Economía Computacional: Tarea 1

Isidoro Garcia

2023

```
library(tidyverse)
library(data.table)
library(RCT)
library(knitr)
library(lfe)
library(kableExtra)
library(broom)
```

En esta tarea pondrán en práctica los conceptos de High Dimensional Inference y Regresión. La base de datos muestra las compras de helados Ben & Jerry. Cada fila es una compra. Cada columna es una característica del helado comprado o de la persona que compró.

Limpieza de datos

Carga los datos en BenAndJerry.csv.

```
# Carga la base de datos
base<-read.csv('BenAndJerry.csv')
```

1. ¿Cuales son las columnas de la base? Muestra una tabla con ellas

```
var<-data.frame("Variables"=names(base))
kbl(list(var[1:17,],var[18:34,]),booktabs = T)
```

2. ¿A qué nivel está la base? Esto es, cuál es la variable que define la base de manera única. Si no la hay, crea una y muestra que es única a nivel de la base (Muestra el código)

3. ¿Qué variables tienen valores vacíos? Haz una tabla con el porcentaje de vacíos para las columnas que tengan al menos una observación vacía

```
nas<-apply(base, 2, function(x) sum(is.na(x))/nrow(base)*100)
nas[nas!=0]<-paste(round(nas[nas!=0],2),"%")
kbl(data.frame("Porcentaje_NAs"= nas[nas!=0]),booktabs = T) %>% kable_styling(position = "center")
```

x	x
quantity	female_head_education
price_paid_deal	marital_status
price_paid_non_deal	male_head_occupation
coupon_value	female_head_occupation
promotion_type	household_composition
size1_descr	race
flavor_descr	hispanic_origin
formula_descr	region
household_id	scantrack_market_identifier
household_size	fips_state_code
household_income	fips_county_code
age_of_female_head	type_of_residence
age_of_male_head	kitchen_appliances
age_and_presence_of_children	tv_items
male_head_employment	female_head_birth
female_head_employment	male_head_birth
male_head_education	household_internet_connection

	Porcentaje_NAs
promotion_type	59.07 %
female_head_occupation	10.32 %
scantrack_market_identifier	18.51 %
tv_items	0.15 %

4. Haz algo con los valores vacíos (Se deben reemplazar por algún valor? Eliminar de la base?). Justifica tu respuesta.

```
base<-na.omit(base)
attach(base)
```

5. Muestra una tabla de estadisticas descriptivas de la base. Esta debe tener cada columna númerica con algunas estadísticas descriptivas (N, media, min, p05, p25, p50, p75, p90, p95, max).

```
res<-summary_statistics(base)
kbl(res,booktabs = T,digits = 2,format.args = list(big.mark=",")) %>%
  kable_styling(full_width = T,font_size = 3)
```

variable	mean	n	0	0.05	0.1	0.25	0.5	0.75	0.9	0.95	1
quantity	1.32	6,986	1	1.00	1.0	1	1.0	1.00	2	2.00	12.00
price_paid_deal	4.25	6,986	1	2.34	2.5	3	3.5	4.49	7	8.78	28.88
price_paid_non_deal	0.00	6,986	0	0.00	0.0	0	0.0	0.00	0	0.00	0.00
coupon_value	0.37	6,986	0	0.00	0.0	0	0.0	0.25	1	2.00	12.95
promotion_type	1.44	6,986	1	1.00	1.0	1	1.0	2.00	3	3.00	4.00
household_id	13,063,215.23	6,986	2,000,358	2,039,847.00	2,074,972.0	8,051,970	8,286,409.0	30,057,669.50	30,270,061	30,348,658.00	30,438,498.00
household_size	2.49	6,986	1	1.00	1.0	2	2.0	3.00	4	5.00	9.00
household_income	21.89	6,986	3	11.00	15.0	18	23.0	26.00	27	29.00	30.00
age_of_female_head	6.38	6,986	1	3.00	4.0	5	7.0	8.00	9	9.00	9.00
age_of_male_head	4.42	6,986	0	0.00	0.0	0	5.0	7.00	8	9.00	9.00
age_and_presence_of_children	7.31	6,986	1	1.00	2.0	6	9.0	9.00	9	9.00	9.00
male_head_employment	2.84	6,986	0	0.00	0.0	0	3.0	3.00	9	9.00	9.00
female_head_employment	4.55	6,986	1	1.00	1.0	3	3.0	9.00	9	9.00	9.00
male_head_education	3.04	6,986	0	0.00	0.0	0	4.0	5.00	6	6.00	6.00
female_head_education	4.51	6,986	1	3.00	3.0	4	5.0	5.00	6	6.00	6.00
marital_status	1.83	6,986	1	1.00	1.0	1	1.0	3.00	4	4.00	4.00
male_head_occupation	5.13	6,986	1	1.00	1.0	1	4.0	8.00	12	12.00	12.00
female_head_occupation	5.50	6,986	1	1.00	1.0	1	3.0	12.00	12	12.00	12.00
household_composition	2.23	6,986	1	1.00	1.0	1	1.0	5.00	5	5.00	8.00
race	1.27	6,986	1	1.00	1.0	1	1.0	1.00	2	3.00	4.00
hispanic_origin	1.96	6,986	1	2.00	2.0	2	2.0	2.00	2	2.00	2.00
region	2.59	6,986	1	1.00	1.0	1	3.0	4.00	4	4.00	4.00
scantrack_market_identifier	21.28	6,986	1	1.00	2.0	9	18.0	32.00	43	48.00	52.00
fips_state_code	25.55	6,986	1	6.00	6.0	9	25.0	37.00	48	53.00	56.00
fips_county_code	76.00	6,986	1	3.00	5.0	21	53.0	97.00	161	191.00	810.00
type_of_residence	2.07	6,986	1	1.00	1.0	1	1.0	2.00	5	6.00	7.00
kitchen_appliances	4.06	6,986	1	1.00	1.0	4	4.0	4.00	7	7.00	8.00
tv_items	1.93	6,986	1	1.00	1.0	1	2.0	3.00	3	3.00	3.00
household_internet_connection	1.19	6,986	1	1.00	1.0	1	1.0	1.00	2	2.00	2.00

6. ¿Hay alguna númerica que en verdad represente una categórica? ¿Cuáles? Cambialas a factor

7. Revisa la distribución de algunas variables. Todas tienen sentido? Por ejemplo, las edades?

8. Finalmente, crea una variable que sea el precio total pagado y el precio unitario

Exploración de los datos

Intentaremos comprender la elasticidad precio de los helados. Para ello, debemos entender:

- La forma funcional base de la demanda (i.e. como se parecen relacionarse q y p).
- Qué variables irían en el modelo de demanda y cuáles no para encontrar la elasticidad de manera 'insesgada'.
- Qué variables cambian la relacion de q y p. Esto es, que variables alteran la elasticidad.

Algo importante es que siempre debemos mirar primero las variables más relevantes de cerca y su relación en:

- Relación univariada
- Relaciones bivariadas
- Relaciones trivariadas

Importante: Las gráficas deben estar bien documentadas (título, ejes con etiquetas apropiadas, etc). Cualquier gráfica que no cumpla con estos requisitos les quitaré algunos puntos.

9. Cómo se ve la distribución del precio unitario y de la cantidad demandada. Haz un histograma.

10. Grafica la $q(p)$. Que tipo de relación parecen tener?

11. Grafica la misma relación pero ahora entre $log(p+1)$ y $log(q+1)$

Usemos la transformación logarítmica a partir de este punto. Grafiquemos la demanda inversa.

12. Grafica la curva de demanda por tamaño del helado. Parece haber diferencias en la elasticidad precio dependiendo de la presentación del helado? (2 pts)

13. Grafica la curva de demanda por sabor. Crea una variable con los 3 sabores más populares y agruga el resto de los sabores como 'otros'. Parece haber diferencias en la elasticidad precio dependiendo del sabor?

Estimación

14. Estima la regresión de la curva de demanda de los helados. Reporta la tabla de la regresión

Algunos tips:

- No olvides borrar la variable que recien creamos de sabores. Incluirla (dado que es perfectamente colineal con flavor), sería una violación a supuesto GM 3 de la regresión.
- No olvides quitar `quantity`, `price_unit`, `price_deal` y otras variables que sirven como identificadora. Tambien quitar `fips_state_code` y `fips_county_code`.
- Empecemos con una regresión que incluya a todas las variables.

Nota: La regresión en `R` entiende que si le metes variables de texto, debe convertirlas a un factor. En algunos otros algoritmos que veremos durante el curso, tendremos que convertir manualmente toda la base a una númerica.

Quitemos las fechas

```
base$female_head_birth<-NULL
base$male_head_birth<-NULL
```

15 (2 pts). Cuales son los elementos que guarda el objecto de la regresión? Listalos. Cual es el F-test de la regresión? Escribe la prueba de manera matemática (i.e. como la vimos en clase). (Tip: `summary(fit)` te arroja algo del F-test)

16. Cuál es la elasticidad precio de los helados Ben and Jerry ? Es significativo? Interpreta el coeficiente

17. Cuántos p-values tenemos en la regresión. Haz un histograma de los p-values.

18 (4pts). Realiza un ajuste FDR a una $q = 0.10$. Grafica el procedimiento (con y sin zoom-in a p-values<0.05). Cuantas variables salían significativas con $\alpha = 0.05$? Cuantas salen con FDR?

Tip: crea el ranking de cada p-value como `resultados %>% arrange(p.value) %>% mutate(ranking = row_number)`

19 (2pts). Repite el ejercicio pero ahora con Holm-Bonferroni. Comparalo vs FDR. En este caso cuantas variables son significativas? Haz la grafica comparativa (solo con zoom-in)