

```
##
## Attaching package: 'car'
##
## The following object is masked from 'package:s20x':
##
##     levene.test
```

```
#read the dataset from an existing .csv file
url<-"http://www.dataapple.net/wp-content/uploads/2013/04/grapeJuice.csv"
df <- read.csv(url,header=T)
#list the name of each variable (data column) and the first six rows of the dataset
head(df)
```

```
##   sales price ad_type price_apple price_cookies
## 1   222  9.83       0        7.36          8.80
## 2   201  9.72       1        7.43          9.62
## 3   247 10.15       1        7.66          8.90
## 4   169 10.04       0        7.57         10.26
## 5   317  8.38       1        7.33          9.54
## 6   227  9.74       0        7.51          9.49
```

```
# basic statistics of the variables
str(df)
```

```
## 'data.frame':    30 obs. of  5 variables:
##  $ sales        : int  222 201 247 169 317 227 214 187 188 275 ...
##  $ price        : num  9.83 9.72 10.15 10.04 8.38 ...
##  $ ad_type      : int  0 1 1 0 1 0 1 0 1 0 ...
##  $ price_apple  : num  7.36 7.43 7.66 7.57 7.33 7.51 7.57 7.66 7.39 8.29 ...
##  $ price_cookies: num  8.8 9.62 8.9 10.26 9.54 ...
```

```
summary(df)
```

```
##      sales           price           ad_type     price_apple
##  Min.   :131.0   Min.   : 8.200   Min.   :0.0   Min.   :7.300
##  1st Qu.:182.5   1st Qu.: 9.585   1st Qu.:0.0   1st Qu.:7.438
##  Median :204.5   Median : 9.855   Median :0.5   Median :7.580
##  Mean   :216.7   Mean   : 9.738   Mean   :0.5   Mean   :7.659
##  3rd Qu.:244.2   3rd Qu.:10.268   3rd Qu.:1.0   3rd Qu.:7.805
##  Max.   :335.0   Max.   :10.490   Max.   :1.0   Max.   :8.290
```

```
##  price_cookies
##  Min.   : 8.790
##  1st Qu.: 9.190
##  Median : 9.515
##  Mean   : 9.622
##  3rd Qu.:10.140
##  Max.   :10.580
```

Exploración de los datos

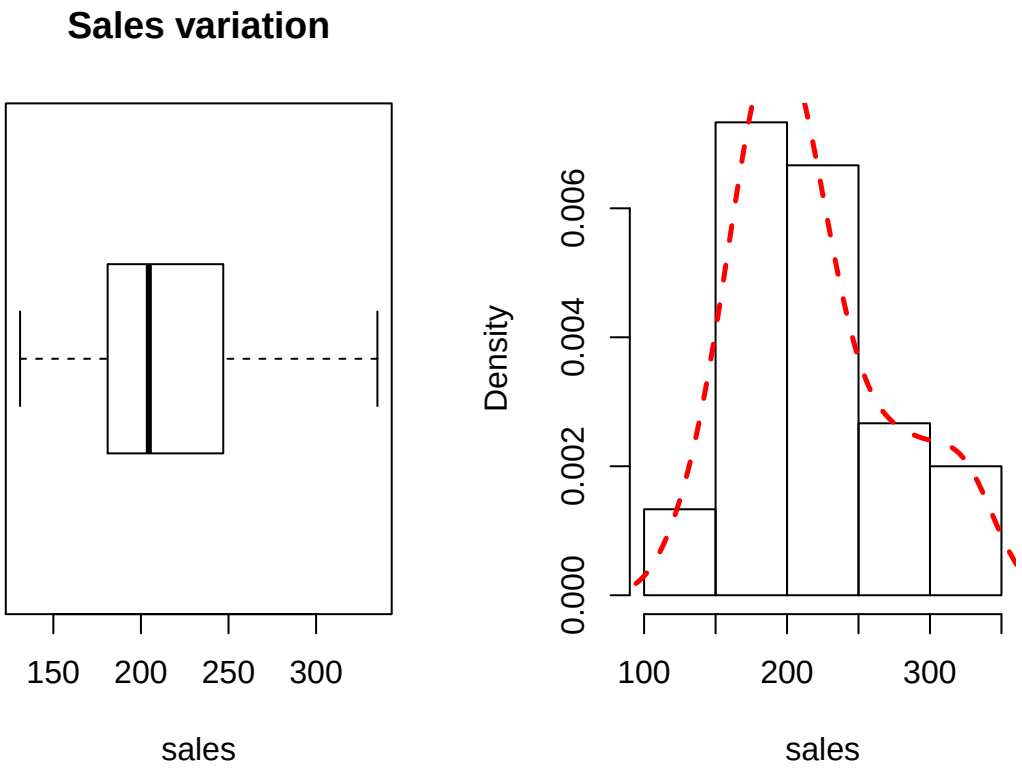
El cuadro resumen anterior nos proporciona los estadísticos descriptivos de la base de datos. Por ejemplo, el valor medio de las ventas es 216.7 unidades, el valor mínimo es de 131, y el valor máximo es 335. Por favor, ignore las estadísticas de la "ad_type" existe ya que es una variable categórica.

Podemos explorar más a fondo la distribución de la los datos de las ventas visualizándolos de la siguiente manera:

```
#exploración datos

#Fijar dos gráficos por fila
par(mfrow = c(1,2))

# boxplot: comprobar la presencia de observciones distorsionantes
boxplot(df$sales,horizontal = TRUE, xlab="sales", main="Sales variation")
# histogram: comprobar la distribución de las variables
hist(df$sales,main="",xlab="sales",prob=T)
lines(density(df$sales),lty="dashed",lwd=2.5,col="red")
```



```r
par(mfrow = c(1,1))
```

No encontramos valores atípicos en el gráfico boxplot de arriba y la distribución de los datos de ventas es más o menos normal. No es necesario aplicar una limpieza más profunda a los datos.

Análisis de la Eficacia del Anuncio

El equipo de marketing quiere averiguar qué anuncio tiene mayor eficacia en la promoción de las ventas, uno utiliza el tema de la producción natural y el otro, el de cuidar de la salud de la familia. Así que se puede colocar el mejor en todas las tiendas de la cadena ABC después del periodo de prueba.

Para averiguar el mejor anuncio, podemos calcular y comparar la media de las ventas con los dos tipos de anuncios

```r
#Análisis de la eficacia del anuncio
#Dividimos la base de datos en dos, según el tipo de anuncio

sales_ad_nature = subset(df,ad_type==0)
sales_ad_family = subset(df,ad_type==1)
#calculate the mean of sales with different ad_type
 mean(sales_ad_nature$sales)
```

```
## [1] 186.6667
```

```r
mean(sales_ad_family$sales)
```

```
## [1] 246.6667
```

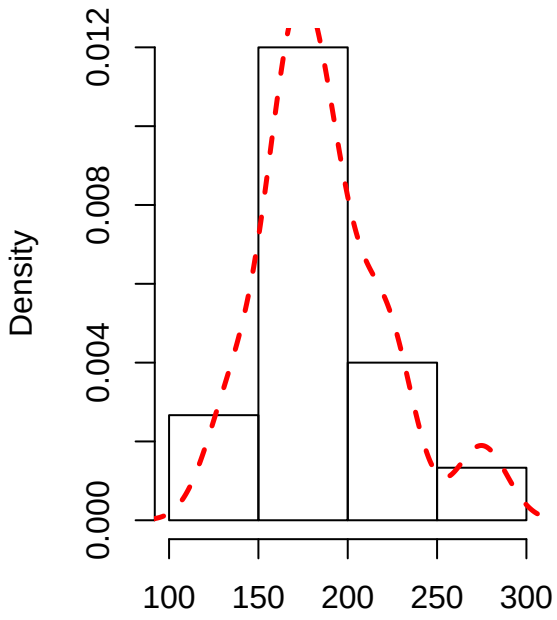
La media de las ventas con el tema de la producción natural tiene una media de 187; la media de las ventas con el tema cuidar de la salud de la familia es de aproximadamente de 247. Parece que este último es mejor. Sin embargo, esto es sólo la conclusión basada en la muestra con sólo 30 observaciones seleccionadas al azar. Para saber qué tan probable es que la conclusión sea correcta para toda la población, es necesario hacer pruebas estadísticas - *t-test* de dos muestras.

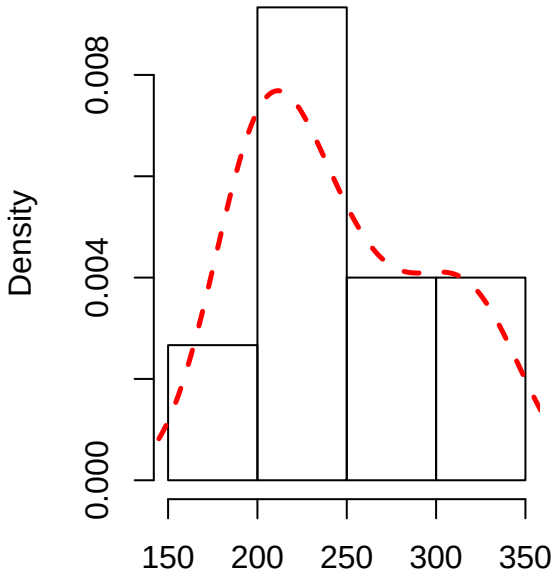
Antes de la realización de las pruebas de la t es importante comprobar las suposiciones de las pruebas t, que asumen que las observaciones tienen una distribución normal e independiente. De lo contrario, los resultados de las pruebas t no son válidos. Las observaciones son independientes, ya que se tomaron muestras al azar. Vamos a comprobar la normalidad por el trazado de las formas de la distribución de los dos grupos de datos de ventas.

```r
#test de diferencias
#Fijar dos gráficos por fila
par(mfrow = c(1,2))

# histogram: explorar normalidad
hist(sales_ad_nature$sales,main="",xlab="Nature production ad",prob=T)
lines(density(sales_ad_nature$sales),lty="dashed",lwd=2.5,col="red")

hist(sales_ad_family$sales,main="",xlab="Family health ad",prob=T)
lines(density(sales_ad_family$sales),lty="dashed",lwd=2.5,col="red")
```






```
par(mfrow = c(1,1))
```

podemos ver que las formas están distribuidas más o menos normalmente. También podemos comprobar la normalidad con la Prueba de Shapiro-Wilk de esta forma:

```
#test de normalidad de las variables
shapiro.test(sales_ad_nature$sales)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sales_ad_nature$sales
## W = 0.9426, p-value = 0.4155
```

```
shapiro.test(sales_ad_family$sales)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sales_ad_family$sales
## W = 0.8974, p-value = 0.08695
```

Los valores p de las pruebas de Shapiro-Wilk son mayores que 0.05, así que no hay pruebas sólidas para rechazar la hipótesis nula de que los dos grupos de datos de ventas se distribuyen normalmente.

Ahora podemos llevar a cabo la prueba de la t ya se cumplen los supuestos de la prueba datos:

```
#Test de diferencias de medias
t.test(sales_ad_nature$sales,sales_ad_family$sales)
```

```
##
##  Welch Two Sample t-test
##
## data:  sales_ad_nature$sales and sales_ad_family$sales
## t = -3.7515, df = 25.257, p-value = 0.0009233
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -92.92234 -27.07766
## sample estimates:
## mean of x mean of y
##  186.6667  246.6667
```

hipótesis alternativa: la verdadera diferencia de medias no es igual a 0 El intervalo de confianza del 95 por ciento: (-92.92234, -27.07766) Estimaciones de la muestra: media de x media de y 186.6667 246.6667

El resultado de la prueba t anterior nos indica que:

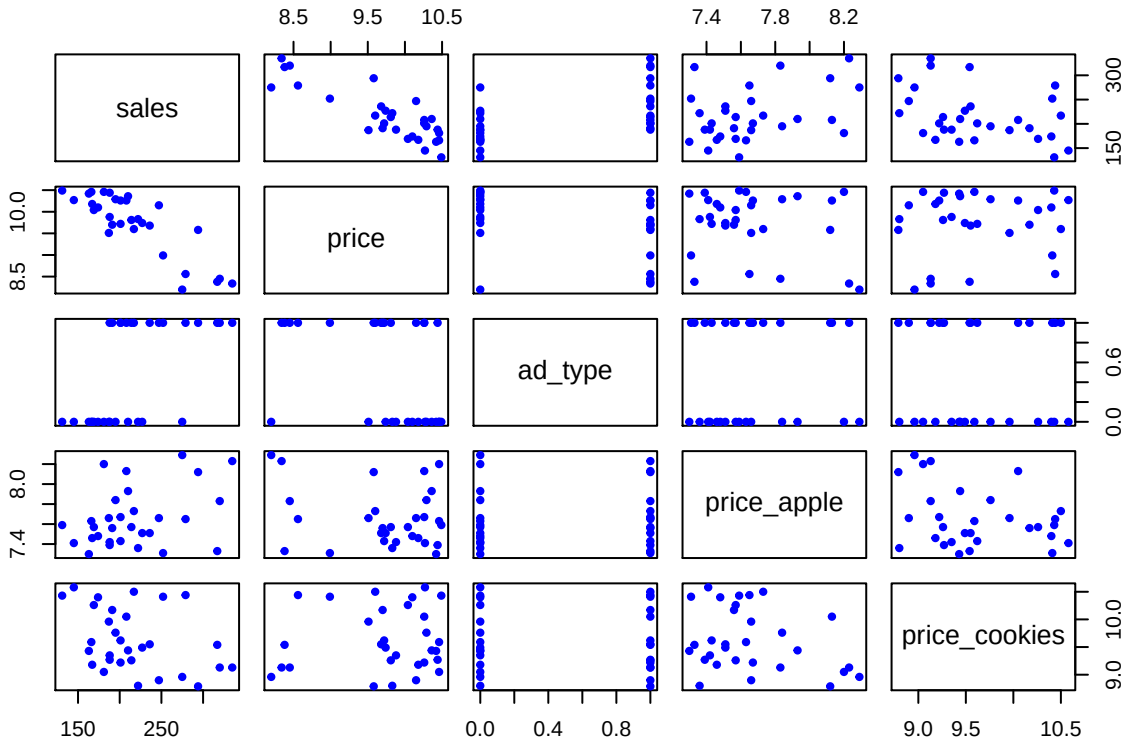
. Tenemos una fuerte evidencia para decir que las medias poblacionales de las ventas con los dos tipos de anuncios son diferentes debido a que el valor de p de la prueba de la t es muy pequeño; . Con una confianza del 95%, se puede estimar que la media de las ventas con el tema de la producción natural de anuncio son menores que las ventas con el tema cuidar de la salud de la familia. . Así que la conclusión es que el anuncio con el tema del cuidado de la salud de la familia es mejor.

Análisis de los impulsores de las ventas y la elasticidad-precio

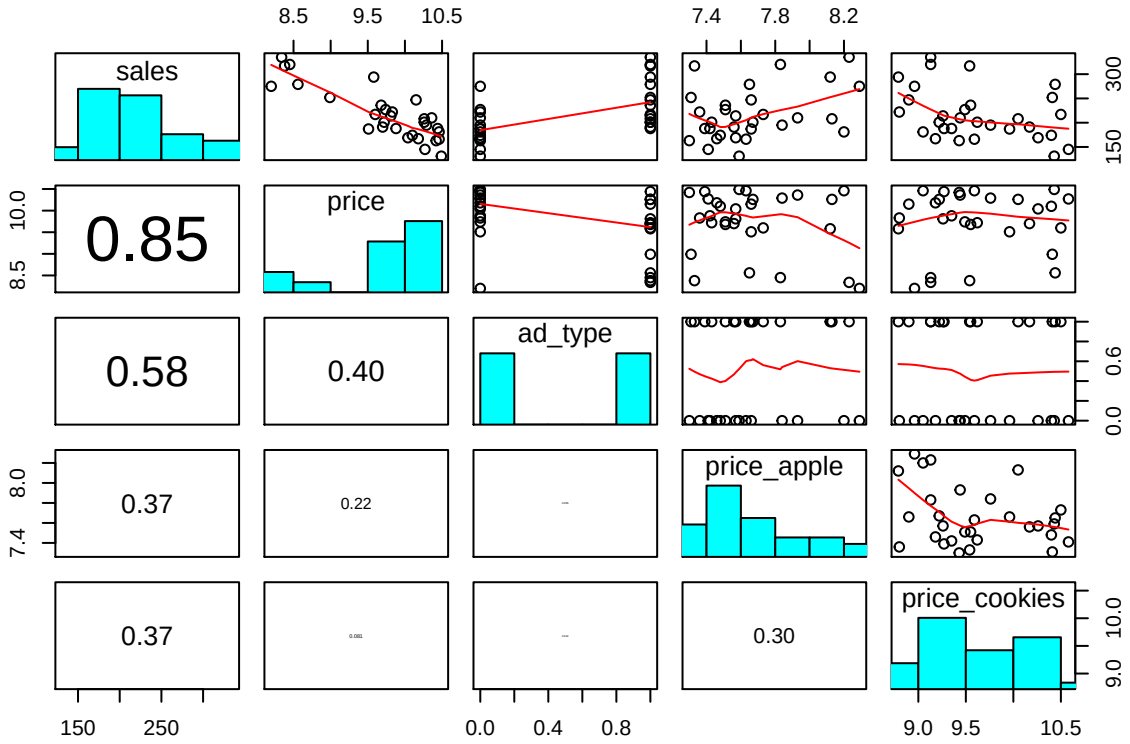
Con la información del conjunto de datos podemos explorar cómo el precio del zumo de uva, el tipo de anuncio, el precio del zumo de manzana, y precio de las galletas influencia las ventas de zumo de uva en una tienda, gracias a un análisis de regresión lineal múltiple. Aquí, "ventas" (sales) es la variable dependiente y las otras variables son variables independientes.

Vamos a investigar la correlación entre las ventas y las otras variables, por la visualización de los coeficientes de correlación de dos endos.

```
# función de las ventas
pairs(df,col="blue",pch=20)
```




```
pairs20x(df)
```



Los coeficientes de correlación entre las ventas y los precios, *ad_type*, *price_apple* y *price_cookies* son de 0.85, 0.58, 0.37, y 0.37, respectivamente, lo que significa que todos ellos pueden tener alguna influencia en las ventas, por lo que podemos tratar de añadir toda las variables independientes en el modelo de regresión de la forma siguiente:

```r
sales.reg<-lm(sales~price+ad_type+price_apple+price_cookies,df)
summary(sales.reg)
```

```
##
## Call:
## lm(formula = sales ~ price + ad_type + price_apple + price_cookies,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.290 -10.488   0.884  10.483  29.471
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    774.813    145.349   5.331 1.59e-05 ***
## price          -51.239      5.321  -9.630 6.83e-10 ***
## ad_type         29.742      7.249   4.103 0.000380 ***
## price_apple     22.089     12.512   1.765 0.089710 .
## price_cookies  -25.277      6.296  -4.015 0.000477 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.2 on 25 degrees of freedom
## Multiple R-squared:  0.8974, Adjusted R-squared:  0.881
## F-statistic: 54.67 on 4 and 25 DF,  p-value: 5.318e-12
```

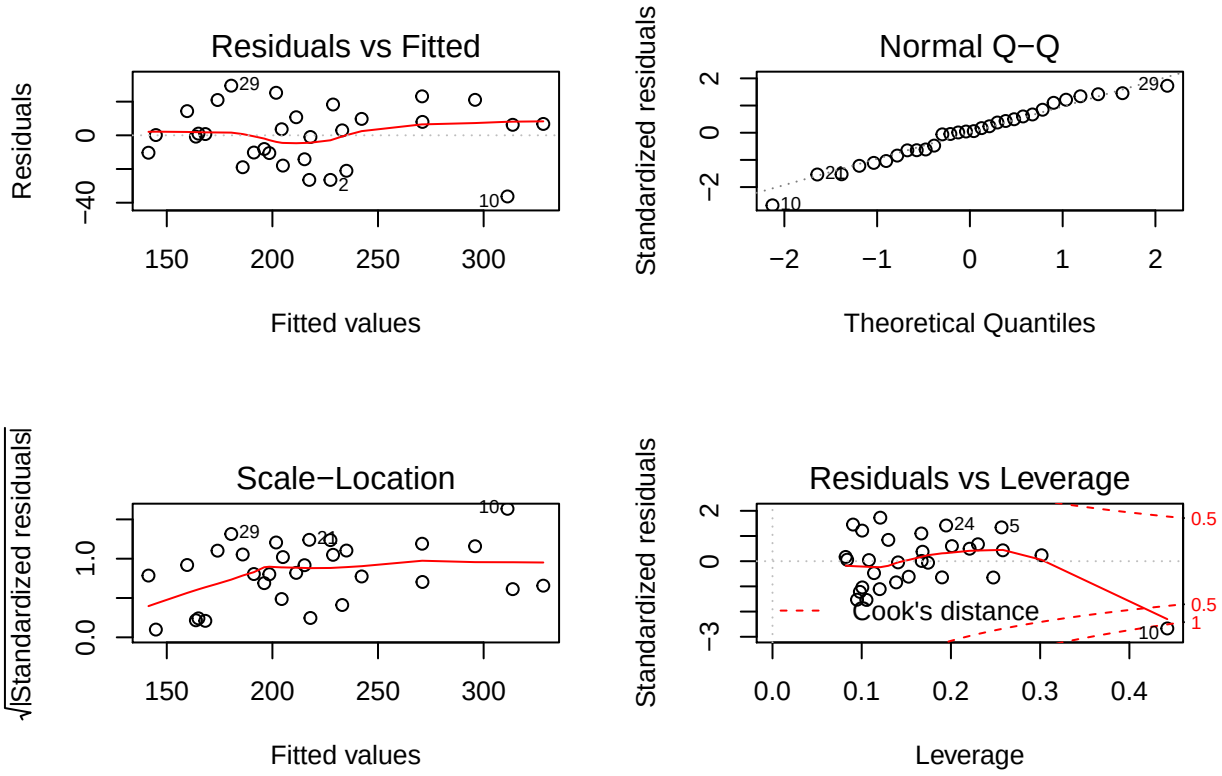
El valor p para las variables *price*, *ad_type*, y *price_cookies* en la última columna del resumen de la estimación es mucho menor que 0,05. Son importantes para explicar las ventas. Estamos seguros de incluir estas tres variables en el modelo.

El valor de p de price_apple es un poco mayor que 0,05, parece que no hay pruebas sólidas de que el precio del zumo de manzana influya en las ventas. Sin embargo, de acuerdo a nuestra experiencia en la vida real, sabemos que cuando el precio de zumo de manzana es menor, los consumidores pueden comprar más zumo de manzana, por lo que las ventas de otros zumos de frutas disminuirán. Así que también podemos agregarla al el modelo para explicar las ventas de zumo de uva.

La R-cuadrado ajustada es 0.881, lo que indica una bondad del ajuste razonable y el 88% de la variación en las ventas se explica por las cuatro variables. El 12% restante se puede atribuir a otros factores o variabilidad inherente. Por favor, tenga en cuenta el R-cuadrado es muy alto aquí, porque el conjunto de datos fue construido a propósito para ilustrar el problema, en lugar de tomarlo de la vida real.

Los supuestos para la regresión del modelo de regresión son que los datos sean aleatorios y que los residuos se distribuyen normalmente y tengan varianza constante. Vamos a comprobar los supuestos visualmente.

```r
#Supuestos del modelo de regresión lineal
# visualizando la relación entre los residuos y otras variables
#fijar los gráficos
par(mfrow=c(2,2))
plot(sales.reg)
```




```
par(mfrow=c(1,1))
```

El gráfico Residuals vs fitted values anterior muestra que los residuos se dispersan alrededor de la línea estimada con un patrón obvio, y el gráfico QQ de normalidad muestra que, básicamente, los residuos se distribuyen normalmente. Se cumplen los supuestos.

Para la regresión múltiple, también es importante comprobar la multicolinealidad entre las variables porque la alta multicolinealidad hará que los coeficientes de las variables independientes sean menos precisos e introducirán grandes errores en las predicciones para la variable dependiente. Podemos investigar la multicolinealidad mostrando los coeficientes de correlación de las variables independientes en pares como lo que hicimos en el comienzo de esta parte. También podemos comprobar la multicolinealidad mediante el siguiente comando en **R**.

```
#controlar la presencia de multicolinearidad
vif(sales.reg)
```

```
##         price       ad_type   price_apple price_cookies
##      1.246084      1.189685      1.149248      1.099255
```

El valor de la prueba VIF para cada variable es cercano a 1, lo que significa que la multicolinealidad entre estas variables es muy baja.

Un modelo de las ventas

Basándose en el análisis anterior, podemos aceptar el resultado de la regresión y construir el modelo de ventas siguiente:

$Sales = 774,81 - 51,24 * Price + 29,74 * ad_{type} + 22,1 * price_{apple} - 25,28 * price_{cookies}$

con el modelo establecido, podemos análisis de la elasticidad-precio (PE) y la elasticidad-precio cruzada (CPE) para predecir las reacciones de los volúmenes de venta frente a cambios en el precio. "La elasticidad precio se define como $\triangle Q/\triangle P$, lo que indica el cambio porcentual en la venas dividida por el cambio porcentual en el precio. La elasticidad-precio cruzada es el cambio porcentual en la cantidad dividida por la variación en el precio de otro producto".

$PE = (\triangle Q/Q)/(\triangle P/P) = (\triangle Q/\triangle P) * (P/Q) = -51,24 * 0,045 = -2,3$

$\triangle Q/\triangle P = -51,24$, el parámetro estimado para la variable *price* $P/Q = 9,738/216,7 = 0,045$, P es el precio, Q es la cantidad de ventas.

El PE señala que un 10% de disminución en el precio va a aumentar las ventas en un 23%, y viceversa.

Calculemos ahora la elasticidd cruzada, CPE, entre el zumo de manzana y las galletas para analizar la forma en que el cambio en el precio del zumo de manzana y en el de las galletas influye en las ventas de zumo de uva.

$CPE_a = (\triangle Q/\triangle P_a) * (P_a/Q) = 22,1 * (7,659/216,7) = 0,78$

$CPE_c = (\triangle Q/\triangle P_c) * (P_c/Q) = -25,28 * (9,622/216,7) = -1,12$

La CPEa(pple) indica que el 10% de disminución en el precio del zumo de manzana disminuirá las ventas de zumo de uva en un 7,8%, y viceversa. Así que el zumo de uva y zumo de manzana son sustitutos. Las CPEc(ookies) indica que 10% de disminución en el precio de las cookies aumentará las ventas en un 11,2%, y viceversa. Así que el zumo de uva y las galletas son productos complementarios. Coloque los dos productos juntos y es probable que aumente las ventas de ambos. También podemos saber que las ventas aumentan 29.74 unidades cuando se utiliza el anuncio del tema 'cuidar la salud de la familia' (*ad_type = 1*).

Precios óptimos y predicción de ventas

En general, las empresas quieren alcanzar beneficios más altos y no sólo mayor cantidad de ventas. Así que, ¿cómo se establecer el precio óptimo para el nuevo zumo de uva que maximiza el beneficio basado en el conjunto de datos recolectados en el periodo de prueba y el modelo de regresión anterior?

Para simplificar la cuestión, podemos fijar el valor del *ad_type = 1*, de *price_apple = 7,659* (valor medio), y el de *price_cookies = 9,738* (valor medio).

El modelo se simplifica de la siguiente manera:

$$Sales = 774,81 \check{} 51,24 * price + 29,74 * 1 + 22,1 * 7,659 \check{} 25,28 * 9,738$$

$$Sales = 772,64 - 51,24 * Precio$$

Asuma que el costo marginal (C) por unidad de zumo de uva es de 5. Entonces podemos calcular la ganancia (Y) por la siguiente fórmula:

$$Y = (price \check{} C) * Sales = (price \check{} 5) * (772,64 \check{} 51,24 * price)$$
$$Y = \check{} 51,24 * price^2 + 1028,84 * price \check{} 3863,2$$

Para obtener el precio óptimo que maximiza Y, podemos usar la siguiente función de R

```
# Decisiones óptimas

f = function(x) -51.24*x^2 + 1028.84*x - 3863.2
optimize(f,lower=0,upper=20,maximum=TRUE)
```

```
## $maximum
## [1] 10.03942
##
## $objective
## [1] 1301.28
```

El precio óptimo es 10.04; la ganancia máxima será de 1301 de acuerdo con el resultado anterior. En realidad, podemos establecer razonablemente el precio para ser 10 ó 9.99. Además podemos utilizar el modelo para predecir las ventas, dado un precio de 10.

```
# predicir ventas
inputData <- data.frame(price=10,ad_type=1,price_apple=7.659,price_cookies=9.738)
predict(sales.reg,inputData,interval="p")
```

```
##        fit      lwr      upr
## 1 215.1978 176.0138 254.3817
```

La previsión de ventas será 215 unidades con un rango variable entre 176 y 254 con un 95% confianza en la tienda experimental. Sobre la base de la previsión y otros factores, la empresa ABC puede preparar el inventario de todas sus tiendas después del periodo de prueba.

Resumen

En este artículo, utilizando el lenguaje de código abierto R, presentamos cómo probar las diferencias de eficacia entre los diferentes tipos de anuncios; cómo analizar la elasticidad precio y elasticidad-precio cruzada de un producto; y cómo establecer el precio óptimo para maximizar el beneficio y luego pronosticar las ventas dado el precio.