Dominick’s Finer Foods

Duvan Santiago Castro - Miguel Ángel Malagón

2022-10-04

# Introducción

A lo largo de las siguientes secciones encontrará los diferentes modelos estimados y sus resultados correspondientes, que permitirá entender la demanda de cerveza de la cadena de restaurantes Dominick’s Finer Foods (DFF). Para dicha interpretación nos centraremos en 2 tipos de cervezas con sus respectivos precios.

Inicialmente realizaremos una serie de filtros que nos permitan seleccionar una muestra de 200 observaciones para el desarrollo de los modelos.

# Primeros pasos - Alistamiento datos

Para la extracción de los datos, vamos a utilizar la base de datos wber, provenientes de (DFF) y les aplicaremos los siguientes filtros:

* Store = 109: Nos centraremos en la tienda 109
* Precio > 0
* Unidades > 0
* Seleccionar 2 tipos de marcas:
  + BUDWEISER BEER
  + BUDWEISER LIGHT BEER

#intalacion de las librerías necesarias  
  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4  
## ✔ tibble 3.1.7 ✔ stringr 1.4.0  
## ✔ tidyr 1.2.0 ✔ forcats 0.5.1  
## ✔ readr 2.1.2

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

datos\_cervezas <- read.csv(file.choose())

beer <- filter(datos\_cervezas,STORE==109,(MOVE>0& PRICE>0),(UPC==1820000016|UPC==1820000106))

# Cambio los datos   
  
beer$UPC <- replace(beer$UPC, beer$UPC==1820000016, 0)  
beer$UPC <- replace(beer$UPC, beer$UPC==1820000106, 1)  
  
# Seleccionar las variables necesarias   
  
beer <- select(beer,unidades="MOVE",precio="PRICE",marca="UPC")  
head(beer)

## unidades precio marca  
## 1 18 3.49 0  
## 2 20 3.49 0  
## 3 24 3.49 0  
## 4 10 3.59 0  
## 5 26 3.59 0  
## 6 14 3.59 0

una vez realizados los filtros necesarios, hacemos un cambio los tipos de datos con el fin de establecer el siguiente modelo:

$$

Demanda=*{0}+*{1}precio+\_{2}marca\ marca:=\ :marca=0: marca=1:

$$

# Modelo - Estimación

Para la estimación del modelo planteado anteriormente vamos se define un vector para la demanda que representa la cantidad de unidades vendidas de cada producto, así mismo, se define una matriz de diseño que contiene, una columna de unos asociada al intercepto, otra columna de precios y tipo de marca definida como una variable Dummy.

En los siguientes apartados vamos a regresar diferentes modelos con el fin de revisar cual de los 3 ( Lin - Lin), (log -log) o (Log -Lig) se ajusta mejor a los datos.

## Modelo Lin - Lin

El primer modelo que vamos a revisar es un modelo lineal de forma matricial.

$$ Y=X+e\ Y= X=\ \

$$ Una vez realizada la estimación, los coeficientes asociados al modelo principal previamente mencionado, son los siguientes:

y <- as.matrix(beer$unidades)  
x <- as.matrix(cbind(1,beer$precio,beer$marca))  
colnames(x) <-c("b0","precio","marca")   
betas <- (solve(t(x)%\*%x)%\*%t(x)%\*%y)  
betas

## [,1]  
## b0 59.336716  
## precio -12.902995  
## marca -3.167463

quedando de esta manera el siguiente modelo:

$$ =59,33-12,90*Precio-3.16*Marca

$$ *Análisis:* - La demanda de Cervezas de la tienda 109 cuyas variables precio y marca, muestran una serie de comportamientos donde por un aumento el precio de 1und, la demanda disminuye 12,90 y cuando es 1 es decir “Budweiser beer”, la demanda se reduce un -3,16. de otro modo es cero.

## Modelo Log - Lin

Ahora vamos a plantear un modelo Log-Lin y revisaremos los coeficientes asociados a cada variable explicativa de la demnada de cervezas de la tienda 109 de (DFF)

$$ ln(consumo)=*{0}+*{1}precio+\_{2}marca+e\

$$ Dado que estamos trabajando las regresiones de forma matricial, antes de correr la regresión la variable consumo, representada en unidades vendidas se le hace una tranformación logaritmica y luego si se corre la respectiva regresión.

y <- as.matrix(beer$unidades)  
x <- as.matrix(cbind(1,beer$precio,beer$marca))  
colnames(x) <-c("b0","precio","marca")   
lny <- log(y)  
betas2 <- (solve(t(x)%\*%x)%\*%t(x)%\*%lny)  
betas2

## [,1]  
## b0 8.3438034  
## precio -1.6152663  
## marca -0.4562355

interpretación:

## Modelo Log - Log

Ahora vamos a plantear un modelo Log-Log en donde nos aplicaremos una transformación logaritmica a los precios, pero no la a la variable dummmy dado que el log(0) no existe, por lo tanto sera un modelo log-log en el par de conusmo y precio, como se muestra en el siguiente planteamiento.

$$ ln(consumo)=*{0}+*{1}ln(precio)+\_{2}marca+e\

$$
Dado que las regresiones se estan trabajando de manera matricial, la matriz de diseño X y la matriz de Y "Consumo" se trabajaria de la siguiente manera:
$$

X=

, Y=

$$ Una vez realizamos la regresión, nos da los siguientes coeficientes asociados a cada variable.

y <- as.matrix(beer$unidades)  
logx <- as.matrix(cbind(1,log(beer$precio),beer$marca))  
colnames(x) <-c("b0","precio","marca")   
lny <- log(y)  
  
betas3 <- (solve(t(logx)%\*%logx)%\*%t(logx)%\*%lny)  
betas3

## [,1]  
## [1,] 10.3112304  
## [2,] -6.0767715  
## [3,] -0.4560181

Interpretación:

# Matriz de Varianza Covarianza para los estimadores

Teniendo en cuenta lo anterior, notamos que hay diferentes interpretaciones para cada modelo, pero seria conveniente también conocer la variación de dichos coeficientes para cada modelo, por esta razón vamos a calcular la matriz de varianzas-covarianzas para los diferentes estimadores de cada modelo, para eso utilizaremos la siguiente formula:

$$ =2(XtX)\ \ ^2= T=observaciones,K=parámetros

$$ Teniendo en cuenta que la estimación de la varianza se realiza apartir de los errores estimados, en cada modelo vamos a realizar dicho procedimiento para ajustarlo al tipo de modelo.

## Modelo Lin - Lin

Antes de calcular la matriz de varianza covarianza vamos a estimar los errores

y <- as.matrix(beer$unidades)  
x <- as.matrix(cbind(1,beer$precio,beer$marca))  
colnames(x) <-c("b0","precio","marca")   
betas <- (solve(t(x)%\*%x)%\*%t(x)%\*%y)  
y\_gorro <- x%\*%betas  
e\_gorro1 <- y-y\_gorro  
  
# Calcular sigma^2  
t <- as.numeric(nrow(y))  
k <- as.numeric(ncol(x))  
var1 <- as.numeric((t(e\_gorro1)%\*%e\_gorro1)/(t-k))  
var1

## [1] 15.56234

Una vez que tenemos calcula la varianza estimada, vamos a calcular nuestra matriz var\_cov

x <- as.matrix(cbind(1,beer$precio,beer$marca))  
var\_cov1 <- var1\*(solve(t(x)%\*%x))  
var\_cov1

## [,1] [,2] [,3]  
## [1,] 20.42811843 -5.284351036 -0.064441587  
## [2,] -5.28435104 1.371728987 -0.001718266  
## [3,] -0.06444159 -0.001718266 0.143446045

## Modelo Log - Lin

Ahora tenermos que tener encuenta que tipo de modelo estamos estimando para así poder contruir nuestra matriz de var-cov, en donde a diferencia del anterior modelo, tenemos que hacer la siguiente modificación:

$$
ln(consumo)=\beta\_{0}+\beta\_{1}precio+\beta\_{2}marca\\
\widehat{ln(consumo)}=X\hat{\beta} \quad \implies
\widehat{consumo}=e^{\widehat{ln(consumo)}}
$$

y <- as.matrix(beer$unidades)  
x <- as.matrix(cbind(1,beer$precio,beer$marca))  
colnames(x) <-c("b0","precio","marca")   
lny <- log(y)  
betas2 <- (solve(t(x)%\*%x)%\*%t(x)%\*%lny)  
  
log\_y\_gorro2 <- x%\*%betas2  
y\_gorro2 <- exp(log\_y\_gorro2)  
e\_gorro2 <- y-y\_gorro2  
  
# Calcular sigma^2  
t <- as.numeric(nrow(y))  
k <- as.numeric(ncol(x))  
var2 <- as.numeric((t(e\_gorro2)%\*%e\_gorro2)/(t-k))  
var2

## [1] 16.48976

x <- as.matrix(cbind(1,beer$precio,beer$marca))  
var\_cov2 <- var2\*(solve(t(x)%\*%x))  
var\_cov2

## [,1] [,2] [,3]  
## [1,] 21.6455042 -5.599264704 -0.068281896  
## [2,] -5.5992647 1.453475298 -0.001820664  
## [3,] -0.0682819 -0.001820664 0.151994516

## Modelo Log - Log

y <- as.matrix(beer$unidades)  
logx <- as.matrix(cbind(1,log(beer$precio),beer$marca))  
colnames(x) <-c("b0","precio","marca")   
lny <- log(y)  
  
betas3 <- (solve(t(logx)%\*%logx)%\*%t(logx)%\*%lny)  
  
log\_y\_gorro3 <- logx%\*%betas3  
y\_gorro3 <- exp(log\_y\_gorro3)  
e\_gorro3 <- y-y\_gorro3  
  
# Calcular sigma^2  
t <- as.numeric(nrow(y))  
k <- as.numeric(ncol(x))  
var3 <- as.numeric((t(e\_gorro3)%\*%e\_gorro3)/(t-k))  
var3

## [1] 16.53493

logx <- as.matrix(cbind(1,log(beer$precio),beer$marca))  
var\_cov3 <- var3\*(solve(t(logx)%\*%logx))  
var\_cov3

## [,1] [,2] [,3]  
## [1,] 37.65849535 -27.885747058 -0.065219265  
## [2,] -27.88574706 20.690605483 -0.007629536  
## [3,] -0.06521927 -0.007629536 0.152411432

# R ajustado de cada especificación

Hasta el momento hemos realizado la estimación de los diferentes modelos, nos centramos en conocer sus diferentes matrices de varianzas - Covarianzas, pero no hemos revisado que tanto las variables explicativas, explican la variabilidad del consumo de cervezas para las marcas seleccionadas.

Para conocer dicha que tan ajustado es del modelo en términos de variabilidad utilizaremos el R^2 ajustado con la siguiente formula:

$$ R\_{ajustado}2=1-(1-R2)\*()\ R^2=1-

$$ Dado que tenemos diferentes modelo vamos a calcularlo para cada uno de ellos.

R2ADJ <- function(y,e\_gorro,n,k,R22){  
 y <- as.matrix(y)  
 e\_gorro <- as.matrix(e\_gorro)  
 mean\_y <- sum(y)/length(y)  
 R\_2 <- 1- (sum((e\_gorro)^2)/(sum((y-mean\_y)^2)))  
 R2ADJ <- 1-(((n-1)/(n-k-1))\*(1-R\_2))   
 return(R2ADJ)  
}  
n <- nrow(beer)  
k <- ncol(x)

## Modelo Lin - Lin

El primer modelo que vamos a revisar es el de Lin-Lin en donde podemos evidenciar que la variabilidad del modelo explica en 0.3033 la variabilidad del consumo.

y <- as.matrix(beer$unidades)  
x <- as.matrix(cbind(1,beer$precio,beer$marca))  
R2ADJ(y = y,e\_gorro = e\_gorro1,n = n,k = k)

## [1] 0.3033914

## Modelo Log - Lin

Por otro lado, tenemos el modelo LOg-lin en donde únicamente aplicamos una transformación logarítmica a la variable del consumo, una vez calculado el R^2 ajustado notamos que el modelo explica en *0.261878* la variabilidad del consumo, cabe resaltar que antes de calcular el consumo estimado para el cálculo de los errores se aplica exponente al logaritmo del consumo para dejarlo en términos lineales.

y <- as.matrix(beer$unidades)  
x <- as.matrix(cbind(1,beer$precio,beer$marca))  
  
R2ADJ(y = y,e\_gorro = e\_gorro2,n = n,k = k)

## [1] 0.261878

## Modelo Log - Log

Finalmente, procedemos a calcular el R^2 ajustado del modelo Log - Log, el cual nos dice que dicho modelo explica en 0.2598559 la variabilidad del consumo, cabe resaltar nuevamente que antes de calcular el consumo estimado para el cálculo de los errores se aplica exponente al logaritmo del consumo para dejarlo en términos lineales.

y <- as.matrix(beer$unidades)  
x <- as.matrix(cbind(1,beer$precio,beer$marca))  
  
R2ADJ(y = y,e\_gorro = e\_gorro3,n = n,k = k)

## [1] 0.2598559

Teniendo en cuenta los resultados anteriores, podemos evidenciar que el modelo que mejor explica la variabilidad del consumo, es el modelo Lin - Lin.

# Distancia de Cook

Aunque la muestra es representativa y aleatoria, pueden existir valores que afecten el cálculo de los betas, por esta razón vamos a utilizar el teorema de Cook o distancia de cook, que nos mide la influencia de un dato (i) sobre los coeficientes.

$$ D\_{i}=\ MCE=\_{t=1}T(y\_t-)2

$$

cook <- function(y,x){  
   
 x <- as.matrix(x)  
 y <- as.matrix(y)  
 coefi <- solve(t(x)%\*%x)%\*%t(x)%\*%y  
 distancia <- c()  
 t <- length(y)  
 y\_gorro <- x%\*%coefi  
 k <- ncol(x)  
 MCE <- ((1/t-k)\*sum((y-y\_gorro)^2))  
   
 for (i in 1:n) {  
  
 xc <- x[-i,]  
 yc <- y[-i,]  
 coefi2 <- solve(t(xc)%\*%xc)%\*%t(xc)%\*%yc  
 distancia[i] <- (t((coefi2-coefi))%\*%t(x)%\*%x%\*%(coefi2-coefi))/(k\*MCE)  
 xc <- x  
 yc <- y   
 }  
   
 tabla <- cbind(y,x,distancia)  
 return(head(tabla))  
   
}  
y <- as.matrix(beer$unidades)  
x <- as.matrix(cbind(1,beer$precio,beer$marca))  
cook(y =y,x = x)

## distancia  
## [1,] 18 1 3.49 0 -3.772995e-06  
## [2,] 20 1 3.49 0 -8.963271e-06  
## [3,] 24 1 3.49 0 -2.597714e-05  
## [4,] 10 1 3.59 0 -1.636821e-06  
## [5,] 26 1 3.59 0 -3.036151e-05  
## [6,] 14 1 3.59 0 -1.747195e-07

Una vez que calculamos las diferentes distancias para cada observación, se revisa a través de un filtro las distancias mayores a uno, en donde nos dio que ningún dato es mayor a uno, por lo tanto los datos que tenemos no son influyentes a la hora de estimar los betas.

# Pronóstico

Finalmente a través de nuestro modelo Lin-Lin vamos pronosticar la demanda que se tendrían con un precio de 2/3, sin embargo debemos tener en cuenta que la otra variable que explica la demanda es un avariable Dummy con la marca de cerveza en donde:

$$
marca=0:\text{Budweiser beer} \\\quad marca=1:\text{Budweiser light beer}
$$

$$ =59,33-12,90*Precio-3.16*Marca\ \ \ =59,33-12,90*(2/3)-3.16*(0)\ =50.73\ \ \ =59,33-12,90*(2/3)-3.16*(1)\ =47.57\

$$

A lo que podemos concluir que ante un precio alto la demanda disminuye cuando se trata de la cerveza tipo Budweiser light beer, a diferencia de la otra cerveza.