Instituto Tecnológico Autónomo de México

Propedéutico

Análisis estadístico de los decesos de COVID19 en la CDMX

Marco Antonio Ramos Juárez

142244

Rodrigo Juárez Jaramillo

145804

Ciudad de México 2021

Contents

Introducción

Hasta el siglo XXI, los seres humanos han sido testigos de tres pandemias mortales: SARS síndrome respiratorio de Oriente Medio (MERS) y COVID-19. Todos estos virus, que son responsables de causar infecciones agudas del tracto respiratorio (IRA), son de naturaleza altamente contagiosa y/o han causado una alta mortalidad. COVID-19 apareció por primera vez en Wuhan, China, en diciembre de 2019 y se extendió rápidamente por todo el mundo.

Además, de acuerdo con el New York Times, el covid-19 ha enfermado a más de 183 millones de personas en todo el mundo. Más de 4 millones de personas han muerto hasta el día de hoy. La pandemia de coronavirus ha afectado nuestras vidas, nuestra economía y casi todos los rincones del mundo. Con los enormes volúmenes de datos que se producen a diario, necesitamos crear información, precisa para diseñar estrategias efectivas para poder terminar con esta pandemia mundial.

Al 02 de julio de 2021, México es el cuarto país con mayor número de defunciones acumuladas, estando por debajo de Estados Unidos, Brasil e India. De acuerdo a un informe de Statista: Estados Unidos encabeza la clasificación al superar los 620,600 decesos, seguido de Brasil con alrededor de 520.200. Para el 2 julio de 2021, había más de 183. 5millones de casos confirmados de COVID-19 en todo el mundo. Es importante recalcar que día con día el número de personas fallecidas en el mundo van aumentando, por lo cual, resulta imperativo seguir investigando más sobre las causas de muerte y tomar acciones preventivas para evitar defunciones alrededor del mundo.

En el contexto de una crisis sanitaria como la que estamos viviendo, es fundamental que los análisis relacionados se hagan con rigor de tal manera que los hallazgos sean veraces y precisos. Esto es de especial importancia en el contexto de la divulgación de información falsa o de desinformación en redes sociales y medios de comunicación. Por ello, el proposito de esta practica final es el anlizar las bases de datos oficiales de contagios y decesos de COVID19 y aplicar los métodos vistos durante la clase con el fin de obtener hallazgos robustos, interesantes y útiles relacionados con el desarrollo de la pandemia en México.

Datos

Extracción

Los datos con los que trabajaremos provienen de la base de datos agregada oficial del Gobierno Federal. Esta base de datos incluye todos los casos diarios asociados a COVID-19 que se han registrado en la CDMX. Cuenta con contenido desagregado por sexo, edad, nacionalidad, padecimientos asociados entre otras variables.

```
data <- read.csv("data.csv")
```

Limpieza

En primer lugar, analizamos la clase de cada variable:

```
kable(data.frame(sapply(data, class)), col.names = "Clase", booktabs = T,
    longtable = T) %>% kable_styling(latex_options = c("scale_down",
    "repeat_header"), position = "center")
```

```
## Warning in styling_latex_scale_down(out, table_info): Longtable cannot be
## resized.
```

	Clase
X.1	integer
X	integer
fecha_actualizacion	character
id_registro	character
origen	character
sector	character
entidad_um	character
sexo	character
entidad_nac	character
entidad_res	character
municipio_res	character
tipo_paciente	character
fecha_ingreso	character
fecha_sintomas	character
fecha_def	character
intubado	character
neumonia	character
edad	integer
nacionalidad	character

(continued)

	Clase
embarazo	character
habla_lengua_indig	character
indigena	character
diabetes	character
epoc	character
asma	character
inmusupr	character
hipertension	character
otra_com	character
cardiovascular	character
obesidad	character
renal_cronica	character
tabaquismo	character
otro_caso	character
toma_muestra_lab	character
resultado_lab	character
toma_muestra_antigeno	character
resultado_antigeno	character
clasificacion_final	character
migrante	character
pais_nacionalidad	character
pais_origen	character
uci	character

Notamos que todas las variables pertenecen a la clase *character* (con expeción de la edad y los id de cada fila) lo cual habla de que debemos clasificarlas de manera adecuada para aprovechar la información por lo que si queremos aprovechar esas variables, el siguiente paso es asignarles la clase correcta. En el siguiente *chunk* clasificamos las fechas y las variables categóricas de manera correcta.

```
# notamos que el formato de las fechas es AAAA/MM/DD
data$fecha_def <- as.Date(data$fecha_def)
data$fecha_sintomas <- as.Date(data$fecha_sintomas)
data$fecha_ingreso <- as.Date(data$fecha_ingreso)

# creamos un dataframe con las variables relacionadas con
```

```
# caracteristicas de las personas
df8 <- data %>% select(sexo, tipo_paciente, intubado, fecha_def,
    neumonia, edad, diabetes, epoc, asma, inmusupr, hipertension,
    obesidad, renal_cronica, tabaquismo)
# df8$sexo<-ifelse(df8$sexo=='MUJER',1,0) #Si es mujer tomará
# valor de 1, 0 en caso de ser hombre
df8$diabetes <- ifelse(df8$diabetes == "NO", 0, 1)
df8$epoc <- ifelse(df8$epoc == "NO", 0, 1)
df8$asma <- ifelse(df8$asma == "NO", 0, 1)
df8$inmusupr <- ifelse(df8$inmusupr == "NO", 0, 1)
df8$hipertension <- ifelse(df8$hipertension == "NO", 0, 1)
df8$obesidad <- ifelse(df8$obesidad == "NO", 0, 1)
df8$renal_cronica <- ifelse(df8$renal_cronica == "NO", 0, 1)
df8$tabaquismo <- ifelse(df8$tabaquismo == "NO", 0, 1)
```

Missing values

En segundo lugar, checamos la cantidad de valores faltantes de cada variable con almenos un NA.

```
missing_data <- sapply(data, function(x) sum(length(which(is.na(x)))))

options(scipen = 999)  #quito notación científica
missing_data <- data.frame(missing_data)
missing_data <- add_rownames(missing_data, var = "variable")
```

```
## Warning: `add_rownames()` was deprecated in dplyr 1.0.0.
## Please use `tibble::rownames_to_column()` instead.
```

```
# imprimo solo las variables con nas en porcentaje
output_missing_data <- missing_data %>% filter(missing_data >
    0) %>% mutate(missing_data = missing_data * 100/nrow(data))

kable(output_missing_data, booktabs = T, col.names = c("Variable",
    "Porcentaje"), caption = "Valores faltantes") %>% kable_styling(position = "center",
    latex_options = "repeat_header")
```

Lo primero que notamos es que las variables *entidad_res* y *municipio_res*, referentes a la entidad y municpio de residencia cuentan con una alta proporción de valores faltantes (la mayoría de sus valores está faltante). De igual manera, la fecha de defunción también cuenta con una alta proporción. En cuanto a sector y país de origen, los valores faltantes realmente son muy pocos y podemos prescindir de dichas observaciones.

Table 2. Valores faltantes

Variable	Porcentaje
sector	0.0012109
entidad_res	85.4328979
municipio_res	85.4329316
fecha_def	97.9585041
pais_origen	0.1057210

Nuestra primera hipótesis esque los valores faltante en *entidad_res* y *municipio_res* en realidad se refieren a personas cuya residencia está en CDMX. Nuestra segunda hipótesis (que convertiremos en supuesto para poder trabajar) esque los valores faltantes en la fecha de defunción corresponden a personas que no han perecido.

```
# como podemos observar no hay CDMX
summary(factor(data$entidad_res))
```

```
##                         DURANGO                      GUANAJUATO
##                             121                             888
##                        GUERRERO                         HIDALGO
##                            1756                            5888
##                         JALISCO                          MÉXICO
##                             691                          408582
##              MICHOACÁN DE OCAMPO                         MORELOS
##                            1133                            3008
##                         NAYARIT                       NUEVO LEÓN
##                              86                             390
##                          OAXACA                          PUEBLA
##                            1032                            3087
##                       QUERÉTARO                     QUINTANA ROO
##                            1103                             359
##                  SAN LUIS POTOSÍ                         SINALOA
##                             347                             216
##                         SONORA                          TABASCO
##                             171                             345
##                      TAMAULIPAS                        TLAXCALA
##                             330                            1031
## VERACRUZ DE IGNACIO DE LA LLAVE                         YUCATÁN
##                            2131                             219
##                       ZACATECAS                          NA's
```

```
##                154                            2539850
```

Creación de variables de interés

Asimismo podemos crear algunas variables de interés que nos serán muy útiles en el análisis: días desde el primer sintoma a la hospitalización y días desde la hospitalización hasta el decesos (en caso de)

```
## time to hospital
data <- data %>% mutate(tiempo_enf_to_hosp = fecha_ingreso -
    fecha_sintomas)
data$tiempo_enf_to_hosp <- as.numeric(data$tiempo_enf_to_hosp)

## time to hospital and death
data <- data %>% mutate(tiempo_to_death = fecha_def - fecha_ingreso)
data$tiempo_to_death <- as.numeric(data$tiempo_to_death)

## time from infection to death
data <- data %>% mutate(tiempo_inf_death = fecha_def - fecha_sintomas)
data$tiempo_inf_death <- as.numeric(data$tiempo_inf_death)
```

Conclusiones

De la extracción y limpieza podemos concluir que:

- Existen muchos *missing values* en cuanto a las variables geográficas que muy probablemente se refieren a gente habitante de la CDMX.
- Existen muchos *missing values* en cuanto a la fecha de deceso pero probablemente es porque es gente que no ha perecido.
- La mayoría de las variables con categóricas y las observaciones son a nivel individuo.
- Convertimos las variables de fecha y categóricas al formato correcto.

EDA

En esta parte realizaremos un análisis exploratorio de las principales variables de interés.

Género

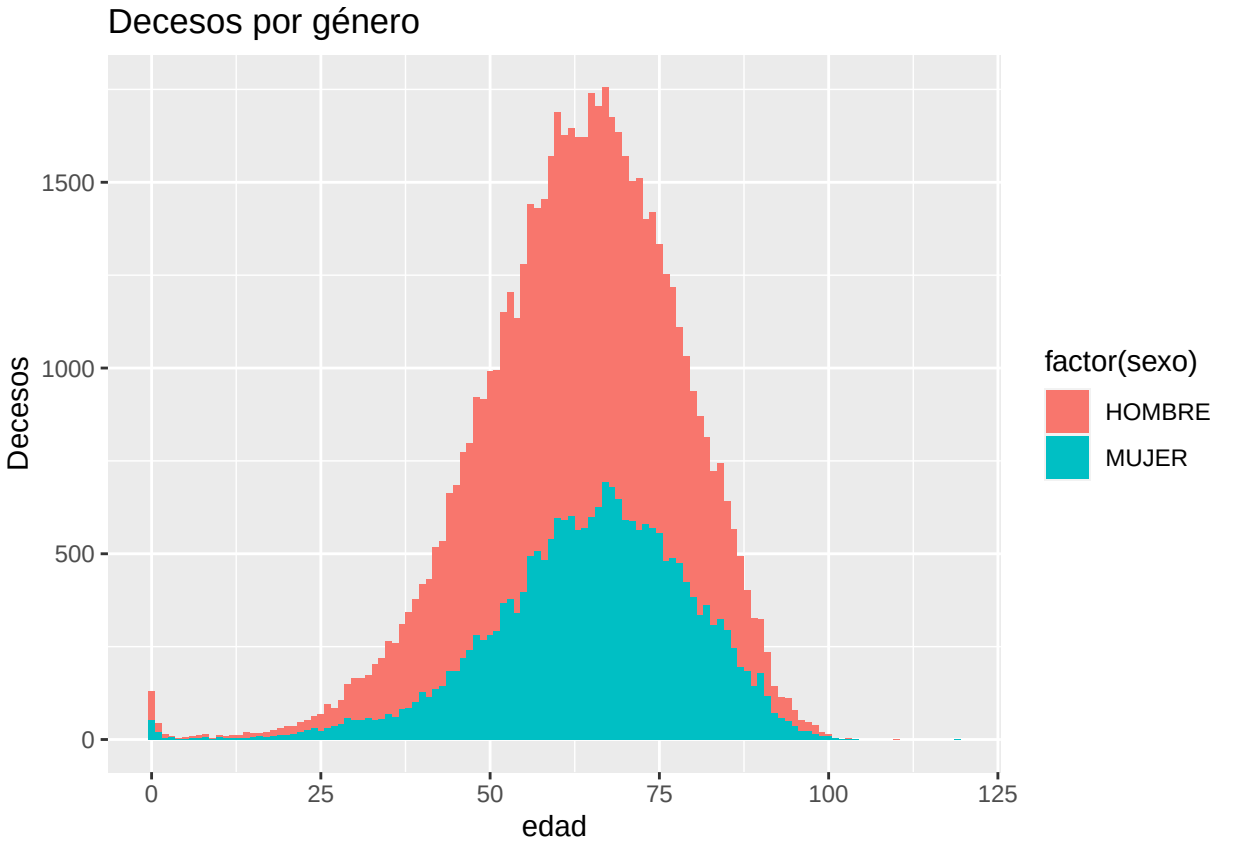
Lo que notamos en género es que los decesos han sido más en los hombres, aunque las distribuciones se parezcan, notamos que las medias son ligeramente distintas: la de mujeres parece ser en mayor

edad y su distribución parece tener una cola más larga hacia la izquierda y una cola con mayor pendiente hacia la derecha con respecto a la distribución de los hombres.

```
# data wrangling
df8$fecha_def <- format(df8$fecha_def, format = "%V")
df8["def"] <- 1
df9 <- subset(df8, fecha_def != "NA")
df10 <- df9 %>% group_by(sexo, edad) %>% summarise(sum_def = sum(def,
    na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'sexo'. You can override using the `.groups` argument.
```

```
# plot
ggplot(df10, aes(x = edad, y = sum_def, fill = factor(sexo))) +
    geom_bar(stat = "identity") + ggtitle("Decesos por género") +
    ylab("Decesos")
```

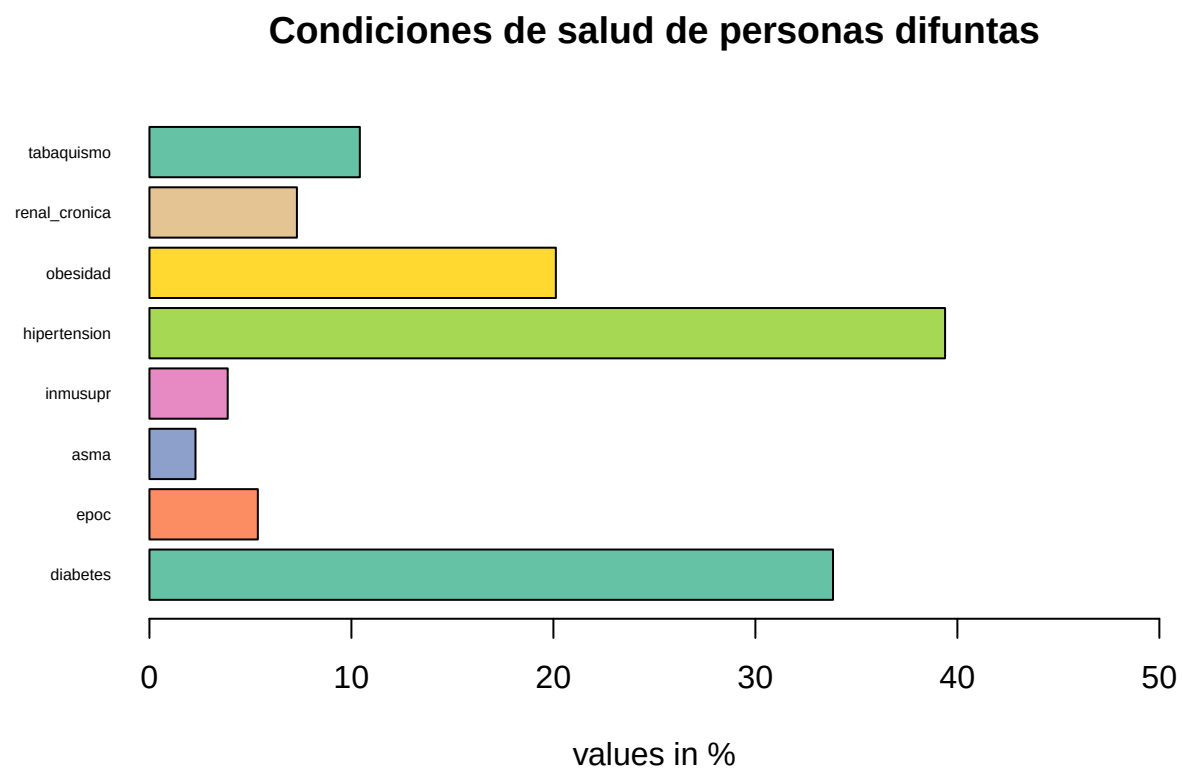


Comorbidades

En cuanto a comorbidades vemos algunas relaciones muy interesantes: notamos que la hipertensión y diabetes son prevalentes en los decesos (con casi el 40% el 33%). De manera secundaría la obesidad

y el tabaquismo aparecen en el 21% y 12% de los decesos aproximadamente.

```
condiciones<-c(sum(df9$diabetes),sum(df9$epoc),sum(df9$asma),sum(df9$inmusupr),sum(df9$hiperte
porc_condiciones<-condiciones*100/sum(df9$def)
nom_cond<-c('diabetes','epoc','asma','inmusupr','hipertension','obesidad','renal_cronica','taba

data2 <- data.frame(
  name=nom_cond ,
  value=porc_condiciones
  )
coul <- brewer.pal(7, "Set2")
barplot(height=data2$value, names=data2$name,
        col=coul,
        xlab="values in %",
        main="Condiciones de salud de personas difuntas",
        xlim=c(0,50),
        horiz=TRUE,las=1,cex.names=.5

        )
```



Tiempos

En cuanto a los tiempos, notamos pasa hasta una semana desde le primer sintoma hasta la hospitalización. En cuanto a decesos, notamos que estos ocurren desde el segundo día de la hospitalización

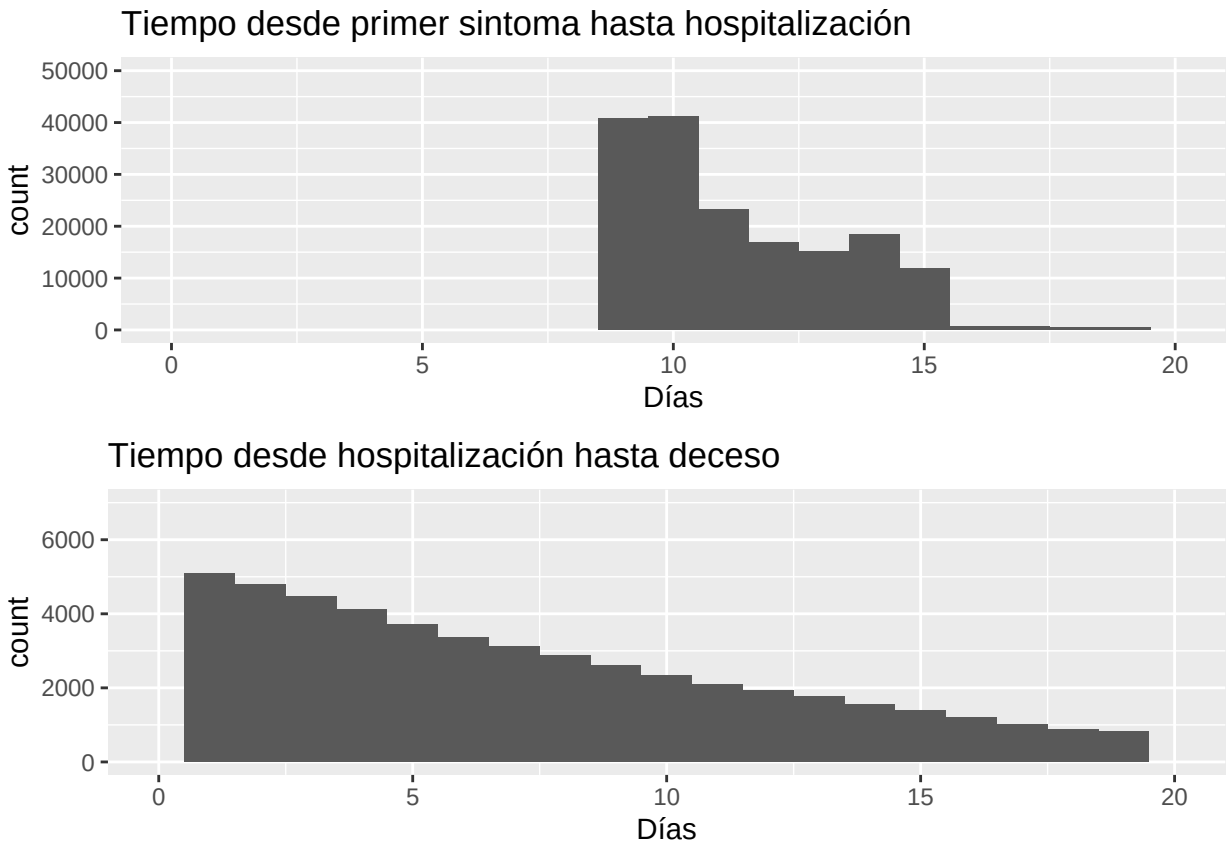
```
times <- data %>% select(tiempo_enf_to_hosp, tiempo_to_death)

a <- ggplot(times, aes(x = tiempo_enf_to_hosp)) + geom_histogram(binwidth = 1) +
    ggtitle("Tiempo desde primer sintoma hasta hospitalización") +
    xlab("Días") + xlim(0, 20) + ylim(0, 50000)

b <- ggplot(times, aes(x = tiempo_to_death)) + geom_histogram(binwidth = 1) +
    ggtitle("Tiempo desde hospitalización hasta deceso") + xlab("Días") +
    xlim(0, 20) + ylim(0, 7000)

grid.arrange(a, b, ncol = 1)
```

```
## Warning: Removed 2475 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 10 rows containing missing values (geom_bar).
```

```
## Warning: Removed 2917805 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



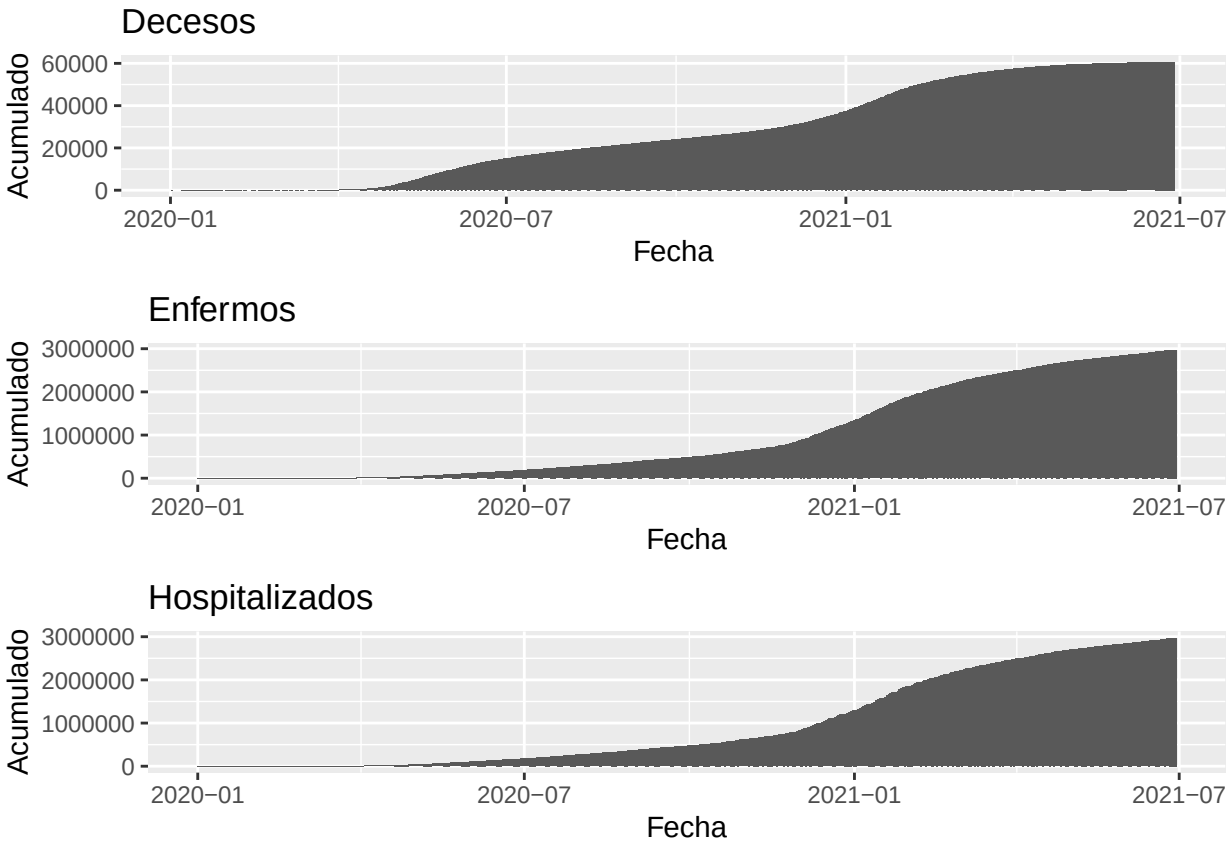
Variables acumuladas

En cuanto a las variables acumuladas (decesos, hospitalizaciones y enfermos), notamos que las 3 tienen un punto de inflexión aproximadamente en diciembre 2020. Asimismo notamos que la curva de decesos es la que se ha mostrado menos sensible de las 3. Asimismo notamos que las 3 curvas tienen pequeñas tiras blancas, estas representan los días que no se recopilan datos o bien los días de corte, por ejemplo, probablemente los fines de semana para los decesos.

```
sintomas <- data %>% group_by(fecha_sintomas) %>% summarise(cantidad = n()) %>%
    mutate(cum = cumsum(cantidad))

ingreso <- data %>% group_by(fecha_ingreso) %>% summarise(cantidad = n()) %>%
    mutate(cum = cumsum(cantidad))

muerte <- data %>% group_by(fecha_def) %>% summarise(cantidad = n()) %>%
    mutate(cum = cumsum(cantidad))

a <- ggplot(data = muerte, aes(x = fecha_def, y = cum)) + geom_bar(stat = "identity") +
    ggtitle("Decesos") + xlab("Fecha") + ylab("Acumulado")
```

```
b <- ggplot(data = sintomas, aes(x = fecha_sintomas, y = cum)) +
    geom_bar(stat = "identity") + ggtitle("Enfermos") + xlab("Fecha") +
    ylab("Acumulado")

c <- ggplot(data = ingreso, aes(x = fecha_ingreso, y = cum)) +
    geom_bar(stat = "identity") + ggtitle("Hospitalizados") +
    xlab("Fecha") + ylab("Acumulado")

grid.arrange(a, b, c, ncol = 1)
```

```
## Warning: Removed 1 rows containing missing values (position_stack).
```



Decesos por semana

```
year_def <- format(data$fecha_def, format = "%Y")  #Obtenemos los años de las fechas
week_def <- format(data$fecha_def, format = "%V")  #Obtenemos las semanas de las fechas
defun <- data.frame(year_def, week_def, def = data$fecha_def)  #creamos un data frame con los
df2 <- subset(defun, year_def != "NA")  # eliminamos los datos en los cuales no hubo defuncion
df2["def"] <- 1  #Agregamos una columna de 1's para hacer la suma de defunciones por semana
```

```
tibble_df2 <- as_tibble(df2)
df3 <- df2 %>% group_by(year_def, week_def) %>% summarize(sum_def = sum(def,
    na.rm = TRUE))  #df3 contiene la suma de defunciones por semana
```

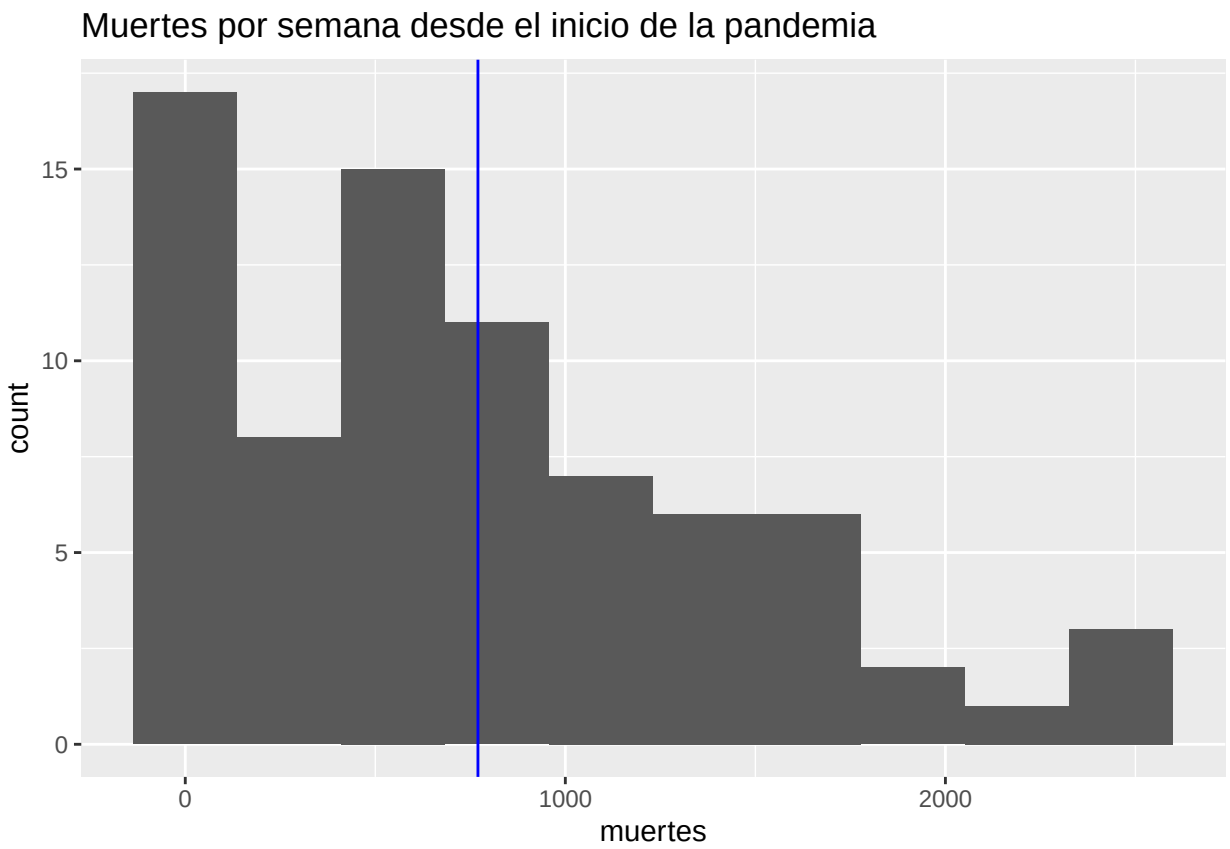
```
## `summarise()` has grouped output by 'year_def'. You can override using the `.groups` argumer
```

```
df4 <- subset(df3, week_def != "53")  #eliminamos datos que no son lógicos, como semana 53 del
df5 <- df4[1:(76), ]  #eliminamos última semana para después hacer una estimación de la misma,
sem <- df5$week_def
muertes <- df5$sum_def
semana <- as.double(sem)
anio <- df5$year_def

df6 <- data.frame(anio, semana, muertes)  #A las semanas del año 2021, le agregamos 52 para di
a <- 1:24
for (i in seq_along(a)) {
    df6[52 + i, 2] <- df6[i, 2] + 52
}
```

```
summary(df6)
```

```
##      anio                semana          muertes
##  Length:76          Min.   : 1.00   Min.   :   1.0
##  Class :character   1st Qu.:19.75   1st Qu.: 199.2
##  Mode  :character   Median :38.50   Median : 641.0
##                     Mean   :38.50   Mean   : 770.1
##                     3rd Qu.:57.25   3rd Qu.:1139.8
##                     Max.   :76.00   Max.   :2464.0
```

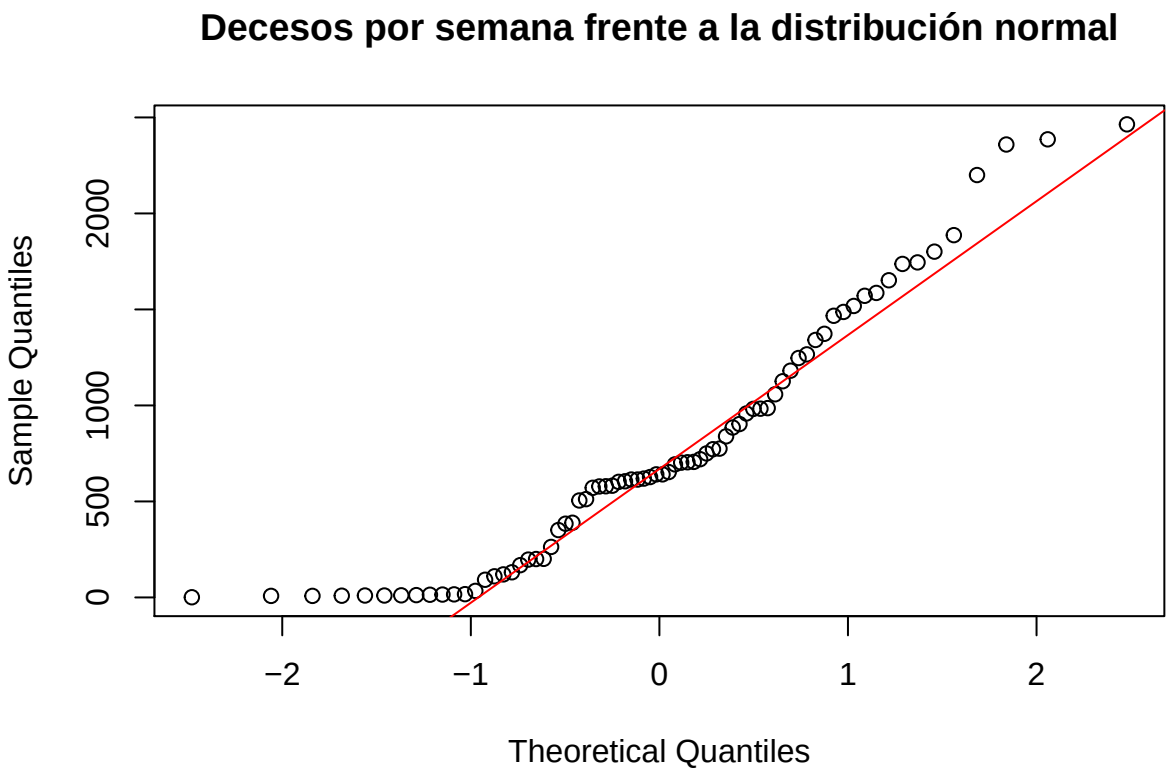
```
ggplot(df6, aes(x = muertes)) + geom_histogram(bins = 10) + ggtitle("Muertes por semana desde
    geom_vline(xintercept = mean(muertes), color = "blue")
```



Lo que notamos principalmente esque la distribución de las muertes por semana está sesgada hacia la derecha de la media, marcada con una línea azul con valor de 770 muertes. Recordando las graficas anteriores sobre decesos, contagios y hospitalizaciones acumuladas es fácil entender que este histograma tenga esta forma, marcando muchas semanas por debajo de la media, y algunas menos semanas pero con una cantidad de muertes muy elevada a la izquierda.

Esto es más evidente si usamos un qqplot contrastando contra la distribución normal.

```
qqnorm(df6$muertes, main = "Decesos por semana frente a la distribución normal")
qqline(df6$muertes, col = "red")
```

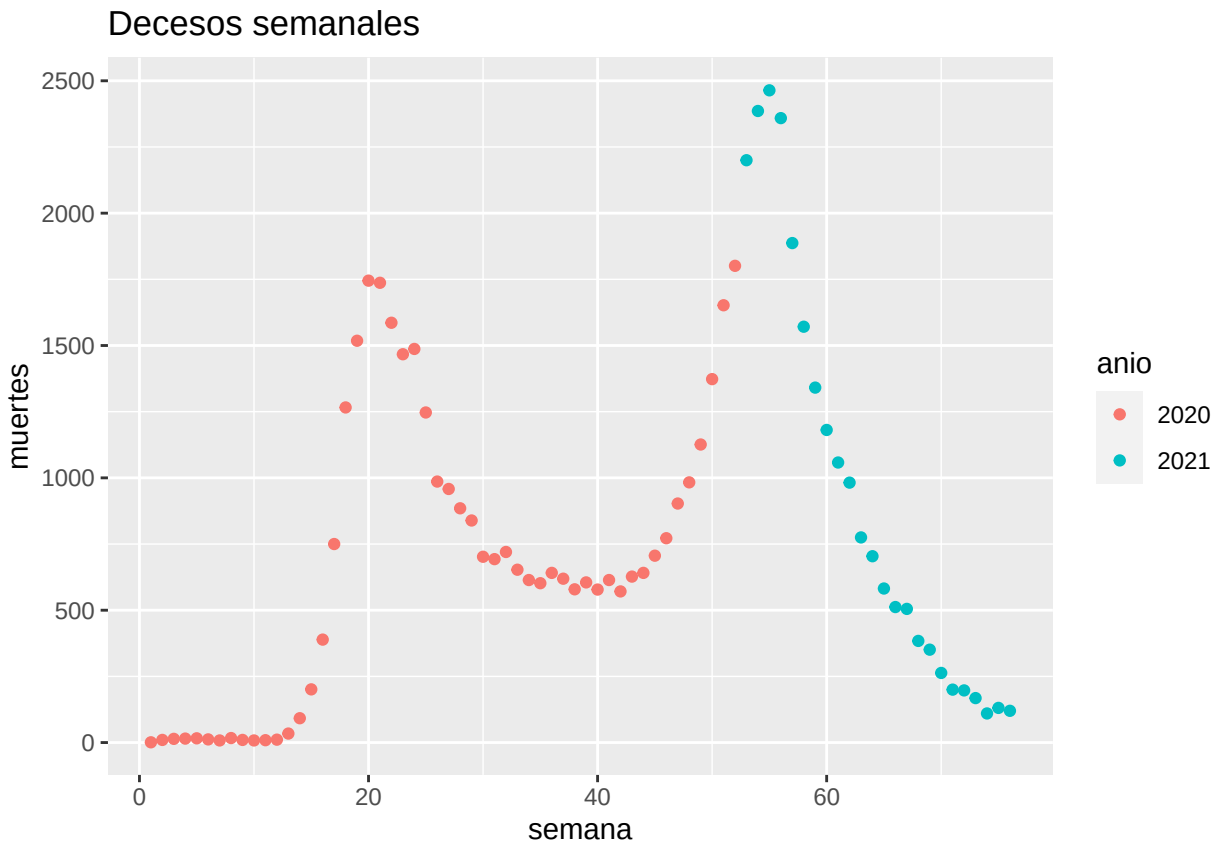


Aplicación de herramientas de clase

Interpolación, modelado y error

La primeras herramientas que usaremos serán aquellas de interpolación. En esta sección probaremos distintos modelos con polinomios de distinto grado con el fin de encontrar aquel que logre la predicción de los decesos de la última semana con el menor error cuadrático medio. El objetivo es poder ajustar un modelo de manera exitosa a la curva de decesos semanales que se presenta a continuación.

```
ggplot(df6) + geom_point(aes(x = semana, y = muertes, color = anio)) +
    ggtitle("Decesos semanales")
```

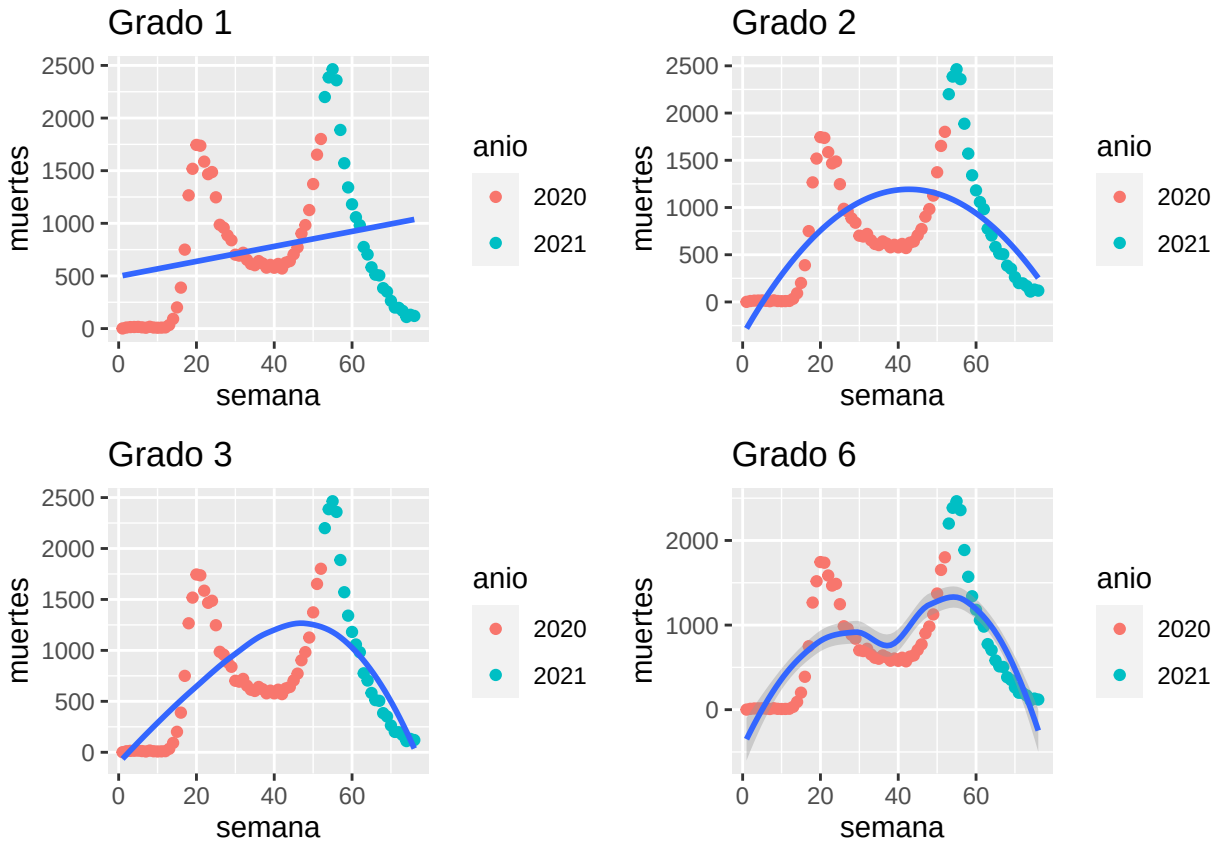


A continuación se muestran los ajustes polinomiales:

```
x1 <- df6$semana
y1 <- df6$muertes
p1 <- polyfit(x1, y1, 1)  #polinomio grado 1
p2 <- polyfit(x1, y1, 2)  #polinomio grado 2
p3 <- polyfit(x1, y1, 3)  #polinomio grado 3
p6 <- polyfit(x1, y1, 6)  #polinomio grado 6

# interpolación grado 1
df6$approx1 <- polyval(p1, x1)
a <- ggplot(df6) + geom_point(aes(x = semana, y = muertes, color = anio)) +
    geom_smooth(aes(x = semana, y = approx1)) + ggtitle("Grado 1")

# interpolación grado 2
df6$approx2 <- polyval(p2, x1)
b <- ggplot(df6) + geom_point(aes(x = semana, y = muertes, color = anio)) +
    geom_smooth(aes(x = semana, y = approx2)) + ggtitle("Grado 2")

# interpolación grado 3
```

```
df6$approx3 <- polyval(p3, x1)
c <- ggplot(df6) + geom_point(aes(x = semana, y = muertes, color = anio)) +
    geom_smooth(aes(x = semana, y = approx3)) + ggtitle("Grado 3")

# interpolación grado 6
df6$approx6 <- polyval(p6, x1)
d <- ggplot(df6) + geom_point(aes(x = semana, y = muertes, color = anio)) +
    geom_smooth(aes(x = semana, y = approx6)) + ggtitle("Grado 6")


grid.arrange(a, b, c, d, ncol = 2)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Habiendo visto los modelos graficamente, el siguiente paso es realizar la predicción y medir el error. De manera concreta, vamos a hacer una extrapolación del número de muertes en la semana 77. Es decir, queremos, con base en nuestros modelos, predecir el número de muertos en la semana 77.

En la siguiente tabla mostramos el error con respecto a la predicción de la semana 77 así como el error cuadrático medio del ajuste del modelo con respecto a todas las semanas.

```
Est_p1 <- polyval(p1, 77)
Est_p2 <- polyval(p2, 77)
Est_p3 <- polyval(p3, 77)
Est_p6 <- polyval(p6, 77)

abs_errors <- c(Est_p1, Est_p2, Est_p3, Est_p6)  #error en termino de decesos

# definimos la funcion error cuadratico medio
M_S_E <- function(est, obj) {
    sqsum = 0
    for (i in 1:length(est)) {
        sqsum <- (obj[i] - est[i])^2 + sqsum
        ecm <- sqsum/length(est)
    }
    print(ecm)
}

ECM_p1 <- M_S_E(df6$approx1, df6$muertes)  # ECM PARA POLINOMIO G1
```

```
## [1] 388691
```

```
ECM_p2 <- M_S_E(df6$approx2, df6$muertes)  # ECM PARA POLINOMIO G2
```

```
## [1] 255668.8
```

```
ECM_p3 <- M_S_E(df6$approx3, df6$muertes)  # ECM PARA POLINOMIO G3
```

```
## [1] 244051.9
```

```
ECM_p6 <- M_S_E(df6$approx6, df6$muertes)  # ECM PARA POLINOMIO G6
```

```
## [1] 97094.37
```

```
mse <- c(ECM_p1, ECM_p2, ECM_p3, ECM_p6)

# creo dataframe con errores
errors <- data.frame(abs_errors, mse)
# output de tabla de errores
kable(errors, booktabs = T, col.names = c("Error en decesos para semana 77",
    "Error cuadrático medio general"), caption = "Comparación de modelos") %>%
    kable_styling(position = "center", latex_options = "repeat_header")
```

Table 3. Comparación de modelos

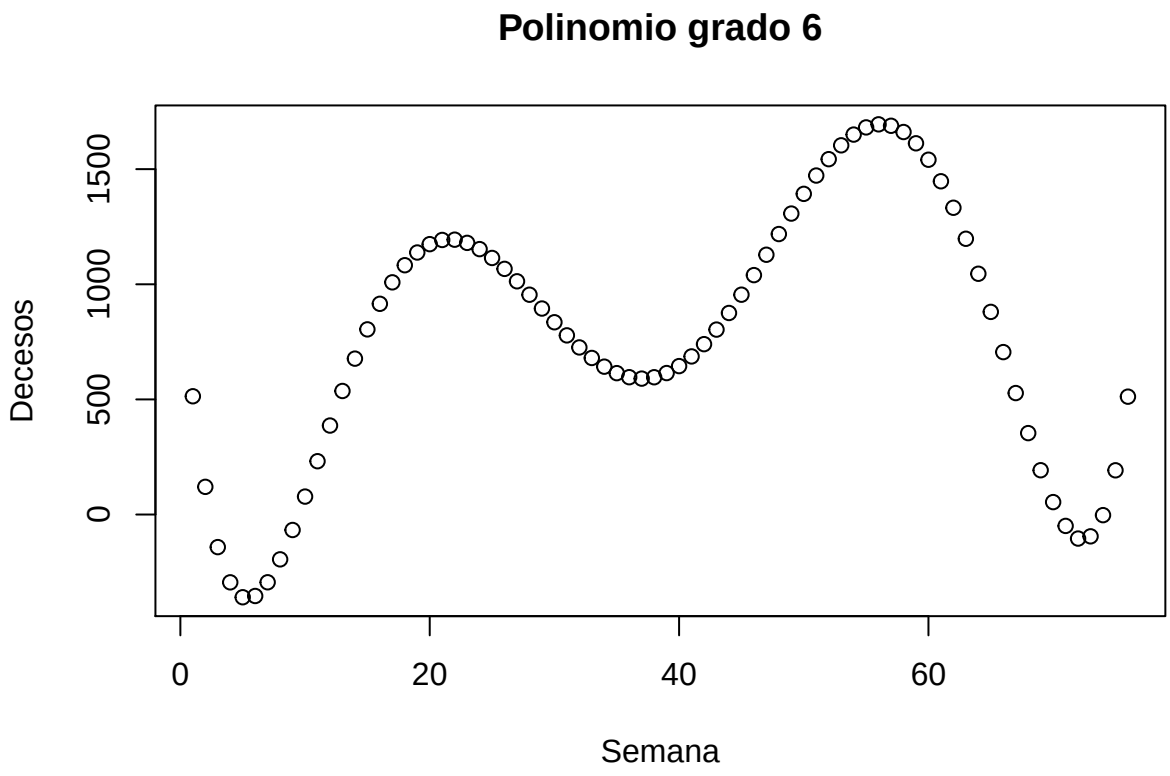
Error en decesos para semana 77	Error cuadrático medio general
1043.9653	388691.03
195.5798	255668.83
-113.0184	244051.89
982.5259	97094.37

La semana 77 tuvo 118 defunciones, por lo cual, el modelo que mejor extrapoló nuestra observación en este caso, fue el modelo del polinomio grado 2, el cual tuvo una menor diferencia en la extrapolación. Sin embargo, concluimos que la mejor estimación para nuestro modelo, es la del modelo del polinomio grado 6 ya que tiene un error cuadrático medio menor frente a los modelos de grado 1, 2 y 3.

Area bajo la curva

Necesitamos conocer el número de defunciones de la semana 1 a la 76. Supongamos que no conocemos el número de defunciones por semana, por lo cual, debemos de estimarlo a partir de la función obtenida de la interpolación del polinomio grado 6.

```
eq1 = function(x) {
    (0.00000606019424051342 * x^6) - (0.00140040473599997 * x^5) +
        (0.121907555490365 * x^4) - (4.94516796308746 * x^3) +
        (92.5593257205104 * x^2) - (638.201246371754 * x) + (1064.09435543837)
}
# la eq1 tiene los coeficientes del polinomio de grado 6 que
# se utilizaron para la interpolación.
plot(1:76, eq1(1:76), main = "Polinomio grado 6", xlab = "Semana",
    ylab = "Decesos")
```




```
n <- 6
a <- 1
b <- 76
f <- function(x) {
    (0.00000606019424051342 * x^6) - (0.00140040473599997 * x^5) +
        (0.121907555490365 * x^4) - (4.94516796308746 * x^3) +
        (92.5593257205104 * x^2) - (638.201246371754 * x) + (1064.09435543837)
}
cc <- gaussLegendre(n, a, b)
Q <- sum(cc$w * f(cc$x))
Q
```

```
## [1] 57939.74
```

```
integral <- integrate(f, a, b)
integral
```

```
## 57939.74 with absolute error < 0.00000000069
```

Si comparamos el valor de nuestra estimación de la integral, con el valor de la integral real, tenemos que el error relativo de nuestra estimación de la integral es:

```
error_reltivo_estimacion <- abs((Q - 57939.74)/57939.74) * (100)
error_reltivo_estimacion
```

```
## [1] 0.00000839447
```

El error de estimación es muy bajo, es del 0.0000084%

Por otro lado, si comparamos el valor de nuestra estimación, contra el valor real del total de suma de defunciones de nuestra base de datos, podemos obtener error relativo de nuestra aproximación, dado que

```
num_def <- sum(df2$def)  # Número total de defunciones
error_reltivo_totalDef <- abs((Q - num_def)/num_def) * (100)
error_reltivo_totalDef
```

```
## [1] 4.534791
```

El error relativo de nuestra estimación de defunciones totales es del 4.5347%

Prueba de hipotesis

Contingency Analysis

Existe la hipótesis que una persona es intubada si presenta cierto tipo de condiciones: asma, diabetes, epoc, etc. Para un nivel de significancia de .05, vamos a rechazar o aceptar la hipótesis de independencia de estas variables.

Formalmente:

$$H_0 : \text{las dos variables son independientes(estar intubado y estado de salud)}$$

$$H_1 : \text{las dos variables no son independientes}$$

```
df10 <- subset(df8, intubado != "NO APLICA")
df11 <- subset(df10, intubado != "NO ESPECIFICADO")
df_intubado <- select(df11, sexo, intubado, diabetes, epoc, asma,
    inmusupr, hipertension, obesidad, renal_cronica, tabaquismo)
df_chi <- df_intubado %>% group_by(intubado) %>% summarise(sum_diab = sum(diabetes,
    na.rm = TRUE), sum_EPOC = sum(epoc, na.rm = TRUE), sum_asma = sum(asma,
    na.rm = TRUE), sum_inmusupr = sum(inmusupr, na.rm = TRUE),
    sum_hipertension = sum(hipertension, na.rm = TRUE), sum_obesidad = sum(obesidad,
        na.rm = TRUE), sum_renal_cronica = sum(renal_cronica,
```

```
        na.rm = TRUE), tabaquismo = sum(tabaquismo, na.rm = TRUE))
df_chi
```

```
## # A tibble: 2 x 9
##   intubado sum_diab sum_EPOC sum_asma sum_inmusupr sum_hipertension sum_obesidad
##   <chr>       <dbl>    <dbl>    <dbl>        <dbl>            <dbl>        <dbl>
## 1 NO          34596     4946     3146         5778            40376        23436
## 2 SI           7613     1016      514          942             8593         5613
## # ... with 2 more variables: sum_renal_cronica <dbl>, tabaquismo <dbl>
```

```
observed_table <- matrix(c(34596, 4946, 3146, 5778, 40376, 23436,
    7618, 12620, 7613, 1016, 514, 942, 8593, 5613, 1439, 2735),
    nrow = 2, ncol = 8, byrow = T)
rownames(observed_table) <- c("NO INTUBADO", "INTUBADO")
colnames(observed_table) <- c("Diabetes", "EPOC", "Asma", "Inmunosuprimido",
    "Hipertension", "Obesidad", "Renal cronico", "Tabaquismo")
observed_table
```

```
##             Diabetes EPOC Asma Inmunosuprimido Hipertension Obesidad
## NO INTUBADO    34596 4946 3146            5778        40376    23436
## INTUBADO        7613 1016  514             942         8593     5613
##             Renal cronico Tabaquismo
## NO INTUBADO          7618      12620
## INTUBADO             1439       2735
```

```
chi_test <- chisq.test(observed_table)
chi_test
```

```
##
##  Pearson's Chi-squared test
##
## data:  observed_table
## X-squared = 175.11, df = 7, p-value < 0.00000000000000022
```

El valor p es menor que el nivel de significancia (0.05). Por tanto, podemos rechazar la hipótesis nula y concluir que las dos variables (estar intubado y condición médica) no son independientes.

Letalidad por tipo de hospital

La segunda prueba que realizaremos es probar si xiste un efecto diferenciado de la letalidad dependiendo la institución donde se hospitalizó a la persona. A continuación se muestran las tasas de letalidad por tipo de institución de salud.

```
source <- data %>% select(resultado_lab, fecha_def, sector) %>%
    filter(resultado_lab == "POSITIVO A SARS-COV-2")  #

# aqui cuento la cantidad de personas atendidas en cada
# sector
people_by_hospital <- source %>% group_by(sector) %>% summarize(count_pplo = n())
# filter(fecha_def != 'NA' )

# Aqui asumo que la persona murio si solo si hay fecha de
# defuncion, con base en eso calculo la cantidad de
# fallecidos
death_by_hospital <- source %>% filter(!is.na(fecha_def)) %>%
    group_by(sector) %>% summarize(count_dths = n())

# match
letalidad <- merge(death_by_hospital, people_by_hospital, by = "sector") %>%
    mutate(letalidad = count_dths/count_pplo)

kable(letalidad, col.names = c("Sector", "Decesos", "Personas atendidas",
    "Tasa de letalidad"), booktabs = T) %>% kable_styling(position = "center")
```

Sector	Decesos	Personas atendidas	Tasa de letalidad
CRUZ ROJA	1	87	0.0114943
ESTATAL	41	246	0.1666667
IMSS	19553	84291	0.2319702
ISSSTE	2335	7660	0.3048303
PEMEX	994	5574	0.1783280
PRIVADA	446	8520	0.0523474
SEDENA	536	2842	0.1885996
SEMAR	230	3042	0.0756082
SSA	8218	243092	0.0338061

```
## letalidad global
(global_letality <- sum(letalidad$count_dths)/sum(letalidad$count_pplo))
```

```
## [1] 0.09104724
```

Para averiguar probaremos la siguiente hipotesis

¿Es letalidad en el sector j mayor a la letalidad del país?

$$H_n : \mu_j = .091$$

$$H_a : \mu_j > .091$$

Para la prueba de hipótesis asumiremos que la distribución de deceso es binomial y partiendo de eso aprovecharemos el teorema del limite central poder realizar las pruebas de hipotesis con base en el estadístico Z.

$$Z = \frac{p - n\pi}{\sqrt{n\pi(1-\pi)}}$$

donde $p =$ letalidad $\cdot n$

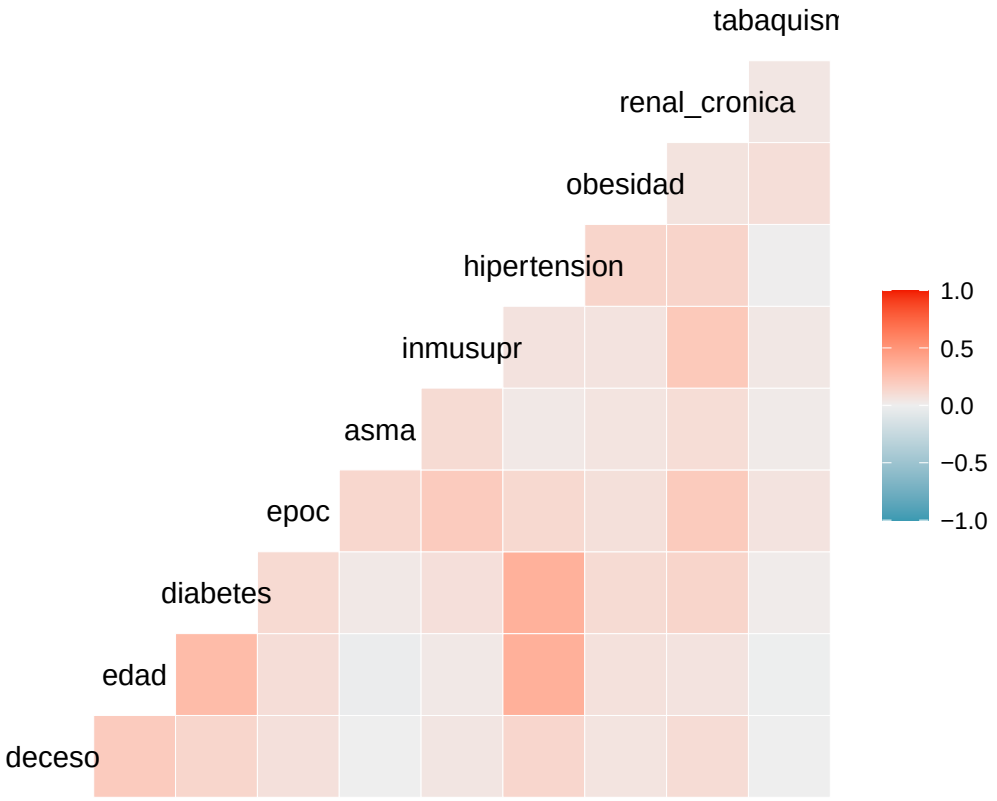
```
letalidad <- letalidad %>% mutate(esperanza = count_pplo * letalidad,
    st_let = sqrt(count_pplo * (1 - letalidad))) %>% mutate(z = (letalidad *
    count_pplo - global_letality * count_pplo)/sqrt(global_letality *
    count_pplo * (1 - global_letality)))

limit <- qnorm(0.995, 0, 1)  # for alpha=.01 one talied
letalidad <- letalidad %>% mutate(prueba = ifelse(z > limit,
    "Se rechaza H0", "No se rechaza H0"))

z_test <- letalidad %>% select(sector, z, prueba)

kable(z_test, col.names = c("Sector", "Z", "Conclusión"), booktabs = T) %>%
    kable_styling(position = "center")
```

Sector	Z	Conclusión
CRUZ ROJA	-2.579361	No se rechaza H0
ESTATAL	4.122844	Se rechaza H0
IMSS	142.222504	Se rechaza H0
ISSSTE	65.040461	Se rechaza H0
PEMEX	22.651540	Se rechaza H0
PRIVADA	-12.417229	No se rechaza H0
SEDENA	18.077804	Se rechaza H0
SEMAR	-2.960037	No se rechaza H0
SSA	-98.104586	No se rechaza H0

Correlaciones y regresión lineal

Finalmente, construiremos un modelo lineal para ver el impacto que tiene cada comorbidad en la probabilidad de perecer devido al COVID19. Primero analizamos la correlación de las variables que usaremos para construir el modelo.

```
df8 <- df8 %>% mutate(deceso = ifelse(is.na(fecha_def), 0, 1))
cor_data <- df8 %>% select(deceso, edad, diabetes, epoc, asma,
    inmusupr, hipertension, obesidad, renal_cronica, tabaquismo)
ggcorr(cor_data)
```



Notamos que la edad está altamente correlaciando con la ocurrencia del deceso; la presencia de asma y de tabaquismo, sorprendentemente, no están tan correlacioandos; el resto de las variables está medianamente correlacionado. Asimismo notamos la fuerte correlación entre edad e hipertensión y diabetes e hipertensión.

Con base en estos hallazgos, construimos un modelo de probabilidad lineal para evaluar la probabilidad de morir dependiendo cada comorbidad. En la siguietne tabla mostramos los resultados, cada coeficiente se interpreta como: la presencia de dicha comorbidad en presencia de las demas condiciones (*ceteris paribus*) incrementa la probabilidad de morir por COVID19 una vez infectado en $\beta_i \cdot 100$ puntos porcentuales. Asimismo, en cuanto a la edad, un año adicional incrementa la probabilidad de morir una vez infectado en

$\beta_{edad} \cdot 100$ puntos porcentuales.

```
model <- lm(deceso ~ ., data = cor_data)
stargazer(model, type = "latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Mon, Jul 05, 2021 - 13:44:57

Table 4

	Dependent variable:
	deceso
edad	0.001^{***}
	(0.00001)
diabetes	0.034^{***}
	(0.0003)
epoc	0.060^{***}
	(0.001)
asma	-0.013^{***}
	(0.001)
inmusupr	0.026^{***}
	(0.001)
hipertension	0.019^{***}
	(0.0003)
obesidad	0.013^{***}
	(0.0003)
renal_cronica	0.106^{***}
	(0.001)
tabaquismo	-0.004^{***}
	(0.0003)
Constant	-0.039^{***}
	(0.0002)
Observations	2,972,918
R^2	0.058
Adjusted R^2	0.058
Residual Std. Error	0.137 (df = 2972908)
F Statistic	$20{,}254.100^{***}$ (df = 9; 2972908)
Note:	*p<0.1; **p<0.05; ***p<0.01

Conclusiones

En cuanto a las conclusiones puntuales de esta practica final se encontró que:

- la presencia de diabetes, enfermedad renal crónica y epoc están altamente correlacionadas con la muerte por COVID19 una vez infectado en presencia de otras comorbidades.
- la presencia de hipertensión está altamente correlacionada con la edad y con la presencia diabetes
- los hospitales del sector ESTATAL,IMSS,ISSSTE, PEMEX y SEDENA presentan una tasa de letalidad mayor a la del país de manera estadísticamente significativa.
- las curvas de decesos, hospitalizaciones y contagios acumulados cuenta con el mismo punto de inflexión alrededor de diciembre del año pasado, sin embargo, la de decesos es la curva menos sensible de las 3.
- respecto a la mortalidad de la enfermedad respecto al sexo de las personas, nos indica que los hombres tienen una mayor defunción absoluta en el rango de edad de 58 a 73, mientras que las mujeres es en el rango de edad de 60 a 75 años.
- de acuerdo con los resultados, de las personas fallecidas por COVID-19, el 39.4% sufría de hipertensión, el 33.8% de diabetes y el 20.12% tenía obesidad. El 7.3% de las personas fallecidas tenían una condición renal crónica.
- además, de acuerdo a nuestros hallazgos, el estar intubado o no estar intubado tienen una dependencia respecto a la condición de salud de las personas hospitalizadas. Esto es consistente con la información proporcionada por la OMS: las personas que padecen afecciones médicas subyacentes, como hipertensión arterial, problemas cardíacos o pulmonares, diabetes, obesidad o cáncer, corren un mayor riesgo de presentar cuadros graves.

Es muy bien sabido que se conoce muy poco de esta enfermedad, por lo cual, es imperativo seguir investigando y analizando la información que existe, con el fin de aprender más sobre este virus. El labor de analizar grandes bases de datos tiene una importante tarea en esta crisis sanitaria: proveer de información confiable y procesada a la población para poder entender los riesgos y de cierta forma, aprender a minimizarlos, promoviendo una cultura de prevención y cuidado de la salud.

El análisis estadístico de los decesos de COVID-19 en México, nos da un panorama de la población que está en riesgo que fallecer según la edad y condición de salud. Nos aporta además un conocimiento sobre el tiempo que pasa una persona en promedio en el hospital antes de fallecer. Nos indica un panorama de esta enfermedad y es un resumen en general de todo el que hay detrás de esta crisis sanitaria mundial ya que detrás de cada estadística, se encuentra la vida de una persona.

Fuentes

- Khan M, Adil SF, Alkhathlan HZ, Tahir MN, Saif S, Khan M, Khan ST. COVID-19: A Global Challenge with Old History, Epidemiology and Progress So Far. Molecules. 2020. Recuperado el 03 de julio de 2021 en https://pubmed.ncbi.nlm.nih.gov/33374759/
- Statista, 2021. Número de personas fallecidas a causa del coronavirus en el mundo al de 2 de julio de 2021, por país. Recuperado el 03 de julio de 2021 en https://es.statista.com/estadisticas/1095779/numero-de-muertes-causadas-por-el-coronavirus-de-wuhan-por-pais/
- Organización Mundial de la Salud. 2021. Información básica sobre la COVID-19 Recuperado el 03 de julio de 2021 en https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19
- The covid Pandemic
- Contingency analysis in r
- How to extract year from date in r
- Delete or drop rows in r
- How to add new columns in r
- Aggregating and analyzing data with dplyr
- Integrate function
- From String to double
- polyfit, polyval en R
- Gauss Legendre
- Spline Function
- How to change the code ues to 1 in r
- Replace NA values with zeros
- Options of barplot