Análisis de Datos Categóricos

Ayudantía 5

Felipe Olivares

Contenido

1. R markdown
2. LPM

Rmarkdown

Rmarkdown permite la creación de informes estáticos que combinan texto con código y resultados, gráficos incluídos, generados con R.

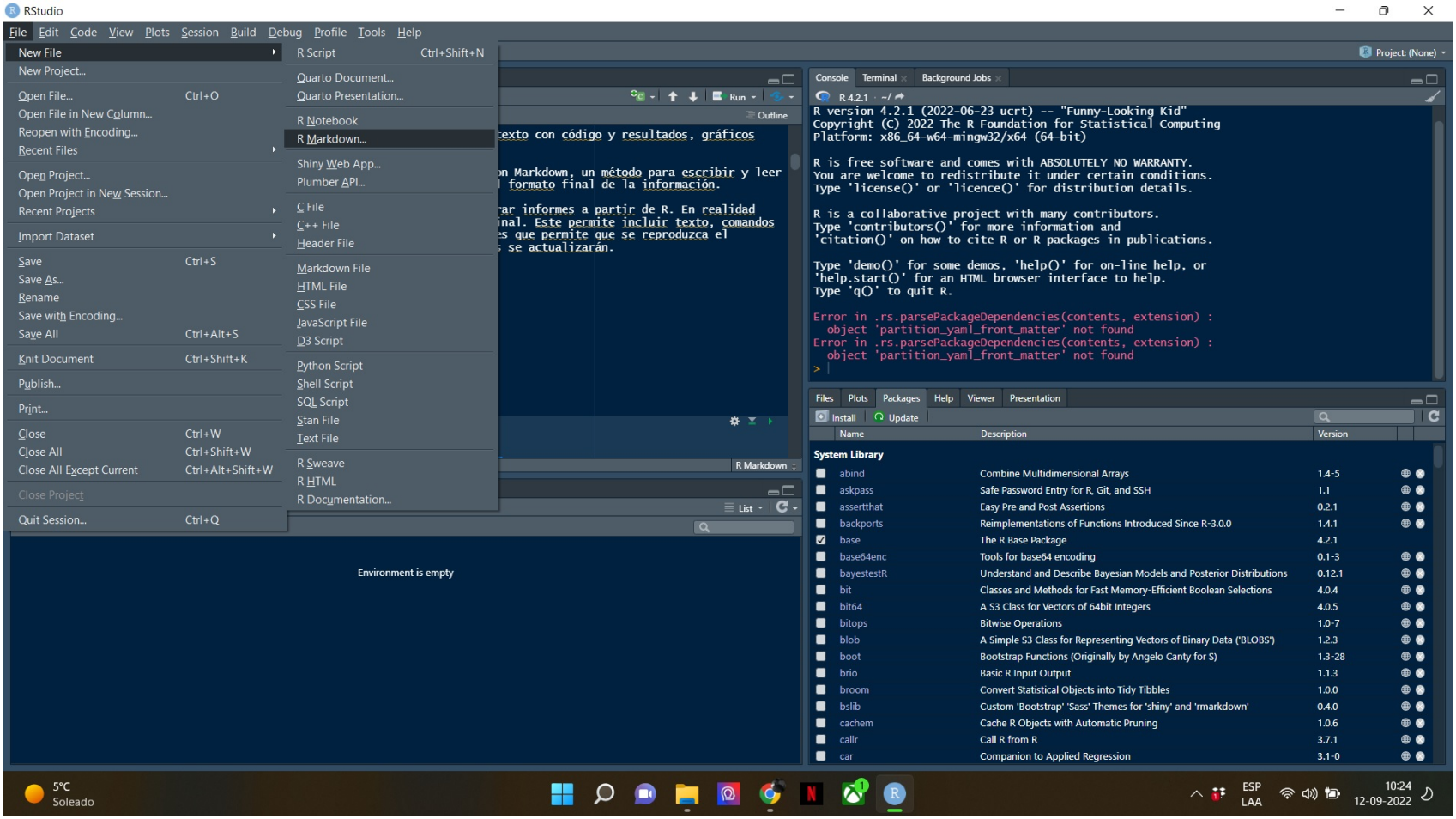
Es la función que vincula el lenguaje open source de R Software con Markdown, un método para escribir y leer de forma rápida en texto plano sin necesidad de preocuparse por el formato final de la información.

Una de las grandes ventajas de usar Rmarkdown es que permite generar informes a partir de R. En realidad utiliza un lenguaje llamado Markdown para enriquecer el informe final. Este permite incluir texto, comandos de R, imágenes y gráficos a un documento. Pero lo más importante es que permite que se reproduzca el análisis realizado y si se incorporan nuevos datos, los resultados se actualizarán.

Rmarkdown

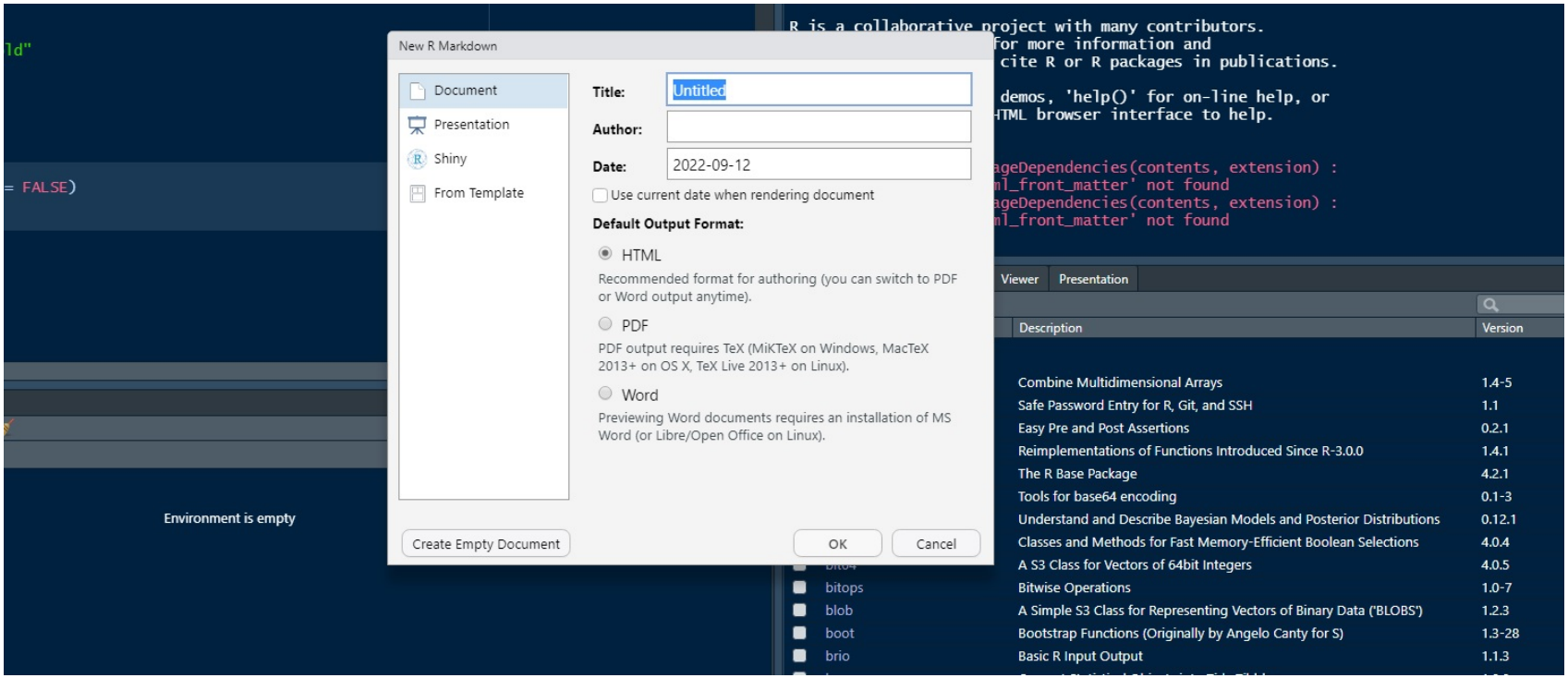
Cómo comenzar?

Tipo de docuneto markdown



Rmarkdown

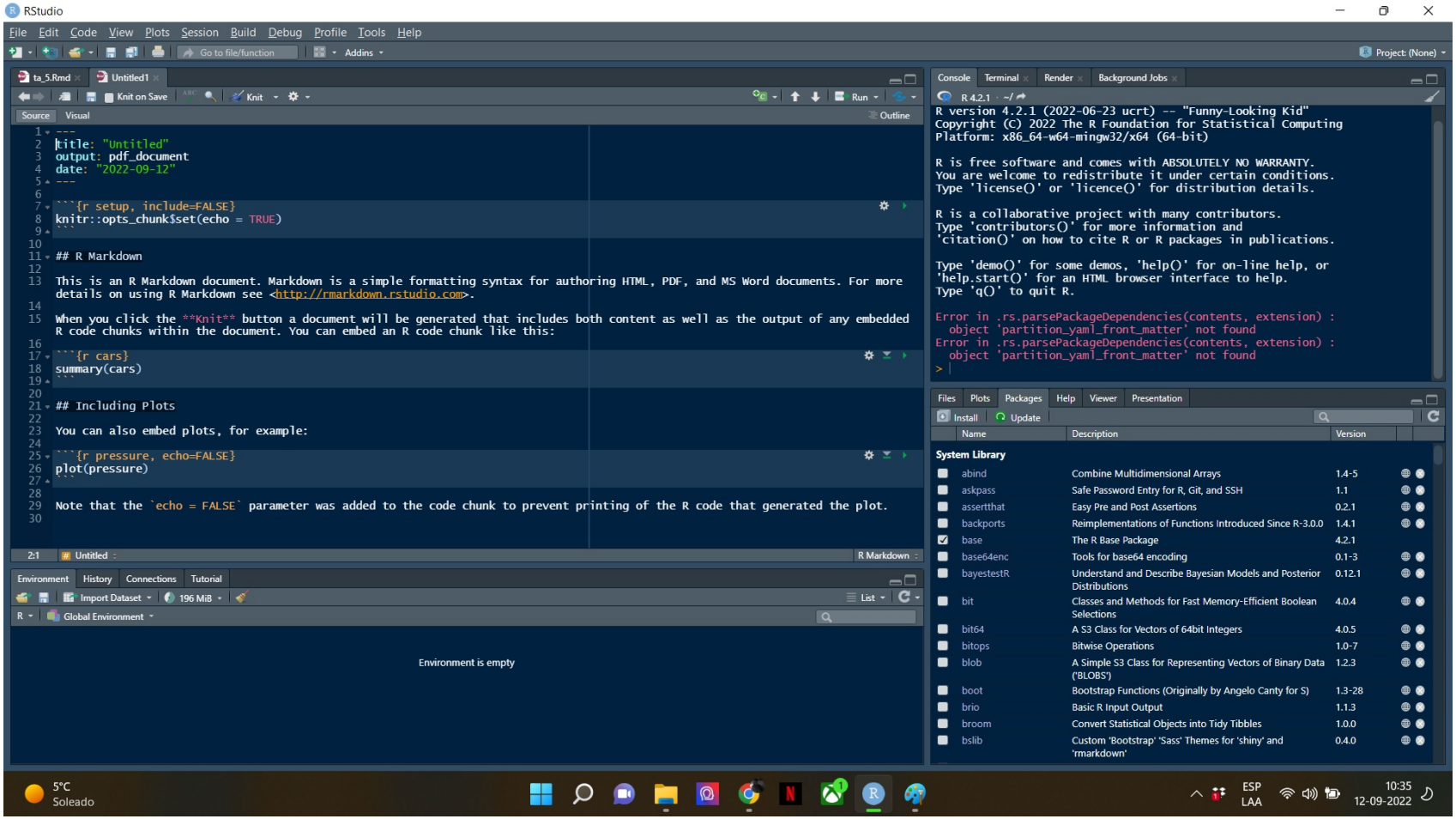
Existen distintos formatos en los cuáles se puede trabajar un formato Markdown. La opción por defecto es crear un documento de texto en html, word o pdf (que son los más utilzidos). Las otras opciones son crear una presentación, un archivo shiny (esta es una clase especial de archivo que permite generar "apps" que se ejecutan a través de un navegador web en un ordenador o dispositivo móvil), o usar una plantilla (template) predefinida.



Rmarkdown

1. La cabecera: como vemos, la plantilla comienza con una cabecera limitada por tres guiones (- - -) por encima y por debajo, donde figuran el título del documento, el autor, la fecha y el formato de salida (html en este caso). En esta cabecera se pueden incluir otras instrucciones para especificar otros formatos de salida, el aspecto de la salida (colores, tamaño de letra), etc.
2. Los chunks: son las cajas grises que contienen código R. Estas cajas están enmarcadas por tres acentos graves (‘‘‘)al inicio y al final. En la primera linea de la caja, junto a los tres acentos y entre llaves se puede asignar un nombre a cada chunk, así como diversas opciones sobre el comportamiento del mismo. Así, por ejemplo, la opción echo=TRUE indica que el contenido de chunk se muestra en la salida, y echo=FALSE que no se muestra.
3. El texto: se escribe directamente en el editor sobre el fondo blanco. Para dar formato al texto se usan una serie de marcas. Así por ejemplo, un hashtag (#) indica que el texto que viene a continuación es un título de primer nivel. Dos hashtags (##) indican un título de segundo nivel. Un texto que se encierre entre parejas de asteriscos (dos asteriscos delante y dos detrás) se muestra en negrilla. Un solo asterisco indica cursivas. En este enlace podemos ver un resumen en español del lenguaje markdown.

Rmarkdown


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
ta_5.Rmd
Untitled1
Knit on Save
Knit
Run
Source
Visual
Outline
---
title: "Untitled"
output: pdf_document
date: "2022-09-12"

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```
R Markdown
This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:
```{r cars}
summary(cars)
```
Including Plots
You can also embed plots, for example:
```{r pressure, echo=FALSE}
plot(pressure)
```
Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.
2:1
Untitled
R Markdown
Environment
History
Connections
Tutorial
Import Dataset
196 MiB
List
Global Environment
Environment is empty
Console
Terminal
Render
Background Jobs
R 4.2.1 ~/
R version 4.2.1 (2022-06-23 ucrt) -- "Funny-Looking Kid"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
Error in .rs.parsePackageDependencies(contents, extension) :
object 'partition_yaml_front_matter' not found
Error in .rs.parsePackageDependencies(contents, extension) :
object 'partition_yaml_front_matter' not found
>
Files
Plots
Packages
Help
Viewer
Presentation
Install
Update
Name
Description
Version
System Library
abind Combine Multidimensional Arrays 1.4-5
askpass Safe Password Entry for R, Git, and SSH 1.1
assertthat Easy Pre and Post Assertions 0.2.1
backports Reimplementations of Functions Introduced Since R-3.0.0 1.4.1
base The R Base Package 4.2.1
base64enc Tools for base64 encoding 0.1-3
bayestestR Understand and Describe Bayesian Models and Posterior Distributions 0.12.1
bit Classes and Methods for Fast Memory-Efficient Boolean Selections 4.0.4
bit64 A S3 Class for Vectors of 64bit Integers 4.0.5
bitops Bitwise Operations 1.0-7
blob A Simple S3 Class for Representing Vectors of Binary Data ('BLOBS') 1.2.3
boot Bootstrap Functions (Originally by Angelo Canty for S) 1.3-28
brio Basic R Input Output 1.1.3
broom Convert Statistical Objects into Tidy Tibbles 1.0.0
bslib Custom 'Bootstrap' 'Sass' Themes for 'shiny' and 'rmarkdown' 0.4.0
5°C
Soleado
ESP
LAA
10:35
12-09-2022

LPM

El modelo lineal de probabilidad, se puede interpretar en términos probabilísticos, en el sentido de que un valor concreto de la recta de regresión mide la probabilidad de que ocurra el hecho objetivo de estudio. Es decir, nuestra variable dependiente se puede considerar como la estimación de la probabilidad de que ocurra el hecho objetivo de estudio Yi = 1 siguiendo el siguiente criterio: Valores próximos a cero se corresponde con una baja probabilidad de ocurrencia del hecho estudiado (menor cuanto más próximos a cero); mientras que a valores próximos a uno se les asigna una probabilidad elevada de ocurrencia (mayor cuanto más próximos a uno).

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots \beta_k x_{ki} + e_i$$

El modelo de regresión más simple aplicado a datos categóricos es el modelo lineal de probabilidad (LPM). Esto es básicamente lo mismo que un modelo de regresión lineal.

Variable dependiente = intercepto + predictores + error aleatorio

LPM

Revisemos los mismos datos de la ayudantía anterior para crear un modelo de regresión lineal en base a predictores que revisamos

```
head(df1)
```

```
# # A tibble: 6 x 16
#   ano   region     educa~1 indig~2 laboral salud pacif~3 disru~4 viole~5 orga
#   <fct> <fct>      <fct>   <fct>   <fct>   <fct> <fct>   <fct>   <fct>   <fct
# 1 2009  Metropoli~ No      No      No      No    Sí      No      No      1 or
# 2 2009  Tarapacá   Sí      No      No      No    Sí      No      No      1 or
# 3 2009  Tarapacá   No      No      Sí      No    Sí      Sí      No      1 or
# 4 2009  O´Higgins  No      No      Sí      No    No      Sí      No      Sin
# 5 2009  Araucanía  No      Sí      No      No    Sí      No      No      Sin
# 6 2009  Araucanía  No      No      No      No    No      Sí      No      Sin
# # ... with 6 more variables: nacional <fct>, macrozona <chr>,
# #   estudiantes <fct>, trabajadores <fct>, ppolicial <fct>, apolicial <fct>,
# #   and abbreviated variable names 1: educacion, 2: indigena, 3: pacifica,
# #   4: disruptiva, 5: violenta, 6: organizacion
# # i Use ‘colnames()‘ to see all variable names
```

LPM

La regresión lineal que vamos a construir la haremos sobre la base de la pregunta acerca del tipo de control policial que realizan las policías (Control negociado o control violento de la protesta) en el caso de las manifestaciones sobre educación en Chile para los años 2009-2019. Para realizar esto, utilizaremos como variable dependiente la acción policial y variables independientes: grupos sociales estudiantes, demandas sobre educación y tipo de protesta disruptiva.

```
str (df1) # siempre importante revisar la estructura de las codificaciones
```

```
# tibble [23,398 x 16] (S3: tbl_df/tbl/data.frame)
#  $ ano         : Factor w/ 11 levels "2009","2010",..: 1 1 1 1 1 1 1 1 1 1 ..
#   ..- attr(*, "label")= chr "Años"
#  $ region      : Factor w/ 16 levels "Tarapacá","Antofagasta",..: 13 1 1 6 9
#   ..- attr(*, "label")= chr "Regiones"
#  $ educacion   : Factor w/ 2 levels "No","Sí": 1 2 1 1 1 1 1 1 1 1 ...
#   ..- attr(*, "label")= chr "Demanda - Educacional"
#  $ indigena    : Factor w/ 2 levels "No","Sí": 1 1 1 1 2 1 2 1 1 1 ...
#   ..- attr(*, "label")= chr "Demanda - Indígenas"
#  $ laboral     : Factor w/ 2 levels "No","Sí": 1 1 2 2 1 1 1 1 1 1 ...
#   ..- attr(*, "label")= chr "Demanda - Laborales"
#  $ salud       : Factor w/ 2 levels "No","Sí": 1 1 1 1 1 1 1 1 1 1 ...
#   ..- attr(*, "label")= chr "Demanda - Salud"
#  $ pacifica    : Factor w/ 2 levels "No","Sí": 2 2 2 1 2 1 2 2 2 2 ...
#   ..- attr(*, "label")= chr "Táctica - Pacífica"
#  $ disruptiva  : Factor w/ 2 levels "No","Sí": 1 1 2 2 1 2 2 1 1 1 ...
#   ..- attr(*, "label")= chr "Táctica - Disruptiva"
```

LPM

Creamos un subset de datos que contenga solo aquellos casos en que existe acción policial. Así mismo, y de acuerdo a lo que observamos previamente en la estructura de los datos, recodificamos las variables de interés para colocarlas en nuestra regresión.

Para efecto de la regresión que modelaremos la función de un control violento de la protesta respecto de un control negociado. Por lo tanto, es importante que el control negociado de la protesta=0 y control violento de la protesta=1 (consideren lo dicho previamente respecto del sentido del efecto que queremos estimar)

```
df2 <- df1 %>% select(apolicial,educacion,estudiantes,,disruptiva) %>%
  mutate(apolicial = if_else(apolicial=="Violencia Policial",1,0),
         educacion = if_else(educacion=="Sí",1,0),
                  estudiantes= if_else(estudiantes=="Sí",1,0),
         disruptiva = if_else(disruptiva=="Sí",1,0))  %>%
  mutate(apolicial = as.numeric(apolicial),
         educacion = as.numeric(educacion),
                 estudiantes = as.numeric(estudiantes),
         disruptiva = as.numeric(disruptiva))  %>%
  na.omit(df2)
```

LPM

```
#Modelo de regresión propuesto
lm1 <-lm(apolicial ~ educacion + estudiantes + disruptiva, data=df2)
summary(lm1)
```

```
#
# Call:
# lm(formula = apolicial ~ educacion + estudiantes + disruptiva,
#     data = df2)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -0.6878 -0.5488  0.3213  0.4512  0.5238
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept)  0.485365   0.010487  46.282  < 2e-16 ***
# educacion   -0.009148   0.022350  -0.409    0.682
# estudiantes  0.139072   0.021381   6.504 8.42e-11 ***
# disruptiva   0.063414   0.012770   4.966 7.03e-07 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.4926 on 6075 degrees of freedom
# Multiple R-squared:  0.01742, Adjusted R-squared:  0.01694
# F-statistic: 35.91 on 3 and 6075 DF,  p-value: < 2.2e-16
```

LPM

Interpretación

recordemos..

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots \beta_k x_{ki} + e_i$$

Entonces decimos que el efecto de x_k sobre y es β_k. ¿Qué significa?

"Un cambio en Δ unidades de x_k se traduce en un cambio en Δ β_k unidades en el valor esperado de y_i"

LPM

Lectura de coeficientes

1- Efecto de educación sobre control violento de la protesta:

R: En promedio, la presencia de demandas por educación disminuye la probabilidad de un control violento de la protesta en 0.009, es decir, la presencia de demandas educativas durante la protesta disminuye un 9% controlando por el resto de la covariables del modelo. Sin embargo, este efecto no es significativo.

2- Efecto de la presencia de estudiantes sobre control violento de la protesta:

R:En promedio, la presencia de estudiantes durante una manifestación aumenta la probabilidad de un control violento de la protesta en 0.14, es decir, la presencia de estudiantes aumenta un 14% controlando por nuestras covariables del modelo. Este efecto es estadísticamente significativo a un 99,9% de confianza y un valor p= 8.42e-11

3- Efecto de tácticas disruptivas sobre el control violento de la protesta:

R:En promedio, la presencia de tácticas disruptivas durante la protesta aumenta la probabilidad de un control violento de la protesta en 0.06, es decir, la presencia de tácticas disruptivas aumenta un 6%, controlando por el resto de la covariables del modelo. Este efecto es estadísticamente significativo a un 99,9% de confianza y un valor p= -7.03e07

LPM

Limitaciones...

Distribución y rango: Los modelos de regresión lineal asumen que la variable dependiente son manifestaciones de distribuciones normales, pero que ocurre cuándo nuestras observaciones no son normales (toma valores 0 y 1). Esto puede provocar que nuestras estimaciones escapen de los rangos 0 y 1 (que son el rango natural de una probabilidad).

Sesgados e inconsistentes: Coeficiente no da en el blanco y además consistentemente convergen en un valor erróneo.

Varianza: Supuesto de varianza constante en caso de las regresiones lineales no se cumple para variables categóricas $= p_i(1 - p_i)$ Sabemos que usar una variable Bernoulli o Binomial no tiene varianzas constantes dada la distribución que tienen los datos.