Predicción de mortalidad

Edson Bernal Mejía

Predicción de mortalidad por COVID-19

A continuación se realizará un modelo de regresión logística para obtener la probabilidad que tiene una persona de morir por COVID-19 a partir de algunas características individuales y su estado de salud, las variables consideradas son las siguientes:

- SEXO
- TIPO_PACIENTE: Es decir, si es un paciente con tratamiento ambulatorio u hospitalizado
- NEUMONIA: Si presentó un cuadro de neuomonía o no
- EDAD
- DIABETES: Si tiene el diagnóstico de Diabetes Mellitus
- EPOC: Si tiene antecedentes de Enfermedad Pulmonar Obstructiva Crónica
- ASMA: Si cuenta con el diagnóstico de asma
- INMUSUPR: Si presenta alguna enfermedad que condicione una inmunosupresión
- HIPERTENSION: Si presenta hipertensión arterial
- CARDIOVASCULAR: Si tiene antecedentes de alguna enfermedad cardiaca, sobre todo aquella que presenta riesgo de infarto
- OBESIDAD: Si tiene un índice de masa corporal mayor de 30
- RENAL_CRONICA: Si cuenta con el antecedente de enfermedad renal crónica o se encuentre en diálisis o hemodiálisis
- TABAQUISMO: Si actualmente fuma

Importación de tabla de datos

Se usará el comando **read_csv** para importar la tabla de formato excel en el objeto **casos_covid19** y se mostrará un previo de los datos contenidos:

```
casos_covid19 <- read_csv("C:/Users/berna/Desktop/Portafolio/221227COVID19MEXICO.csv")
```

```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
clean_names(casos_covid19)
```

```
## # A tibble: 6,330,966 x 40
##    fecha_actualiza~1 id_re~2 origen sector entid~3  sexo entid~4 entid~5 munic~6
##    <date>            <chr>    <dbl>  <dbl> <chr>   <dbl> <chr>   <chr>   <chr>
##  1 2022-12-27        10e0db       1     12 20          2 20      20      067
##  2 2022-12-27        0989f5       2     12 14          1 32      14      071
##  3 2022-12-27        01e27d       2      9 25          2 25      25      001
##  4 2022-12-27        180725       2      9 09          2 09      09      012
##  5 2022-12-27        0793b8       2     12 09          2 09      09      010
##  6 2022-12-27        1a4a8d       1     12 23          2 27      23      008
##  7 2022-12-27        1933c0       1     12 09          2 09      09      007
##  8 2022-12-27        1e6dad       1     12 07          1 06      07      012
##  9 2022-12-27        08c5f9       1      3 15          1 09      15      060
## 10 2022-12-27        045795       1     12 09          2 09      09      010
## # ... with 6,330,956 more rows, 31 more variables: tipo_paciente <dbl>,
## #   fecha_ingreso <date>, fecha_sintomas <date>, fecha_def <date>,
## #   intubado <dbl>, neumonia <dbl>, edad <dbl>, nacionalidad <dbl>,
## #   embarazo <dbl>, habla_lengua_indig <dbl>, indigena <dbl>, diabetes <dbl>,
## #   epoc <dbl>, asma <dbl>, inmusupr <dbl>, hipertension <dbl>, otra_com <dbl>,
## #   cardiovascular <dbl>, obesidad <dbl>, renal_cronica <dbl>,
## #   tabaquismo <dbl>, otro_caso <dbl>, toma_muestra_lab <dbl>, ...
```

```
glimpse(casos_covid19)
```

```
## Rows: 6,330,966
## Columns: 40
## $ FECHA_ACTUALIZACION   <date> 2022-12-27, 2022-12-27, 2022-12-27, 2022-12-27,~
## $ ID_REGISTRO           <chr> "10e0db", "0989f5", "01e27d", "180725", "0793b8"~
## $ ORIGEN                <dbl> 1, 2, 2, 2, 2, 1, 1, 1, 1, 1, 2, 1, 1, 1, 2, 1, ~
## $ SECTOR                <dbl> 12, 12, 9, 9, 12, 12, 12, 12, 3, 12, 6, 12, 12, ~
## $ ENTIDAD_UM            <chr> "20", "14", "25", "09", "09", "23", "09", "07", ~
## $ SEXO                  <dbl> 2, 1, 2, 2, 2, 2, 2, 1, 1, 2, 2, 1, 2, 1, 2, 1, ~
## $ ENTIDAD_NAC           <chr> "20", "32", "25", "09", "09", "27", "09", "06", ~
## $ ENTIDAD_RES           <chr> "20", "14", "25", "09", "09", "23", "09", "07", ~
## $ MUNICIPIO_RES         <chr> "067", "071", "001", "012", "010", "008", "007",~
## $ TIPO_PACIENTE         <dbl> 1, 1, 1, 2, 1, 1, 1, 1, 2, 1, 2, 1, 1, 1, 1, 1, ~
## $ FECHA_INGRESO         <date> 2022-06-23, 2022-08-09, 2022-02-14, 2022-01-19,~
## $ FECHA_SINTOMAS        <date> 2022-06-21, 2022-08-06, 2022-02-14, 2022-01-17,~
## $ FECHA_DEF             <date> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ INTUBADO              <dbl> 97, 97, 97, 2, 97, 97, 97, 97, 2, 97, 2, 97, 97,~
## $ NEUMONIA              <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ EDAD                  <dbl> 28, 57, 81, 33, 43, 49, 27, 34, 63, 57, 3, 43, 6~
## $ NACIONALIDAD          <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ EMBARAZO              <dbl> 97, 2, 97, 97, 97, 97, 97, 2, 2, 97, 97, 2, 97, ~
## $ HABLA_LENGUA_INDIG    <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ INDIGENA              <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ DIABETES              <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ EPOC                  <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ ASMA                  <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ INMUSUPR              <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ HIPERTENSION          <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, ~
## $ OTRA_COM              <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ CARDIOVASCULAR        <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ OBESIDAD              <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ RENAL_CRONICA         <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ TABAQUISMO            <dbl> 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ OTRO_CASO             <dbl> 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 1, 2, 2, ~
## $ TOMA_MUESTRA_LAB      <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 2, 1, 2, 2, ~
## $ RESULTADO_LAB         <dbl> 97, 97, 97, 97, 97, 97, 97, 97, 97, 97, 4, 2, 97~
## $ TOMA_MUESTRA_ANTIGENO <dbl> 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 2, 1, 1, 1, 1, 1, ~
## $ RESULTADO_ANTIGENO    <dbl> 2, 1, 2, 2, 2, 2, 2, 97, 2, 2, 97, 2, 2, 2, 2, 1~
## $ CLASIFICACION_FINAL   <dbl> 7, 3, 7, 7, 7, 7, 7, 6, 7, 7, 5, 7, 7, 3, 7, 3, ~
## $ MIGRANTE              <dbl> 99, 99, 99, 99, 99, 99, 99, 99, 99, 99, 99, 99, ~
## $ PAIS_NACIONALIDAD     <chr> "México", "México", "México", "México", "México"~
## $ PAIS_ORIGEN           <dbl> 97, 97, 97, 97, 97, 97, 97, 97, 97, 97, 97, 97, ~
## $ UCI                   <dbl> 97, 97, 97, 2, 97, 97, 97, 97, 2, 97, 2, 97, 97,~
```

Preparación de datos

Se seleccionarán las variables que se usarán en el análisis, se filtrarán aquellas filas en las que el resultado de la prueba haya sido confirmatorio para COVID-19 y se excluirán aquellos pacientes con edad mayor de 110 años debido a errores de registro, por último, se cambiará el tipo de variable a **factor**, de aquellas variables que lo ameriten y se transformarán los valores inusuales en **NA**, con la finalidad que no sean contemplados durante el análisis

```
covid19 <- casos_covid19 %>%
  dplyr::select(SEXO,TIPO_PACIENTE,CLASIFICACION_FINAL,ENTIDAD_RES,
         NEUMONIA,EDAD,HABLA_LENGUA_INDIG,INDIGENA,DIABETES,EPOC,ASMA,
         INMUSUPR,HIPERTENSION,CARDIOVASCULAR,OBESIDAD,RENAL_CRONICA,
         TABAQUISMO, MIGRANTE, NACIONALIDAD,FECHA_DEF) %>%
  filter(CLASIFICACION_FINAL<=3 & EDAD <=110) %>%
  mutate(SEXO= factor(SEXO),
         EDAD= EDAD-mean(EDAD),
         TIPO_PACIENTE=factor(TIPO_PACIENTE),
         ENTIDAD_RES=factor(ENTIDAD_RES),
         NEUMONIA=factor(case_when(NEUMONIA=="99"~NA_real_,
                                   NEUMONIA=="1"~1,
                                   NEUMONIA=="2"~2)),
         HABLA_LENGUA_INDIG=factor(case_when(HABLA_LENGUA_INDIG=="1" ~ 1,
                                            HABLA_LENGUA_INDIG=="2" ~ 2,
                                            HABLA_LENGUA_INDIG=="99" ~ NA_real_)),
         INDIGENA=factor(case_when(INDIGENA=="1" ~ 1,
                                   INDIGENA=="2" ~ 2,
                                   INDIGENA=="99" ~ NA_real_)),
         DIABETES=factor(case_when(DIABETES=="1" ~ 1,
                                   DIABETES=="2" ~ 2,
                                   DIABETES=="98" ~ NA_real_)),
         EPOC=factor(case_when(EPOC=="1" ~ 1,
                               EPOC=="2" ~ 2,
                               EPOC=="98" ~ NA_real_)),
         ASMA=factor(case_when(ASMA=="1" ~ 1,
                               ASMA=="2" ~ 2,
                               ASMA=="98" ~ NA_real_)),
         INMUSUPR=factor(case_when(INMUSUPR=="1" ~ 1,
                                   INMUSUPR=="2" ~ 2,
                                   INMUSUPR=="98" ~ NA_real_)),
         HIPERTENSION=factor(case_when(HIPERTENSION=="1" ~ 1,
                                       HIPERTENSION=="2" ~ 2,
                                       HIPERTENSION=="98" ~ NA_real_)),
         CARDIOVASCULAR=factor(case_when(CARDIOVASCULAR=="1" ~ 1,
                                         CARDIOVASCULAR=="2" ~ 2,
                                         CARDIOVASCULAR=="98" ~ NA_real_)),
         OBESIDAD=factor(case_when(OBESIDAD=="1" ~ 1,
                                   OBESIDAD=="2" ~ 2,
                                   OBESIDAD=="98" ~ NA_real_)),
         RENAL_CRONICA=factor(case_when(RENAL_CRONICA=="1" ~ 1,
                                        RENAL_CRONICA=="2" ~ 2,
                                        RENAL_CRONICA=="98" ~ NA_real_)),
         TABAQUISMO=factor(case_when(TABAQUISMO=="1" ~ 1,
                                     TABAQUISMO=="2" ~ 2,
                                     TABAQUISMO=="98" ~ NA_real_)),
         MIGRANTE=factor(case_when(MIGRANTE=="99" ~ 2,
                                   MIGRANTE=="2" ~ 1,
                                   MIGRANTE=="1" ~ NA_real_)),
         NACIONALIDAD=factor(NACIONALIDAD),
         FECHA_DEF=factor(case_when(is.na(FECHA_DEF) ~ 0,
                                    TRUE~1))
         )
```

Contrucción de modelo de regresión logística

Se introducirán todas las variables al modelo y se seleccionarán aquellas con valor *p* menor a 0.05; también, se explorará la multicolinialidad en las variables, con la finalidad de tener un mejor modelo.

```
summary(modelocovid1 <- glm(FECHA_DEF~SEXO+TIPO_PACIENTE+NEUMONIA+EDAD+HABLA_LENGUA_INDIG+
                      INDIGENA+DIABETES+EPOC+ASMA+INMUSUPR+HIPERTENSION+CARDIOVASCULAR+OBESIDAD+
                     RENAL_CRONICA+TABAQUISMO+MIGRANTE+NACIONALIDAD,family = binomial (link = "logit"),
                    data = covid19))
```

```
##
## Call:
## glm(formula = FECHA_DEF ~ SEXO + TIPO_PACIENTE + NEUMONIA + EDAD +
##     HABLA_LENGUA_INDIG + INDIGENA + DIABETES + EPOC + ASMA +
##     INMUSUPR + HIPERTENSION + CARDIOVASCULAR + OBESIDAD + RENAL_CRONICA +
##     TABAQUISMO + MIGRANTE + NACIONALIDAD, family = binomial(link = "logit"),
##     data = covid19)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8487  -0.0019  -0.0015  -0.0012   4.9383
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.295e+01  6.305e-01 -20.543  < 2e-16 ***
## SEXO2                3.452e-01  1.706e-02  20.234  < 2e-16 ***
## TIPO_PACIENTE2       1.140e+01  4.990e-01  22.837  < 2e-16 ***
## NEUMONIA2           -9.390e-01  1.683e-02 -55.789  < 2e-16 ***
## EDAD                 3.629e-02  4.561e-04  79.577  < 2e-16 ***
## HABLA_LENGUA_INDIG2 -1.001e-01  1.294e-01  -0.773  0.43926
## INDIGENA2           -2.262e-02  1.185e-01  -0.191  0.84861
## DIABETES2           -1.359e-01  1.943e-02  -6.994 2.67e-12 ***
## EPOC2                1.580e-01  3.651e-02   4.328 1.50e-05 ***
## ASMA2                1.597e-01  6.487e-02   2.463  0.01379 *
## INMUSUPR2           -2.182e-01  4.373e-02  -4.990 6.02e-07 ***
## HIPERTENSION2       -5.784e-02  1.957e-02  -2.956  0.00312 **
## CARDIOVASCULAR2      2.708e-01  3.321e-02   8.154 3.53e-16 ***
## OBESIDAD2           -5.945e-02  2.689e-02  -2.211  0.02705 *
## RENAL_CRONICA2      -4.007e-01  2.773e-02 -14.453  < 2e-16 ***
## TABAQUISMO2          5.410e-02  3.302e-02   1.638  0.10134
## MIGRANTE2            4.021e-01  3.644e-01   1.104  0.26976
## NACIONALIDAD2       -2.947e-01  3.221e-01  -0.915  0.36029
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 282499  on 2997375  degrees of freedom
## Residual deviance:  86377  on 2997358  degrees of freedom
##   (141246 observations deleted due to missingness)
## AIC: 86413
##
## Number of Fisher Scoring iterations: 15
```

```
vif(modelocovid1)
```

```
##              SEXO      TIPO_PACIENTE           NEUMONIA              EDAD
##          1.026393           1.000550           1.015631           1.121999
## HABLA_LENGUA_INDIG           INDIGENA           DIABETES              EPOC
##          2.629446           2.630741           1.240852           1.060059
##              ASMA           INMUSUPR       HIPERTENSION     CARDIOVASCULAR
##          1.008676           1.013027           1.352774           1.052754
##          OBESIDAD      RENAL_CRONICA         TABAQUISMO           MIGRANTE
##          1.054646           1.111214           1.051510           4.552406
##       NACIONALIDAD
##          4.553416
```

Modelo final y predicción de casos nuevos

Se realiza un modelo definitivo incluyendo las variables con significancia y que cumplan los criterios de no multicolinealidad; también, se agregarán casos nuevos para obtener la probabilidad de muerte por COVID-19 usando el modelo realizado.

```
prediccion_mort_covid19 <-glm(FECHA_DEF~SEXO+TIPO_PACIENTE+NEUMONIA+EDAD+DIABETES+
                                EPOC+ASMA+INMUSUPR+HIPERTENSION+CARDIOVASCULAR+
                                OBESIDAD+RENAL_CRONICA+TABAQUISMO,
                              family = binomial,data = covid19)


paciente <- data.frame(SEXO =c("2","1","2","2"),
                       TIPO_PACIENTE = c("1","1","2","2"),
                       NEUMONIA = c("2","2","1","1"),
                       EDAD = c(29,29,80,90),
                       DIABETES =c("2","2","1","1"),
                       EPOC = c("2","2","2","1"),
                       ASMA = c("2","1","1","1"),
                       INMUSUPR = c("2","2","2","1"),
                       HIPERTENSION = c("2","2","1","1"),
                       CARDIOVASCULAR = c("2","2","2","1"),
                       OBESIDAD =c("2","2","1","1"),
                       RENAL_CRONICA =c("2","2","2","1"),
                       TABAQUISMO =c("2","2","1","1"))
```

Se agregaron 4 nuevos pacientes

Paciente 1: Hombre de 29 años, manejo ambulatorio, sin neumonía por COVID-19, sin diabetes, EPOC, asma, inmunosupresión, hipertensión, eventos cardiovasculares, enfermedad renal crónica, tabaquismo y sin obesidad.

Paciente 2: Mujer de 29 años, manejo ambulatorio, sin neumonía por COVID-19, con asma, sin diabetes, EPOC, inmunosupresión, hipertensión, eventos cardiovasculares, enfermedad renal crónica, tabaquismo y sin obesidad.

Paciente 3: Hombre de 80 años, hospitalizado, con neumonía por COVID-19, con antecedentes de diabetes, asma, hipertensión, obesidad y fumador, niega EPOC, inmunosupresión, eventos cardiovasculares previos y enfermedad renal crónica.

Paciente 4: Hombre de 90 años, hospitalizado, con neumonía por COVID-19, con antecedentes de diabetes, EPOC, asma, inmunosupresión, hipertensión arterial, infartos previos, obesidad, enfermedad renal crónica en hemodiálisis y fumador.

La probabilidad de morir de cada paciente se presenta a continuación:

```
format(predict(object = prediccion_mort_covid19, newdata = paciente, type= "response")*100, scientific = FALSE)
```

```
##               1               2               3               4
## " 0.0003774504" " 0.0002271470" "85.5267101123" "91.0460897192"
```

Paciente 1: 0.0004% Paciente 2: 0.0002% Paciente 3: 85.5% Paciente 4: 91%