Economía Computacional: Tarea 1

Isidoro Garcia

2021

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.5     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
library(RCT)
library(knitr)
library(lfe)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
library(broom)
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
library(naniar)
```

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

```
columnas <- (as.data.frame(colnames(base)))

kable(columnas, booktabs=T, align = 'c', col.names = c("Columnas"))
```

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

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)

Así como está la base sin niguna modificación el nivel es la compra. Es decir, cada fila representa una transacción realizada por un hogar. Esto lo podriamos modificar para que la unidad sea el hogar o cualquier otra variable.

No hay una variable explícita que identifique cada observación de manera única pero si hay una manera implícita y es el índice de cada fila. Sin embargo, podemos crear una variable que contenga la información del índice de fila.

```
base<- base %>% rowid_to_column("ID")
```

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

```
kable( (base %>% select_if(~sum(is.na(.)) > 0) %>% miss_var_summary()), booktabs=T, align = 'c', col.nam
```

Variable	Cantidad	%
promotion_type	12980	59.0698098
scantrack_market_identifier	4068	18.5127878
female_head_occupation	2267	10.3167380
tv_items	34	0.1547283

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

Pues dependiendo de la cantidad de valores vacíos, de si hay un patrón en los valores vacíos y las características de cada variable podemos proponer una estrategia, por ejemplo imputación o quitar esas observaciones. En este sentido tenemos que realizar un análisis por variable:

promotion_type

```
summary(factor(base$promotion_type))
```

```
##     1     2     3     4  NA's
##  6509  1106  1258   121 12980
```

En esta variable podría ser que los NAs nos indiquen que sencillamente no hubo ninguna promoción (y eso podría explicar que casi el 60% de sus valores sean NAs). En este caso podemos suponer eso e imputarle un valor de 5 o 0 a cada Na.

```
base$promotion_type[is.na(base$promotion_type)] <- 5
```

scantrack_market_identifier

```
summary(factor(base$scantrack_market_identifier))
```

```
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
##  960  609  269  196  122  118  988  559  310  229  259  802  650  468  136  345
##   17   18   19   20   21   22   23   24   25   26   27   28   29   30   31   32
##  442  666  567  424  137  394  187  569  318  332  199  382  350  240  105  337
##   33   34   35   36   37   38   39   40   41   42   43   44   45   46   47   48
##  406  128  102  138  137  472  311  200  392  499  208  404   79  259  117   72
##   49   50   51   52 NA's
##  251  468  403  191 4068
```

En este caso es más complejo porque es muy probable que cada valor corresponda a un producto. En este caso, lo que podríamos hacer es ver si podemos inferir está información de otras varaibles, de lo contrario imputar sería una muy mala idea pues estaríamos creando ruido en nuestra información.

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).

```
# opción a
 stargazer(base,type="text", summary.logical=FALSE,digits=2)
```

```
##
## ================================================================================================
## Statistic                        N        Mean        St. Dev.       Min    Pctl(25)   Pctl(75)     Ma
```

```
## ------------------------------------------------------------------------------------------------
## ID                              21,974   10,987.50     6,343.49        1      5,494.2   16,480.8    21
## quantity                        21,974     1.28          0.73          1         1         1          1
## price_paid_deal                 21,974     1.74          2.55       0.00      0.00       3.34       28
## price_paid_non_deal             21,974     2.45          2.77          0         0        3.6          7
## coupon_value                    21,974     0.16          0.64          0         0          0          1
## promotion_type                  21,974     3.54          1.82          1         1          5          5
## household_id                    21,974 16,612,005.00 11,685,954.00 2,000,358 8,142,253 30,183,891 30,44
## household_size                  21,974     2.46          1.34          1         2          3          5
## household_income                21,974     21.47         5.72          3        17         26          3
## age_of_female_head              21,974     5.51          2.64          0         4          8          9
## age_of_male_head                21,974     4.76          3.15          0         2          8          9
## age_and_presence_of_children    21,974     7.40          2.75          1         6          9          9
## male_head_employment            21,974     3.09          2.78          0         1          3          9
## female_head_employment          21,974     4.20          3.24          0         2          9          9
## male_head_education             21,974     3.32          2.09          0         2          5          6
## female_head_education           21,974     3.98          1.64          0         3          5          6
## marital_status                  21,974     1.94          1.24          1         1          3          4
## male_head_occupation            21,974     5.11          4.17          1         1          8          9
## female_head_occupation          19,707     5.80          4.71       1.00      1.00      12.00       12
## household_composition           21,974     2.57          2.29          1         1          5          8
## race                            21,974     1.24          0.69          1         1          1          4
## hispanic_origin                 21,974     1.95          0.21          1         2          2          2
## region                          21,974     2.63          1.09          1         2          4          4
## scantrack_market_identifier     17,906     23.05         15.05      1.00      11.00      36.00       52
## fips_state_code                 21,974     27.20         15.89          1        12         39          5
## fips_county_code                21,974     79.67         94.00          1        25        101         81
## type_of_residence               21,974     2.08          1.92          1         1          3          7
## kitchen_appliances              21,974     3.81          1.76          1         4          4          9
## tv_items                        21,940     1.93          0.81       1.00      1.00       3.00        3
## household_internet_connection   21,974     1.16          0.36          1         1          1          2
## ------------------------------------------------------------------------------------------------
```

```
# opción b
b<- summary_statistics(base,probs=c(0,0.05,0.25,0.5,0.75,0.9,0.95,1),na.rm=T)
(b<- b %>% mutate_at(vars(-variable),funs(round(.,2))) %>%
  rename(mín=4) %>%
  rename(máx=11))
```

```
## Warning: `funs()` is deprecated as of dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
## # A tibble: 30 x 11
##    variable    mean     n    mín `0.05` `0.25`  `0.5` `0.75`   `0.9` `0.95`
```

```
##    <chr>      <dbl> <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>   <dbl>  <dbl>
##  1 ID       1.10e+4 21974 1.00e0 1.10e3 5.49e3 1.10e4 1.65e4 1.98e+4 2.09e4
##  2 quantity 1.28e+0 21974 1.00e0 1.00e0 1.00e0 1.00e0 1.00e0 2.00e+0 2.00e0
##  3 price_p~ 1.74e+0 21974 0.     0.     0.     0.     3.34e0 4.50e+0 6.86e0
##  4 price_p~ 2.45e+0 21974 0.     0.     0.     2.99e0 3.56e0 4.99e+0 6.86e0
##  5 coupon_~ 1.60e-1 21974 0.     0.     0.     0.     0.     5.00e-1 1.00e0
##  6 promoti~ 3.54e+0 21974 1.00e0 1.00e0 1.00e0 5.00e0 5.00e0 5.00e+0 5.00e0
##  7 househo~ 1.66e+7 21974 2.00e6 2.05e6 8.14e6 8.40e6 3.02e7 3.03e+7 3.04e7
##  8 househo~ 2.46e+0 21974 1.00e0 1.00e0 2.00e0 2.00e0 3.00e0 4.00e+0 5.00e0
##  9 househo~ 2.15e+1 21974 3.00e0 1.10e1 1.70e1 2.30e1 2.60e1 2.70e+1 2.80e1
## 10 age_of_~ 5.51e+0 21974 0.     0.     4.00e0 6.00e0 8.00e0 8.00e+0 9.00e0
## # ... with 20 more rows, and 1 more variable: máx <dbl>
```

6. Hay alguna númerica que en verdad represente una categorica? Cuales? Cambialas a factor

Sí

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

```
# precio total pagado
base <- base %>% mutate(total_price=price_paid_deal+price_paid_non_deal)

# precio unitario
base <- base %>% mutate(unit_price= (total_price)/quantity)
```

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

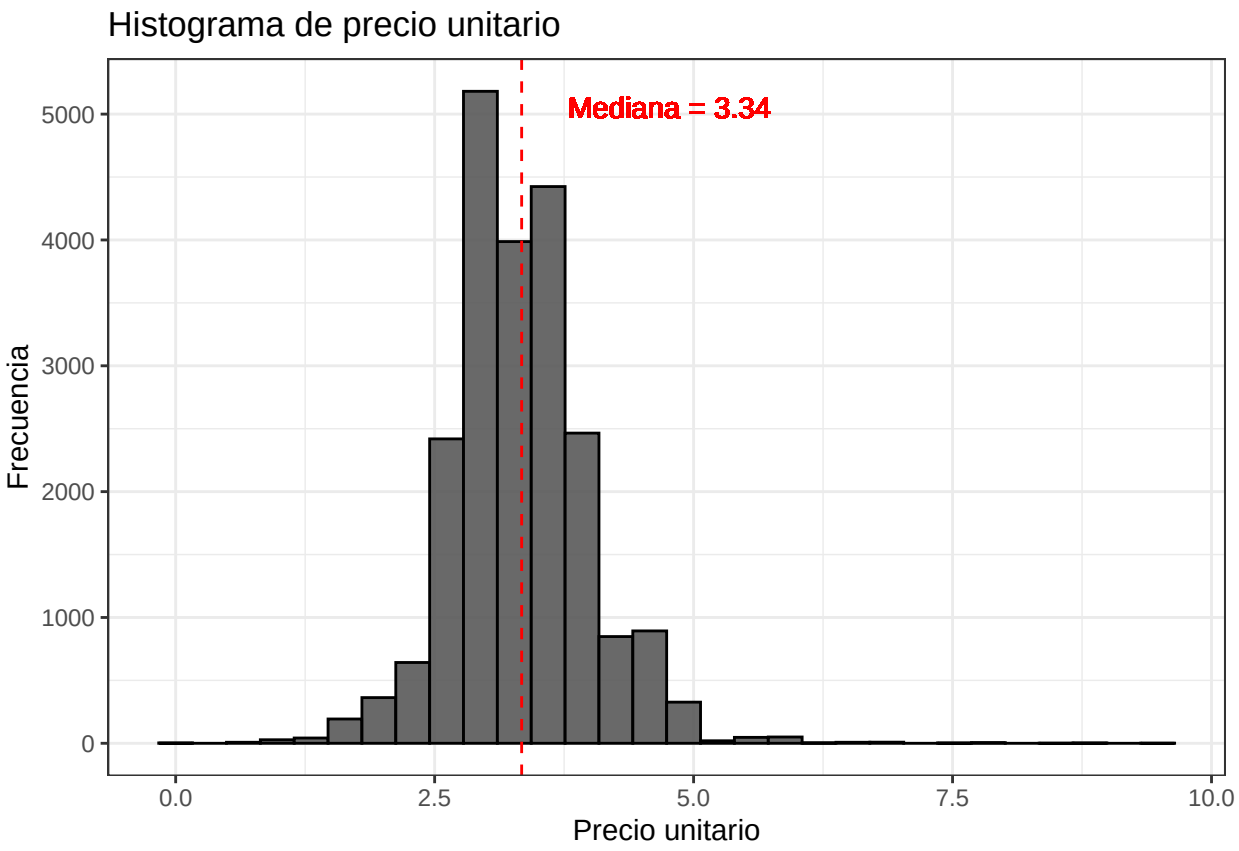
Importante: Las gráficas deben estar bien documentadas (título, ejes con etiquetas apropiadas, etc). Cualquier gráfica que no cumpla con estos requisitos les quitaré algunos puntos.

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

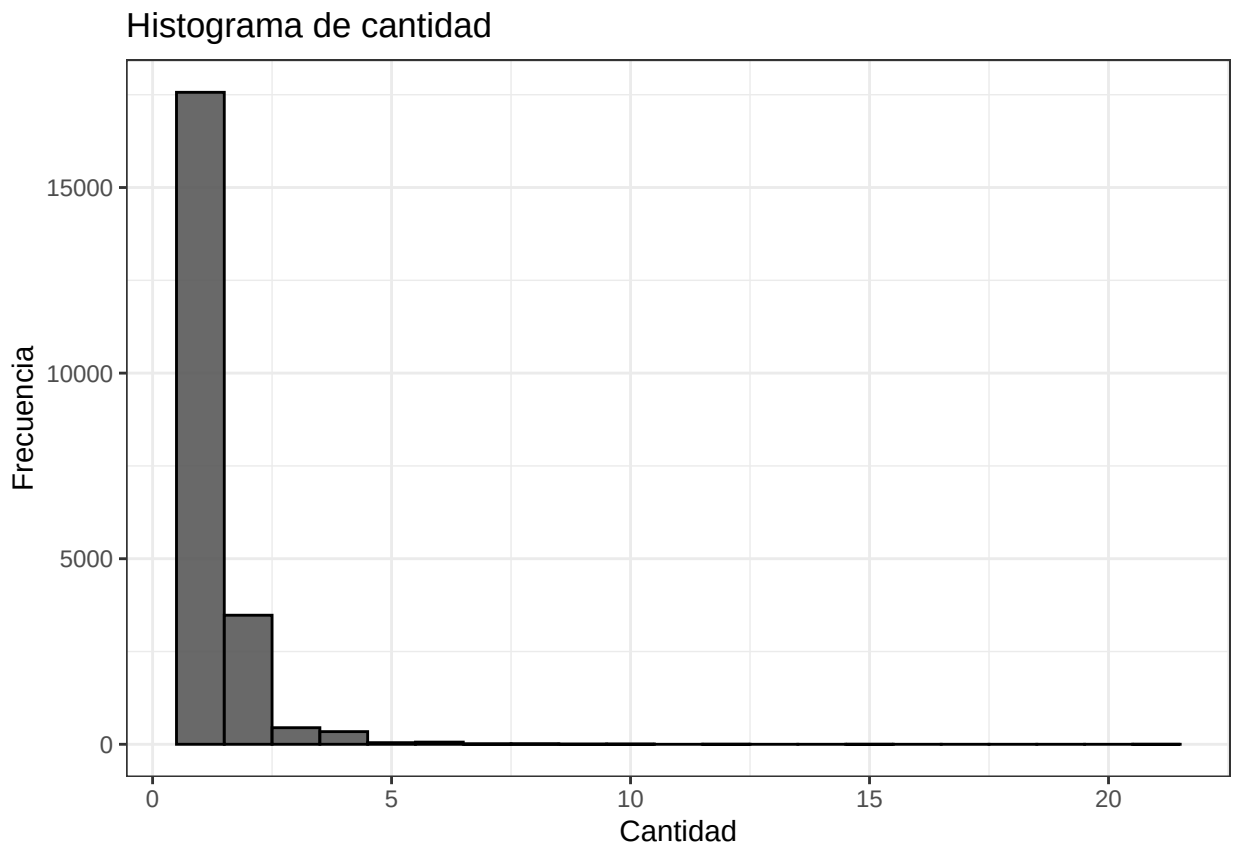
```
median_price <- quantile(base$unit_price)[3]

ggplot(base)+
  geom_histogram(aes(x=unit_price),alpha=0.9,col = 'black')+
  geom_vline(xintercept = median_price,size=0.5,colour="red", linetype = "dashed")+
```

```
  geom_text(aes(x=median_price+2.8, label=paste("Mediana =",median_price), y=4800),size=4, colour="red"
  theme_bw()+
  labs(title="Histograma de precio unitario",x="Precio unitario",y="Frecuencia")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

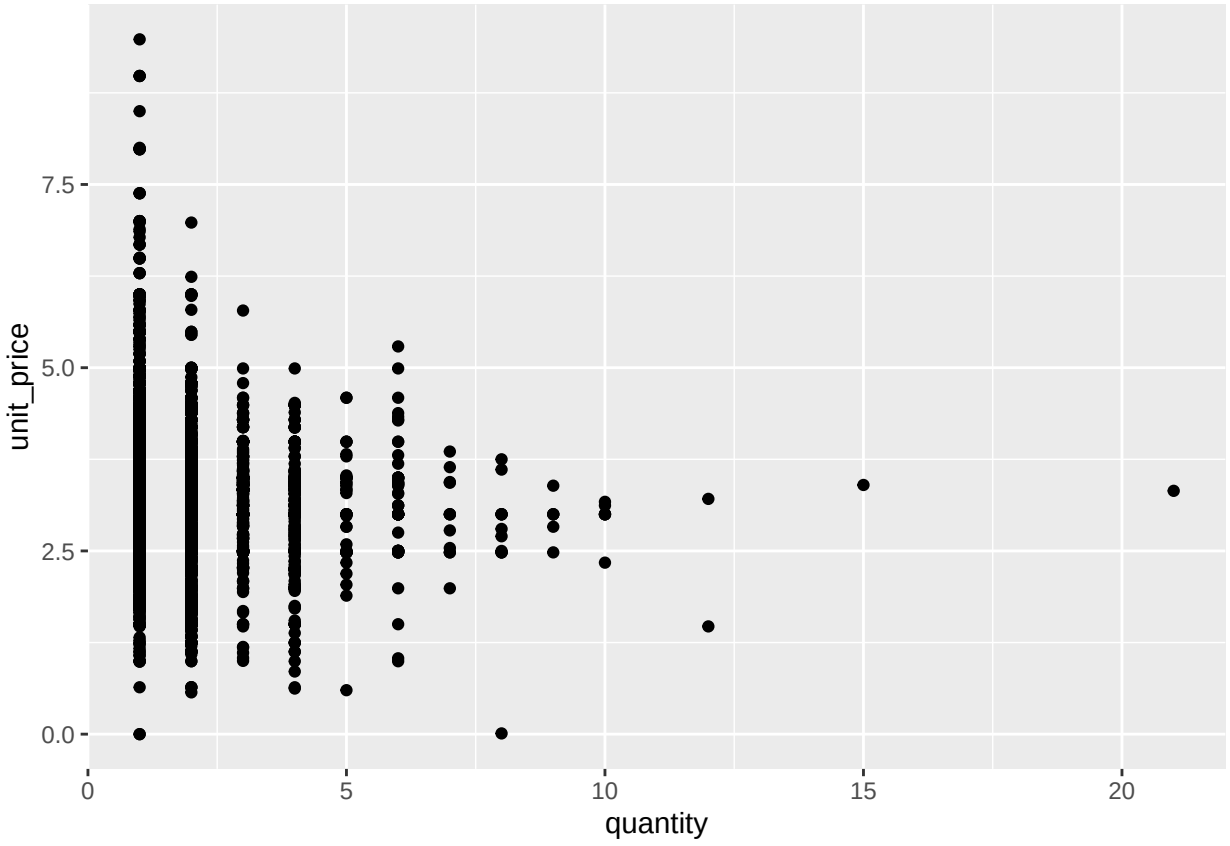



```
ggplot(base)+
  geom_histogram(aes(x=quantity),binwidth=1,alpha=0.9,col = 'black')+
  theme_bw()+
  labs(title="Histograma de cantidad",x="Cantidad",y="Frecuencia")
```



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

```
ggplot(base)+
  geom_point(aes(x=quantity,y=unit_price))
```



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)