Modelo de regresión logística para accidentes de tránsito en zonas urbanas y suburbanas

Elaboró: Martha Aguilar Jiménez
Fecha: Mayo de 2022

Índice

Introducción

La estadística de accidentes de tránsito terrestre en zonas urbanas y suburbanas (ATUS) es un proyecto elaborado por el Instituto Nacional de Estadística y Geografía (INEGI), con el fin de proporcionar un panorama cuantitativo sobre la incidencia de percances viales en el ámbito nacional, así como las consecuencias humanas y materiales que conllevan.

El objetivo de este trabajo es implementar un modelo de regresión logística que permita pronosticar la probabilidad de que existan víctimas mortales en los accidentes de tránsito terrestre en zonas urbanas y suburbanas, mediante las variables incluidas en la base de datos de ATUS. El presente documento se organiza en 4 secciones. La sección 1 muestra la preparación de la base de datos, la siguiente contiene el análisis exploratorio de los datos. Posteriormente, la sección 3 presenta el ajuste del modelo de regresión logística con su respectivo análisis. Por último, la sección 4 cuenta con las conclusiones de este trabajo.

1. Preparación de la base de datos

El **objetivo** de este proyecto es ajustar un modelo de regresión logística que permita pronosticar la probabilidad de que existan víctimas mortales en un accidente de tránsito, lo que denominaremos accidentes fatales de aquí en adelante, mediante la información de los accidentes de tránsito terrestre en zonas urbanas y suburbanas (ATUS) contenida en la página de INEGI.

La base de datos ATUS cuenta con información para diferentes años. Para el presente estudio se utilizan la base más reciente, que corresponde al año 2020. Dicha base puede descargarse a través del siguiente enlace: https://www.inegi.org.mx/programas/accidentes/#Microdatos

La base de datos de 2020 cuenta con 318,046 registros y 40 campos, los cuales se describen a continuación:

Cuadro 1: Descripción de las variables

Nombre del campo	Descripción	Valores
EDO	Estado	Según Entidad Federativa
MES	Mes de la información	Según mes de referencia
ANIO	Año	Año de referencia
MPIO	Clave del municipio	Según municipio de referencia
HORA	Hora del accidente	99.- No especificado
MINUTOS	Minutos del accidente	99.- No especificado
DIA	Día del mes accidente	0.- Certificado cero 32 .- No especificado
DIASEMANA	Día del accidente	0.- Certificado cero 1.- Lunes 2.- Martes 3.- Miércoles 4.- Jueves 5.- Viernes 6.- Sábado 7.- Domingo 8.- No especificado
URBANA	Zona urbana	0.- Accidente en zona suburbana 1.- Accidente en intersección 2.- Accidente en no intersección
SUBURBANA	Zona suburbana	0.- El accidente en zona urbana 1.- Accidente en camino rural 2.- Accidente en carretera estatal 3.- Accidentes en otro camino
TIPACCID	Tipo de accidente	0.- Certificado cero 1.- Colisión con vehículo automotor 2.- Colisión con peatón 3.- Colisión con animal 4.- Colisión con objeto fijo 5.- Volcadura 6.- Caída de pasajero 7.- Salida del camino 8.- Incendio 9.- Colisión con ferrocarril 10.- Colisión con motocicleta 11.- Colisión con ciclista 12.- Otro
AUTOMOVIL	Automóvil	Vehículos involucrados

Nombre del campo	Descripción	Valores
CAMPASAJ	Camioneta para pasajeros	Vehículos involucrados
MICROBUS	Microbús	Vehículos involucrados
PASCAMION	Camión urbano de pasajeros	Vehículos involucrados
OMNIBUS	Ómnibus	Vehículos involucrados
TRANVIA	Trolebús o tranvia	Vehículos involucrados
CAMIONETA	Camioneta de carga	Vehículos involucrados
CAMION	Camión de carga	Vehículos involucrados
TRACTOR	Tractor con o sin remolque	Vehículos involucrados
FERROCARRI	Ferrocarril	Vehículos involucrados
MOTOCICLET	Motocicleta	Vehículos involucrados
BICICLETA	Bicicleta	Vehículos involucrados
OTROVEHIC	Otro vehículo	Vehículos involucrados
CAUSAACCI	Causa probable	1.- Conductor
		2.- Peatón o pasajero
		3.- Falla del vehículo
		4.- Mala condición del camino
		5.- Otra
CAPAROD	Capa de rodamiento	1.- Pavimentada
		2.- No pavimentada
SEXO	Sexo del responsable	1.- Se fugó
		2.- Hombre
		3.- Mujer
ALIENTO	Aliento alcohólico	4.- Sí
		5.- No
		6.- Se ignora
CINTURON	Uso de cinturón	7.- Sí
		8.- No
		9.- Se ignora
EDAD	Edad del responsable	0 .- Se ignora por que se fugó
		99 .- No especificado
CONDMUERTO	Conductor muerto	Víctimas involucradas
CONDHERIDO	Conductor herido	Víctimas involucradas
PASAMUERTO	Pasajero muerto	Víctimas involucradas
PASAHERIDO	Pasajero herido	Víctimas involucradas
PEATMUERTO	Peatón muerto	Víctimas involucradas
PEATHERIDO	Peatón herido	Víctimas involucradas
CICLMUERTO	Ciclista muerto	Víctimas involucradas
CICLHERIDO	Ciclista herido	Víctimas involucradas
OTROMUERTO	Otro muerto	Víctimas involucradas
OTROHERIDO	Otro herido	Víctimas involucradas

El primer paso de la preparación de la base fue descartar a los registros con información no especificada (por lo general se representan con 9) o que en la variable TIPACCID son iguales a 0 (Certificado 0), ya que no cuentan con información y por lo tanto no aportan al ajuste del modelo de regresión logística. El código implementado en R para el filtrado de datos fue el siguiente:

```
## Preparación de la base ##

#Se limpia la memoria
rm(list = ls())
```

```r
# Se cargan las librerías de interés
library(foreign)

# Se establece el directorio de trabajo
setwd("C:/Regresión_log_ATUS")

# Se cargan los datos de interés
Bd<-read.dbf(file = "atus_20.DBF", as.is = FALSE)

# Se eliminan los datos con no especificado en alguna de las variables o 0 en TIPACCID
Tr<-Bd[((Bd$HORA %in% 99) | (Bd$MINUTOS %in% 99) | (Bd$DIA %in% 32) |
          (Bd$DIASEMANA %in% c(0, 8)) | (Bd$TIPACCID %in% 0) |
          (Bd$CAUSAACCI %in% 0) | (Bd$CAPAROD %in% 0) | (Bd$SEXO %in% 0) |
          (Bd$ALIENTO %in% c(0, 6)) | (Bd$CINTURON %in% c(0, 9)) |
          (Bd$EDAD %in% 99)) == FALSE, ]
```

Después de aplicar el filtro en la base de datos se obtuvieron 79,961 registros para los cuales se generó la **variable respuesta** "exito", la cual consiste en identificar si en el accidente de tránsito hubo alguna víctima mortal, ya sea conductor, pasajero, peatón, ciclista u otro tipo de involucrado. La generación de la variable respuesta se realizó mediante el siguiente código:

```r
## Preparación de la base ##
Tr[, "exito"]<-ifelse((Tr$CONDMUERTO > 0) | (Tr$PASAMUERTO > 0) |
                        (Tr$PEATMUERTO > 0) | (Tr$CICLMUERTO > 0) |
                        (Tr$OTROMUERTO > 0), 1, 0)
table(Tr$exito)
```

A partir de la variable anterior se contabilizaron 620 éxitos y 79,341 fracasos. Dado que el objetivo es implementar un modelo de regresión logística que permita pronosticar la probabilidad de identificar un accidente fatal y no la proporción o conteo de fallecidos en accidentes, se seleccionaron al azar 300 casos de éxitos y 200 para fracasos. El código implementado para la selección de los registros se muestra a continuación:

```r
# Se va a balancear el número de éxitos y fracasos de manera aleatoria
Exito<-Tr[Tr$exito == 1, ]
Fracaso<-Tr[Tr$exito == 0, ]
set.seed(9959)
Ntr<-rbind(Exito[sample(1:nrow(Exito), 300), ], Fracaso[sample(1:nrow(Fracaso), 200), ])
```

Los registros que se utilizaron para el estudio pueden consultarse en el archivo "Información del proyecto.xlsx", en la hoja con nombre "Base". Una vez que se cuenta con la base de datos para el ajuste del modelo de regresión logística, se procedió a realizar un análisis exploratorio de la información, el cual se muestra en la siguiente sección.

2. Análisis exploratorio

En el siguiente gráfico se presenta el porcentaje de accidentes fatales por grupos de horarios. En el gráfico se puede apreciar que, los mayores porcentajes de accidentes fatales se tienen en los horarios de 19 a 22 hrs, de 6 a 10 hrs y 15 a 18 hrs con un 23%, 21% y 20% respectivamente, en contraste con el resto de los horarios que cuentan cada uno de ellos con el 12% en promedio.

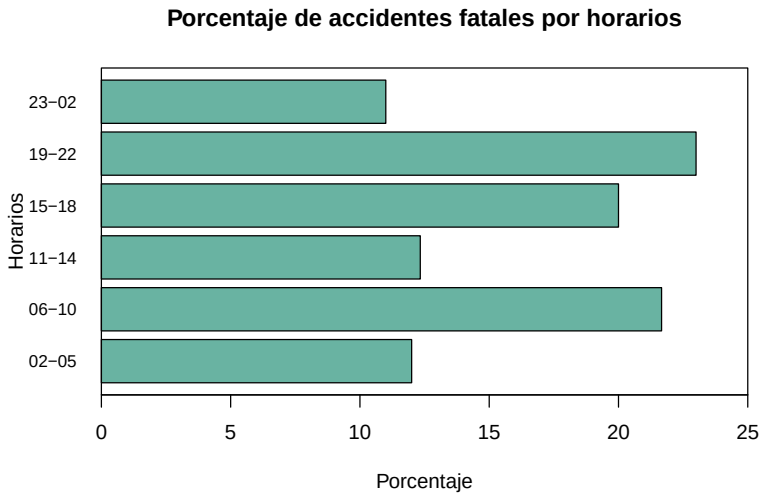

Figura 1: Análisis de la variable HORA

La siguiente gráfica muestra los grupos de edad del responsable del accidente fatal. En el grupo de edad de 16 a 30 años es donde se presenta el mayor porcentaje de accidentes fatales, con un 40.6%, en contraste con el grupo de edad de 15 o menos, el cual cuenta con el 1.7%.

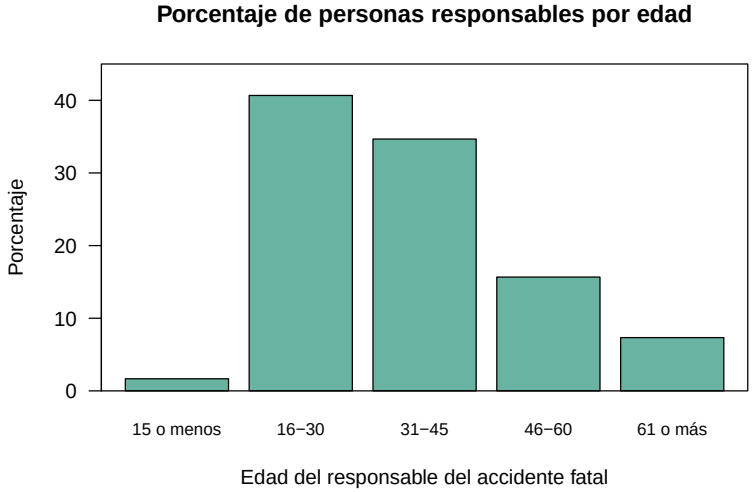

Figura 2: Análisis de la variable EDAD

La próxima gráfica muestra los accidentes fatales según la entidad federativa de ocurrencia. Como se aprecia en la gráfica, Sonora (clave 26) es la entidad federativa con más accidentes fatales, representando casi la quinta parte (17.6%) de accidentes fatales. Los siguientes estados con más accidentes fatales son Jalisco (clave 14) y Michoacán (clave 16) con el 11.3% y 7.3%, respectivamente.

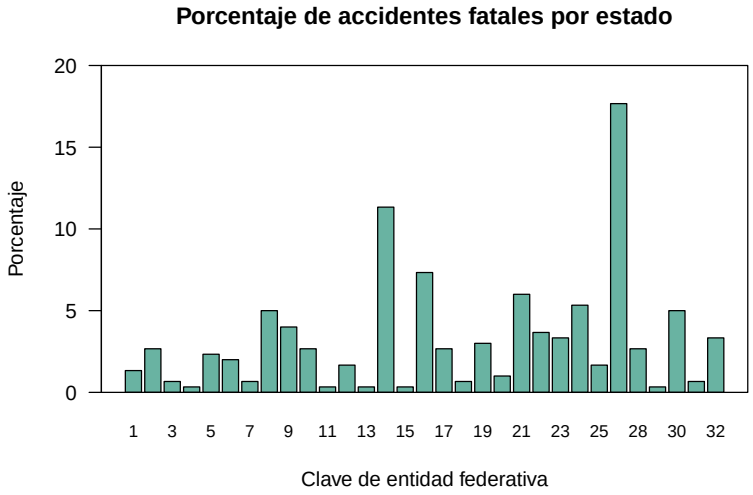

Figura 3: Análisis de la variable EDO

En la siguiente gráfica se muestra el porcentaje de accidentes fatales por tipo de accidente. En el gráfico se aprecia que, los mayores porcentajes de accidentes fatales se presentan cuando existe una colisión con motocicleta, con un peatón o con vehículo automotor, también cuando existe una volcadura, cada uno de estos tipos cuenta con el 17% de los accidentes fatales. Los menores porcentajes se presentan cuando existe una colisión con ferrocarril o con algún animal (0.3% y 1% respectivamente).

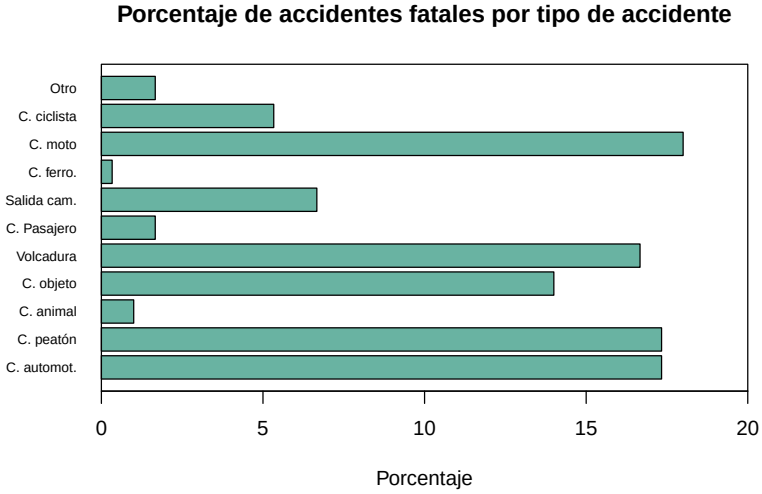

Figura 4: Análisis de la variable TIPACCID

En la siguiente gráfica se muestra el porcentaje según el tipo de zona urbana. Como se aprecia en el gráfico, poco más de la mitad de los accidentes fatales ocurren en una intersección (54%) y dos quintas partes de los siniestros con muertos ocurren en zonas suburbanas (39%); el resto (7%) ocurre en una no intersección.

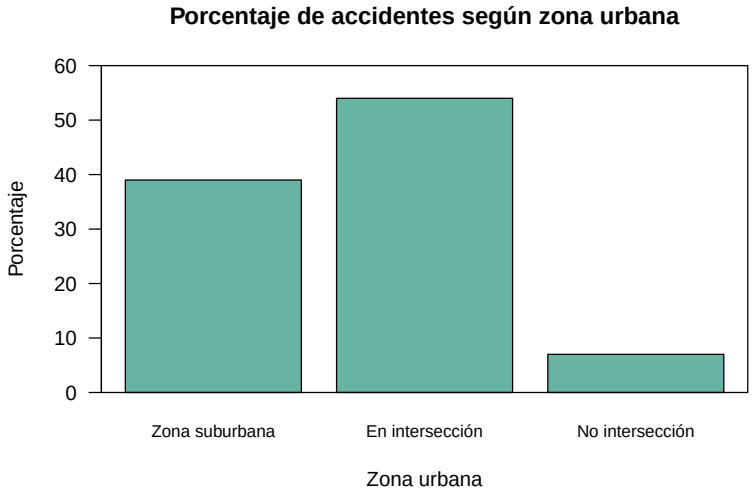

Figura 5: Análisis de la variable URBANA

Los siguientes gráficos contienen el porcentaje de accidentes fatales según el uso del cinturón de seguridad y si el responsable del accidente tenía aliento alcohólico. En el gráfico del uso de cinturón de seguridad se aprecia que, el 60% de los accidentes fatales no utilizaron cinturón de seguridad, mientras que, en el 86% de los casos, el responsable no tenía aliento alcohólico.

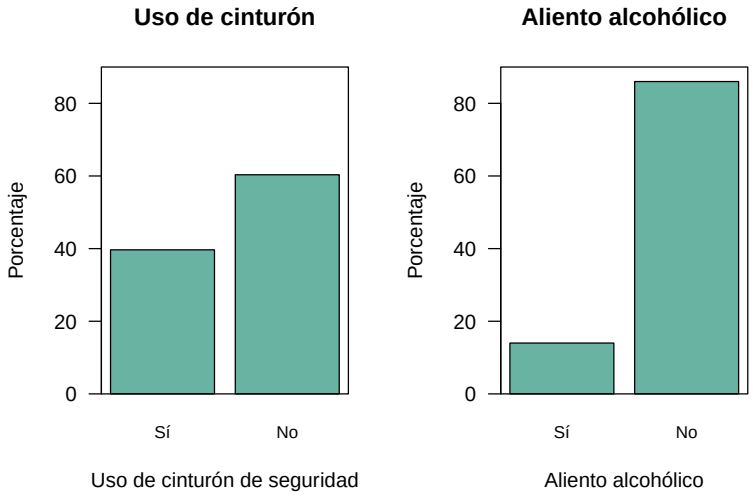

Figura 6: Análisis de las variables CINTURON y ALIENTO

Una vez que se realizó el análisis exploratorio, se concluye que, existen variables en donde ocurre una mayor cantidad de accidentes fatales en comparación con la misma categoría de la variable, por ejemplo TIPACCID, EDAD, EDO, entre otras. La siguiente parte del análisis consiste en el ajuste del modelo de regresión logística, el cual se muestra en la siguiente sección.

3. Ajuste del modelo de regresión logística

3.1 Ajuste inicial del modelo

Con la intención de realizar un comparativo del error de predicción entre un modelo lineal multivariado y un modelo con regularización LASSO, se dividió la base en 75% para datos de entrenamiento y el 25% restante para calcular el error de predicción (datos de prueba). El código implementado en R para la división de los datos puede consultarse al final del documento.

Recordando que la variable respuesta es el campo "exito" (1 = Accidente fatal, 0 = Accidente no fatal), se comenzó con el ajuste del modelo de regresión para determinar si al menos una de las variables predictoras está relacionada con la variable respuesta. Es importante mencionar que, se excluye a la variable SUBURBANA, ya que es una subcategoría del campo URBANA y hace que estas variables predictoras sean dependientes, lo que implica problemas en la estimación de los coeficientes de regresión.

Las variables consideradas como predictoras fueron: EDO, MES, HORA, DIASEMANA, URBANA, TIPACCID, CAUSAACCI, CAPAROD, SEXO, ALIENTO, CINTURON y EDAD. El resto de las variables no fueron incluidas, debido a que TIPACCID se clasifica a partir de las variables AUTOMOVIL, CAMPASAJ, MICROBUS, PASCAMION, OMNIBUS, TRANVIA, CAMIONETA, CAMION, TRACTOR, FERROCARRI, MOTOCICLET, BICICLETA y OTROVEHIC, otras 5 variables se utilizaron para la definición de la variable respuesta y el resto está relacionada con víctimas heridas.

Para esta prueba se utiliza un nivel de significancia de 0.05 ($\alpha = 0.05$). Las hipótesis son las siguientes:

H_o : $\beta_i = 0$ dado que $\beta_0 \neq 0$ para i = 1, 2, . . . , 88

H_a : Al menos un coeficiente β_i es diferente de 0. Dado que $\beta_0 \neq 0$

Esta prueba se hace a través de la diferencia de las dos devianzas (nula - residual), la cual posee una distribución chi-cuadrada con p grados de libertad, donde p es el número de parámetros del modelo sin contar el intercepto. La salida de R fue la siguiente:

```
##
## Call:
## glm(formula = cbind(conteo$exitos, conteo$fracasos) ~ EDO + MES +
##     HORA + DIASEMANA + URBANA + TIPACCID + CAUSAACCI + CAPAROD +
##     SEXO + ALIENTO + CINTURON + EDAD, family = binomial(link = logit),
##     data = conteo)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3.09204  -0.00080   0.00000   0.00031   3.08851
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.919e+00  1.128e+03  -0.009 0.992982
## EDO10        4.703e+01  7.299e+03   0.006 0.994859
## EDO11        2.169e+01  1.128e+03   0.019 0.984654
## EDO12        5.820e+01  1.011e+04   0.006 0.995405
## EDO14        2.406e+01  1.128e+03   0.021 0.982979
## EDO16        2.672e+01  1.128e+03   0.024 0.981100
## EDO17        2.520e+01  1.128e+03   0.022 0.982175
## EDO18        2.845e+01  1.135e+03   0.025 0.980004
## EDO19        1.654e+01  1.128e+03   0.015 0.988300
## EDO2         1.563e+01  1.128e+03   0.014 0.988940
## EDO20        3.902e+00  1.128e+03   0.003 0.997240
## EDO21        3.298e+01  1.128e+03   0.029 0.976672
```

```
## EDO22         7.260e-01  1.128e+03   0.001 0.999486
## EDO23         2.392e+01  1.128e+03   0.021 0.983076
## EDO24         3.294e+01  1.128e+03   0.029 0.976697
## EDO25         3.154e+01  1.293e+04   0.002 0.998054
## EDO26         3.256e+01  1.128e+03   0.029 0.976966
## EDO28         4.451e+01  6.275e+03   0.007 0.994340
## EDO29         3.400e+01  2.925e+04   0.001 0.999073
## EDO3          4.634e+01  1.766e+04   0.003 0.997907
## EDO30         9.528e+00  1.128e+03   0.008 0.993259
## EDO31         1.920e+01  1.128e+03   0.017 0.986418
## EDO32         3.524e+01  1.128e+03   0.031 0.975076
## EDO4          1.382e+01  9.808e+04   0.000 0.999888
## EDO5          2.357e+01  1.128e+03   0.021 0.983323
## EDO6          1.234e+01  1.129e+03   0.011 0.991283
## EDO7         -3.237e+01  2.530e+04  -0.001 0.998979
## EDO8          2.699e+01  1.128e+03   0.024 0.980903
## EDO9          4.275e+01  1.128e+03   0.038 0.969763
## MES10         1.251e+01  4.558e+00   2.744 0.006067 **
## MES11        -1.194e+00  2.781e+00  -0.429 0.667618
## MES12         5.319e+00  3.264e+00   1.630 0.103192
## MES2         -7.571e-01  2.769e+00  -0.273 0.784557
## MES3         -3.050e+00  2.950e+00  -1.034 0.301181
## MES4          8.622e+00  3.557e+00   2.424 0.015360 *
## MES5          7.227e+00  4.257e+00   1.698 0.089592 .
## MES6         -2.725e+00  2.687e+00  -1.014 0.310516
## MES7         -8.625e+00  3.609e+00  -2.390 0.016842 *
## MES8          4.998e+00  3.273e+00   1.527 0.126778
## MES9          7.355e+00  3.265e+00   2.253 0.024288 *
## HORA1         3.154e+01  5.937e+03   0.005 0.995761
## HORA10        2.970e+00  2.904e+00   1.023 0.306417
## HORA11       -8.558e+00  4.569e+00  -1.873 0.061043 .
## HORA12       -1.938e+01  6.663e+00  -2.909 0.003627 **
## HORA13       -2.126e+01  7.865e+00  -2.704 0.006856 **
## HORA14       -2.060e+01  6.850e+00  -3.008 0.002633 **
## HORA15       -1.284e+01  4.519e+00  -2.842 0.004490 **
## HORA16       -5.234e+00  3.030e+00  -1.727 0.084125 .
## HORA17       -9.035e+00  4.373e+00  -2.066 0.038807 *
## HORA18       -1.088e+01  4.480e+00  -2.429 0.015156 *
## HORA19       -9.569e+00  3.975e+00  -2.407 0.016081 *
## HORA2         4.501e+01  4.566e+03   0.010 0.992134
## HORA20        7.013e+00  3.960e+00   1.771 0.076522 .
## HORA21        4.582e+00  3.928e+00   1.167 0.243388
## HORA22       -2.388e+00  3.089e+00  -0.773 0.439349
## HORA23        2.064e+00  2.730e+00   0.756 0.449647
## HORA3         1.037e+01  1.395e+02   0.074 0.940753
## HORA4         3.010e+00  8.553e+00   0.352 0.724902
## HORA5         1.226e+01  4.610e+00   2.660 0.007822 **
## HORA6        -8.153e+00  3.690e+00  -2.209 0.027154 *
## HORA7         6.234e+00  5.158e+00   1.209 0.226771
## HORA8        -9.883e+00  4.406e+00  -2.243 0.024882 *
## HORA9        -3.674e+00  3.769e+01  -0.097 0.922332
## DIASEMANA2    5.044e+00  2.018e+00   2.500 0.012423 *
## DIASEMANA3    9.309e+00  2.937e+00   3.169 0.001529 **
## DIASEMANA4    3.539e+00  2.177e+00   1.626 0.103949
```

```
## DIASEMANA5  -2.805e+00  2.410e+00  -1.164 0.244592
## DIASEMANA6   4.769e+00  2.055e+00   2.321 0.020301 *
## DIASEMANA7   2.018e+00  1.978e+00   1.020 0.307584
## URBANA1     -2.513e+01  6.913e+00  -3.635 0.000278 ***
## URBANA2     -3.408e+01  9.335e+00  -3.651 0.000261 ***
## TIPACCID10   1.754e+01  4.750e+00   3.692 0.000222 ***
## TIPACCID11   6.808e+01  6.578e+03   0.010 0.991742
## TIPACCID12   1.625e+01  5.221e+00   3.112 0.001858 **
## TIPACCID2    2.181e+01  5.844e+00   3.731 0.000190 ***
## TIPACCID3   -5.145e+00  1.961e+04   0.000 0.999791
## TIPACCID4    9.837e+00  2.554e+00   3.851 0.000117 ***
## TIPACCID5    1.011e+01  3.620e+00   2.794 0.005214 **
## TIPACCID6    1.856e+01  6.281e+00   2.955 0.003126 **
## TIPACCID7   -5.090e-01  2.150e+00  -0.237 0.812810
## CAUSAACCI2   4.212e+01  4.531e+03   0.009 0.992583
## CAUSAACCI3  -1.919e+01  4.983e+01  -0.385 0.700227
## CAUSAACCI4   5.431e+01  8.514e+03   0.006 0.994910
## CAUSAACCI5  -3.532e+00  2.923e+04   0.000 0.999904
## CAPAROD2    -1.847e+01  5.943e+00  -3.108 0.001883 **
## SEXO3        3.549e-02  1.261e+00   0.028 0.977546
## ALIENTO5     3.235e-01  1.690e+00   0.191 0.848234
## CINTURON8    2.454e+00  1.355e+00   1.812 0.070016 .
## EDAD         1.514e-02  4.104e-02   0.369 0.712223
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 504.759  on 373  degrees of freedom
## Residual deviance:  80.693  on 285  degrees of freedom
## AIC: 258.69
##
## Number of Fisher Scoring iterations: 20
```

Al calcular la diferencia de las devianzas se obtiene un valor de 424.0654459, este estadístico de prueba tiene una distribución chi-cuadrada con 88 grados de libertad. La regla de decisión es rechazar H_o si el valor p es menor que el nivel de significancia. En este caso, el valor p para esta prueba es de $3.7013018 \times 10^{-45}$, lo cual es menor al nivel de significancia establecido ($\alpha = 0.05$), lo que implica que se rechace H_o, es decir, tenemos suficiente evidencia para determinar que sí existe relación entre la variable respuesta y al menos una de las variables predictoras.

Con respecto a la prueba individual para las variables predictoras, cuando se modelan en conjunto todas las seleccionadas, los niveles 4, 7, 9 y 10 de la variable MES, los niveles 5, 6, 8, 12 a 15 y 17 a 19 de la variable HORA, las categorías 2, 3 y 6 de DIASEMANA, las clases 1 y 2 de la variable URBANA, los niveles 4 a 6, 10 y 12 de la variable TIPACCID tienen un nivel de significancia de a lo mucho de 0.05, es decir, son significativas para el modelo ajustado.

Por otro lado, al obtener el coeficiente de determinación (R^2) mediante $1 - \frac{deviance}{null.deviance}$, éste arroja un valor de 0.8401349, es decir, aproximadamente el 84.01% de la variación en la variable de respuesta, se debe a las variables predictoras del modelo que se ha ajustado.

3.2 Análisis de multicolinealidad

La siguiente parte del análisis consistió en determinar si hay problemas de multicolinealidad, para ello, se calculó la matriz de correlaciones y los valores VIF del modelo inicial ajustado.

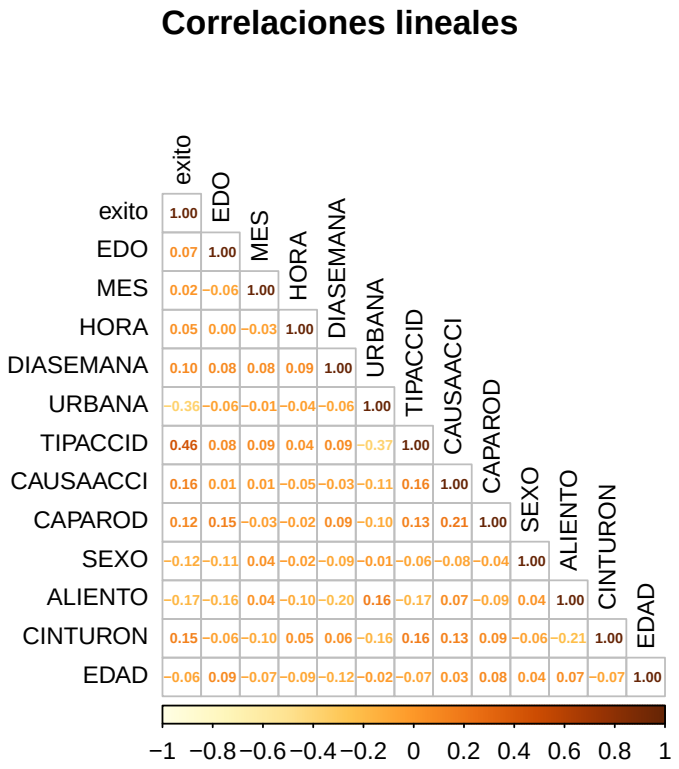

Figura 7: Matriz de correlaciones lineales de Pearson

Como se observa en la gráfica, ninguna variable tiene correlación mayor a 0.8, no obstante, las variables HORA y EDO tienen valores VIF que sobrepasan por mucho el valor de 10 (ver gráfico 1 de la figura 8). Estas variables tienen una correlación con la variable respuesta de 0.05 y de 0.07, respectivamente; la correlación lineal entre ellas es de 0. Dado que, HORA tiene una menor correlación con la variable respuesta, se optó por removerla del modelo y realizar nuevamente el ajuste sin considerar a dicha variable.

Al ajustar el modelo removiendo la variable HORA y calcular nuevamente los VIF, en el segundo gráfico de la figura 8 se puede observar que, las variables EDO, MES y TIPACCID sobrepasan el valor de 10, siendo EDO la que cuenta con un valor mucho más alto en comparación con las otras dos. Debido a lo anterior, se decidió se remover la variable EDO y realizar nuevamente el ajuste del modelo sin incluir esta variable, además de HORA.

Finalmente, al remover HORA y EDO del modelo y calcular nuevamente los VIF, podemos observar en el tercer gráfico de la figura 8, que ninguna variable predictora excede el valor de 10 en cuanto al VIF, ni se tienen variables que tengan valores cercanos a éste, siendo TIPACCID (VIF = 4.9893144) la que presenta el mayor valor en VIF. En conclusión, se remueven las variables HORA y EDO del modelo por problemas de multicolinealidad.

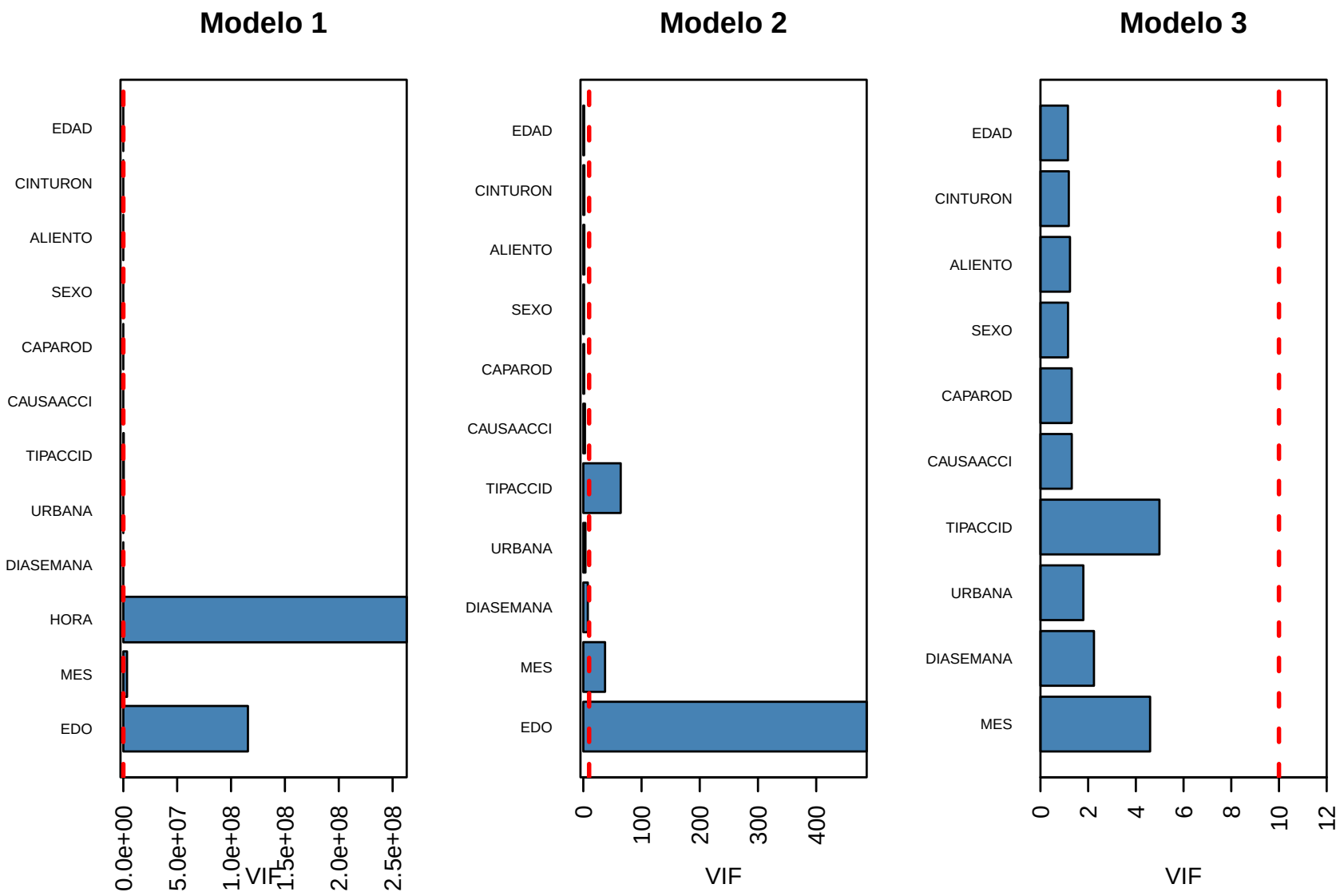

Figura 8: Valores VIF para diferentes modelos

3.3 Selección de variables

Posteriormente, se seleccionó el mejor modelo resultante de la búsqueda del mejor subconjunto usando el criterio de información de Akaike (AIC), ya que se requiere un modelo con mayor poder predictivo. Los 5 mejores modelos seleccionados con este criterio se muestran a continuación:

Cuadro 2: Variables selccionadas en los 5 mejores modelos

MES	DIASEMANA	URBANA	TIPACCID	CAUSAACCI	CAPAROD	SEXO	ALIENTO	CINTURON	EDAD	AIC
No	No	Si	Si	No	No	No	No	No	Si	307.546
No	No	Si	Si	No	No	No	No	No	No	308.166
No	No	Si	Si	Si	No	No	No	No	Si	308.314
No	No	Si	Si	No	No	No	Si	No	Si	308.643
No	No	Si	Si	No	No	No	No	Si	Si	308.698

Con base a la anterior, el mejor modelo obtuvo un estadístico AIC de 307.5462565, dicho modelo contempla a las variables predictoras URBANA, TIPACCID y EDAD. Los coeficientes de regresión estimados se muestran a continuación:

```
##
## Call:
## glm(formula = cbind(conteo_mej$exitos, conteo_mej$fracasos) ~
##     URBANA + TIPACCID + EDAD, family = binomial(link = logit),
##     data = conteo_mej)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9324  -0.4723   0.1746   0.5146   2.0970
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.80131    0.81659   3.431 0.000602 ***
## URBANA1       -3.83146    0.68922  -5.559 2.71e-08 ***
## URBANA2       -4.33351    0.85763  -5.053 4.35e-07 ***
## TIPACCID10     2.75321    0.40319   6.829 8.57e-12 ***
## TIPACCID11    19.17157 1122.80441   0.017 0.986377
## TIPACCID12     2.27861    1.23472   1.845 0.064973 .
## TIPACCID2      4.32414    0.65745   6.577 4.80e-11 ***
## TIPACCID3      0.40033    1.59051   0.252 0.801276
## TIPACCID4      2.25835    0.43844   5.151 2.59e-07 ***
## TIPACCID5      2.36299    0.60864   3.882 0.000103 ***
## TIPACCID6      2.36735    1.25709   1.883 0.059672 .
## TIPACCID7      0.96035    1.04161   0.922 0.356537
## EDAD          -0.01830    0.01141  -1.604 0.108675
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 398.70  on 246  degrees of freedom
## Residual deviance: 175.49  on 234  degrees of freedom
## AIC: 244.97
##
## Number of Fisher Scoring iterations: 16
```

Como se puede observar en la salida anterior, el modelo arroja 6 variables significativas, además del intercepto. Revisando los coeficientes de regresión se tiene lo siguiente:

1.- Para las variables urbana1 (Accidente en intersección) y urbana2 (Accidente en no intersección), los coeficientes son $\hat{\beta}_{urbana1} = -3.83146$ y $\hat{\beta}_{urbana2} = -4.33351$, respectivamente, al ser negativos disminuyen la probabilidad de éxito. Lo mismo ocurre con la variable EDAD.
2.- Para los niveles de la variable TIPACCID todos sus $\hat{\beta}_{tipaccid} > 0$, entonces, aumentan la probabilidad de éxito. El coeficiente de determinación (R^2) arroja un valor de 0.5598474, es decir, aproximadamente el 55.98% de la variación en la variable de respuesta, se debe a las variables predictoras del modelo que se ha ajustado.

3.4 Análisis de datos irregulares

El siguiente paso fue analizar si existen datos atípicos o de alto apalancamiento en el modelo que incluye a las variables URBANA, TIPACCID y EDAD.

Para decidir qué tan desviado deber ser un residual para ser considerado atípico, se pueden emplear los residuales estandarizados de Pearson, valores menores a -3 y mayores a 3 se consideran irregulares.

El siguiente gráfico muestra los valores ajustados vs residuales estandarizados de Pearson.

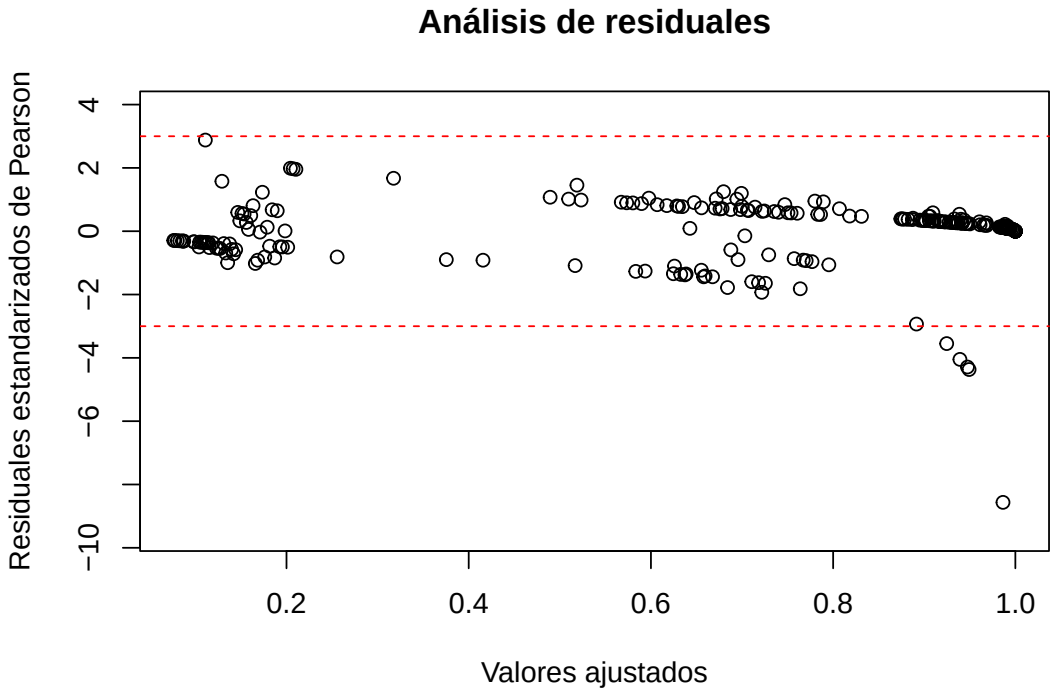

Figura 9: Análisis de datos atípicos

En el gráfico anterior, se puede observar que 5 observaciones sobrepasan el límite de -3 o 3, por lo que podrían considerarse como datos atípicos.

En cuanto al análisis para datos con alta palanca, se determinan que valores h_{ii} (matriz sombrero) son mayores a la medida 4(p + 1)/n, en donde p es el número de variables predictoras del modelo ajustado y n es el número de observaciones. Al efectuar el cálculo en R, se tienen 18 observaciones que se consideran de alta palanca.

El resultado del análisis de datos atípicos y para datos de alta palanca se resume en la siguiente tabla:

Cuadro 3: Datos atípicos y de alto apalancamiento

Tipo	Número de observación
Atípicos	59, 76, 168, 169 y 183
Alta palanca	32, 33, 158, 161, 162, 163, 164, 165, 166, 193, 214, 220, 221, 222, 223, 235, 246 y 247

Para eliminar datos atípicos y de alto apalancamiento es necesario conocer más información con respecto a la base, por ejemplo, si no hubo un error de captura en ese registro. No obstante, para fines del estudio, se optó por descartar dichos datos y ajustar nuevamente el modelo.

```
##
## Call:
## glm(formula = cbind(conteo_mej_f$exitos, conteo_mej_f$fracasos) ~
##     URBANA + TIPACCID + EDAD, family = binomial(link = logit),
##     data = conteo_mej_f)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9352  -0.4943   0.1824   0.5622   2.0567
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.6949     0.8235   3.273 0.001066 **
## URBANA1       -3.7367     0.6893  -5.421 5.94e-08 ***
## URBANA2       -4.2043     0.8598  -4.890 1.01e-06 ***
## TIPACCID10     2.6478     0.4048   6.541 6.10e-11 ***
## TIPACCID11    18.9468  1181.7791   0.016 0.987209
## TIPACCID12     2.2277     1.2331   1.807 0.070814 .
## TIPACCID2      4.2002     0.6584   6.380 1.77e-10 ***
## TIPACCID3      0.2090     1.6722   0.125 0.900556
## TIPACCID4      2.1410     0.4557   4.699 2.62e-06 ***
## TIPACCID5      2.3626     0.6600   3.580 0.000344 ***
## TIPACCID6      2.2999     1.2612   1.824 0.068213 .
## TIPACCID7      0.9081     1.0397   0.873 0.382471
## EDAD          -0.0159     0.0117  -1.360 0.173956
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 371.93  on 235  degrees of freedom
## Residual deviance: 170.09  on 223  degrees of freedom
## AIC: 237.69
##
## Number of Fisher Scoring iterations: 16
```

Al comparar este modelo sin datos irregulares con el que sí los contiene, se puede observar que se mantiene el mismo número de variables significativas, así como el nivel de significancia. De igual forma, el intercepto es significativo, también se conservan los signos de los coeficientes de regresión estimados.

El coeficiente de determinación (R^2), disminuyó un poco con respecto al anterior, ahora arroja un valor de 0.5426722, es decir, aproximadamente el 54.27% de la variación en la variable de respuesta, se debe a las variables predictoras del modelo que se ha ajustado.

3.5 Análisis de los supuestos del modelo

La siguiente parte del análisis consiste en revisar los supuestos del modelo mediante el análisis de residuales, para ello se realizaron los gráficos de los residuales del modelo.

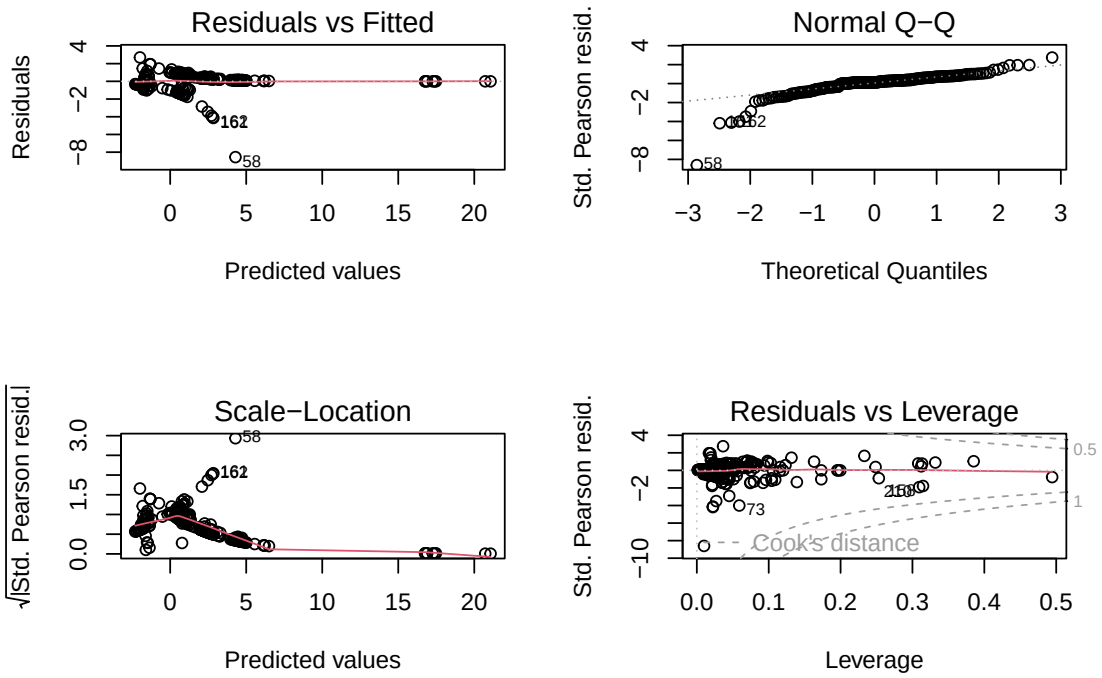

Figura 10: Análisis de residuales

La interpretación de la gráfica es la siguiente:

1.- Residuals vs Fitted: Se observa en la gráfica que existe cierta aleatoriedad, lo que indica que el modelo se ajusta medianamente.
2.- Normal Q-Q: Como podemos observar en el gráfico de qq plot los residuales cumplen el supuesto de normalidad, pues la mayoría están en el rango de [-2, 2], que es donde se concentra aproximadamente el 95% de la densidad de una distribución normal estándar.
3.-Scale-Location: En este gráfico se visualiza que a pesar de que la línea roja no es recta, los datos no siguen un patrón en particular, por lo que podemos pensar en que se cumple el supuesto de homocedasticidad.
4.- Residuals vs Leverage: Como muestra en el gráfico ninguna observación cae fuera de las distancias de Cook, por tanto no se aprecian puntos influyentes (alta palanca) en nuestro modelo de regresión.

3.6 Predicciones

Adicionalmente, se realizaron 2 predicciones con sus respectivos intervalos de confianza al 95% para dos observaciones, una de ellas contiene URBANA = “1”, TIPACCID = “1” y EDAD = 50; la otra con los valores de URBANA = “0”, TIPACCID = “1”, EDAD = 20. Los resultados fueron:

```
##   URBANA TIPACCID EDAD      pred        lwr       upr
## 1      1        1   50 0.1374189 0.08161482 0.2221493
```

```
##   URBANA TIPACCID EDAD      pred       lwr       upr
## 1      0        1   20 0.9150434 0.7253847 0.9777374
```

Con base al modelo ajustado, para la observación 1 la estimación puntual de la predicción de la probabilidad es de 0.1374189, mientras que el intervalo de confianza al 95% es de (0.08161482, 0.2221493) y su interpretación es la siguiente, se tiene una confianza del 95% de que la probabilidad de un accidente fatal se encuentre entre 8.16% y 22.2% cuando las variables predictoras URBANA = 1 (Accidente en intersección), TIPACCID = 1 (Colisión con vehículo automotor) y EDAD = 50 (edad del responsable del accidente); se tiene una probabilidad baja de que ocurra un accidente fatal con estos valores de variables.

En cambio, para la observación 2 la estimación puntual de la predicción de la probabilidad es de 0.9150434, mientras que el intervalo de confianza al 95% es de (0.7253847, 0.9777374) y su interpretación es la siguiente, se tiene una confianza del 95% de que la probabilidad de un accidente fatal se encuentre entre 72.5% y 97.8% cuando las variables predictoras URBANA = 0 (Accidente en zona suburbana), TIPACCID = 1 (Colisión con vehículo automotor) y EDAD = 50 (edad del responsable del accidente); se tiene una probabilidad alta de un accidente fatal con estos valores de variables.

3.7 Ajuste de un modelo con regularización LASSO

Para fines comparativos a cuanto el poder predictivo y de selección de variables, se optó por ajustar un modelo de regresión lineal con regularización LASSO. Es importante mencionar que, se tomaron los mismos datos de entrenamiento que el modelo anterior y que se descartaron las variables EDO y HORA por el análisis de multicolinealidad que se realizó. El ajuste del modelo se presenta a continuación:

```
## 39 x 1 sparse Matrix of class "dgCMatrix"
##                       s0
## (Intercept)  0.744849801
## MES10        0.005224288
## MES11        0.018863311
## MES12        .
## MES2         .
## MES3        -0.105540660
## MES4         .
## MES5        -0.030332450
## MES6         .
## MES7         .
## MES8         .
## MES9         .
## DIASEMANA2   .
## DIASEMANA3   0.049924953
## DIASEMANA4   .
## DIASEMANA5  -0.057296690
## DIASEMANA6   .
## DIASEMANA7   0.011480004
## URBANA1     -0.400922682
## URBANA2     -0.405118120
## TIPACCID10   0.446873730
## TIPACCID11   0.624606045
## TIPACCID12   0.277085183
## TIPACCID2    0.635015030
```

```
## TIPACCID3    .
## TIPACCID4    0.361172495
## TIPACCID5    0.310849163
## TIPACCID6    0.309144765
## TIPACCID7    0.239471691
## TIPACCID9    .
## CAUSAACCI2   .
## CAUSAACCI3  -0.056202913
## CAUSAACCI4   0.310914655
## CAUSAACCI5   .
## CAPAROD2    -0.015516295
## SEXO3       -0.018836490
## ALIENTO5    -0.041567344
## CINTURON8    0.029112785
## EDAD        -0.001968484
```

Como podemos observar en el modelo LASSO algunos coeficientes son exactamente igual a cero, a diferencia de los modelos con regularización Ridge en donde todas las variables predictoras se incluyen en el modelo final. Podemos ver el método de regularización LASSO como un método de selección de variables.

Para seleccionar con qué método de selección de variables nos quedamos, tenemos que calcular el error de predicción, usando el conjunto de prueba.

Cuadro 4: ECM para los modelos ajustados

Modelo ajustado	ECM
Modelo inicial	14.123
Modelo LASSO	0.194

Con base a la anterior, el modelo lineal con regularización LASSO genera un menor error de predicción que el ajustado en un inicio. A diferencia del modelo ajusto en el apartado 3.4, en el modelo con regularización LASSO se incluye la variable ALIENTO, CINTURON, algunos niveles de la variables CAUSAACCI, DIASEMANA y MES; coinciden en la variable URBANA y EDAD, además también en algunos niveles de TIPACCID.

4. Conclusiones

Como aprendizaje se cree que lo más complicado no fue implementar códigos en R para ajustar modelos, sino que preparar los datos y encontrar un sentido fue lo más difícil del proyecto.

En cuanto al modelo final ajustado, las variables seleccionadas fueron la edad del responsable (EDAD), el tipo de accidente (TIPACCID) y el tipo de zona urbana en donde ocurrió el accidente (URABANA). El coeficiente de determinación (R^2) para el modelo que no incluye datos irregulares (atípicos y de alto apalancamiento) fue de 54.27%

En cuanto a los supuestos del modelo, es necesario explorar más documentación relacionada con modelos de regresión logística, ya que el análisis de residuales es más complicado que el que se presenta en un modelo de regresión lineal.

Al comparar el modelo ajustado con el método de regularización LASSO, se obtuvo un menor error de predicción. Revisando las variables seleccionadas, se incluyen un mayor número, no obstante, coincide con las que se seleccionaron en el otro modelo ajustado.

Anexo. Código implementado en R

```
# Limpiar la memoria
rm(list = ls())

# Se cargan las librerías de interés
library(sqldf)
library(ISLR2)
library(dplyr)
library(foreign)
library(DescTools)
library(dplyr)
library(corrplot)
library(car)
library(bestglm)
library(ciTools)
library(xlsx)
library(glmnet)

# Establecer dirección de trabajo
setwd("C:/Regresión_log_ATUS")

### Preparación de la base ###

# Se cargan los datos de interés
# Bd<-read.dbf(file = "atus_20.DBF", as.is = FALSE)

# Se cargan los datos de interés
Bd<-read.dbf(file = "atus_20.DBF", as.is = FALSE)

# Se eliminan los datos con no especificado en alguna de las variables o 0 en TIPACCID
Tr_se<-Bd[((Bd$HORA %in% 99) | (Bd$MINUTOS %in% 99) | (Bd$DIA %in% 32) |
           (Bd$DIASEMANA %in% c(0, 8)) | (Bd$TIPACCID %in% 0) |
           (Bd$CAUSAACCI %in% 0) | (Bd$CAPAROD %in% 0) | (Bd$SEXO %in% 0) |
           (Bd$ALIENTO %in% c(0, 6)) | (Bd$CINTURON %in% c(0, 9)) |
           (Bd$EDAD %in% 99)) == FALSE, ]

## se cre la variable éxito
Tr_se[, "exito"]<-ifelse((Tr_se$CONDMUERTO > 0) | (Tr_se$PASAMUERTO > 0) |
                         (Tr_se$PEATMUERTO > 0) | (Tr_se$CICLMUERTO > 0) |
                         (Tr_se$OTROMUERTO > 0), 1, 0)

# Se va a balancear el número de éxitos y fracasos de manera aleatoria
Exito<-Tr_se[Tr_se$exito == 1, ]
Fracaso<-Tr_se[Tr_se$exito == 0, ]
set.seed(9959)
Tr<-rbind(Exito[sample(1:nrow(Exito), 300), ], Fracaso[sample(1:nrow(Fracaso), 200), ])

# Se carga la base de datos para el ajuste del modelo
Tr<-read.xlsx(file = "Información del proyecto.xlsx", sheetName = "Base")

# Se crea una copia de los datos para el análisis exploratorio
Ae<-Tr
```

```r
# Se transforman las variables que deben ser factor
Tr[, "EDO"]<-as.factor(Tr$EDO)
Tr[, "MES"]<-as.factor(Tr$MES)
Tr[, "HORA"]<-as.factor(Tr$HORA)
Tr[, "DIASEMANA"]<-as.factor(Tr$DIASEMANA)
Tr[, "URBANA"]<-as.factor(Tr$URBANA)
Tr[, "SUBURBANA"]<-as.factor(Tr$SUBURBANA)
Tr[, "TIPACCID"]<-as.factor(Tr$TIPACCID)
Tr[, "CAUSAACCI"]<-as.factor(Tr$CAUSAACCI)
Tr[, "CAPAROD"]<-as.factor(Tr$CAPAROD)
Tr[, "SEXO"]<-as.factor(Tr$SEXO)
Tr[, "ALIENTO"]<-as.factor(Tr$ALIENTO)
Tr[, "CINTURON"]<-as.factor(Tr$CINTURON)

# Datos de entrenamiento y de prueba
n<-nrow(Tr)
set.seed(20)
muestra<-sample(1:n, size = round(0.75*n), replace = FALSE)
entrenamiento<-Tr[muestra, ]
prueba<-Tr[-muestra, ]

### Análisis exploratorio ###

###### Gráfico 1

# Se convierte a numérica la variable HORA
Ae[, "HORA"]<-as.numeric(Ar$HORA)
# Se crea la consulta
consulta<-sqldf('SELECT *, CASE
                           WHEN HORA >= 19 AND HORA <= 22 THEN "19-22"
                           WHEN HORA >= 15 AND HORA <= 18 THEN "15-18"
                           WHEN HORA >= 11 AND HORA <= 14 THEN "11-14"
                           WHEN HORA >=  6 AND HORA <= 10 THEN "06-10"
                           WHEN HORA >=  2 AND HORA <=  5 THEN "02-05"
                           ELSE "23-02"
                           END AS HORA_C
                FROM Ae')

# Se crean las proporciones
prop<-sqldf('SELECT exito, HORA_C, COUNT(exito) as CONTEO
             FROM consulta
             WHERE exito = 1
             GROUP BY exito, HORA_C')
total<-sum(prop$CONTEO)

# Se crea el gráfico de barras
barplot(height = prop$CONTEO*100/300, cex.names = 0.8,
        names = prop$HORA_C, col = "#69b3a2",
        horiz = T, las = 1, xlab = "Porcentaje", ylab = "Horarios",
        main = "Porcentaje de accidentes fatales por horarios", xlim = c(0, 25))
box()

###### Gráfico 2
```

```r
# Se crea la consulta
consulta<-sqldf('SELECT *, CASE
                           WHEN EDAD <= 15 THEN "15 o menos"
                           WHEN EDAD >= 16 AND EDAD <= 30 THEN "16-30"
                           WHEN EDAD >= 31 AND EDAD <= 45 THEN "31-45"
                           WHEN EDAD >= 46 AND EDAD <= 60 THEN "46-60"
                           ELSE "61 o más"
                           END AS EDAD_C
                FROM Ae')

# Se crean las proporciones
prop<-sqldf('SELECT exito, EDAD_C, COUNT(exito) as CONTEO
             FROM consulta
             WHERE exito = 1
             GROUP BY exito, EDAD_C')

# Se crea el gráfico de barras
barplot(height = prop$CONTEO*100/300, cex.names = 0.8,
        names = prop$EDAD_C, col = "#69b3a2",
        horiz = FALSE, las = 1, ylab = "Porcentaje",
        xlab = "Edad del responsable del accidente fatal",
        main = "Porcentaje de personas responsables por edad", ylim = c(0, 45))
box()

###### Gráfico 3

# Se hace numérica la variable EDO
Ae[, "EDO"]<-as.numeric(Ar$EDO)

# Se crean las proporciones
prop<-sqldf('SELECT exito, EDO, COUNT(exito) as CONTEO
             FROM Ae
             WHERE exito = 1
             GROUP BY exito, EDO
             ORDER BY EDO')

# Se crea el gráfico de barras
barplot(height = prop$CONTEO*100/300, cex.names = 0.8,
        names = prop$EDO, col = "#69b3a2",
        horiz = FALSE, las = 1, ylab = "Porcentaje",
        xlab = "Clave de entidad federativa",
        main = "Porcentaje de accidentes fatales por estado", ylim = c(0, 20))
box()

##### Gráfico 4
# Se hace numérica la variable TIPACCID
Ae[, "TIPACCID"]<-as.numeric(Ar$TIPACCID)

# Se crean las proporciones
prop<-sqldf('SELECT exito, TIPACCID, COUNT(exito) as CONTEO
             FROM Ae
             WHERE exito = 1
             GROUP BY exito, TIPACCID
```

```
            ORDER BY TIPACCID')

# Se crea el gráfico de barras
nombre<-c("C. automot.", "C. peatón", "C. animal",
          "C. objeto", "Volcadura", "C. Pasajero",
          "Salida cam.", "C. ferro.",
          "C. moto", "C. ciclista", "Otro")
barplot(height = prop$CONTEO*100/300, cex.names = 0.68,
        names = nombre, col = "#69b3a2",
        horiz = T, las = 1, xlab = "Porcentaje", ylab = "",
        main = "Porcentaje de accidentes fatales por tipo de accidente",
        xlim = c(0, 20))
box()

#### Gráfico 5
# Se crean las proporciones
prop<-sqldf('SELECT exito, URBANA, COUNT(exito) as CONTEO
            FROM Ae
            WHERE exito = 1
            GROUP BY exito, URBANA
            ORDER BY URBANA')

# Se crea el gráfico de barras
nombre<-c("Zona suburbana", "En intersección", "No intersección")
barplot(height = prop$CONTEO*100/300, cex.names = 0.8,
        names = nombre, col = "#69b3a2",
        horiz = FALSE, las = 1, ylab = "Porcentaje", xlab = "Zona urbana",
        main = "Porcentaje de accidentes según zona urbana", ylim = c(0, 60))
box()

#### Gráfico 6 y 7
par(mfrow = c(1, 2))
# Se crean las proporciones
prop<-sqldf('SELECT exito, CINTURON, COUNT(exito) as CONTEO
            FROM Ae
            WHERE exito = 1
            GROUP BY exito, CINTURON
            ORDER BY CINTURON')

# Se crea el gráfico de barras
barplot(height = prop$CONTEO*100/300, cex.names = 0.8,
        names = c("Sí", "No"), col = "#69b3a2",
        horiz = FALSE, las = 1, ylab = "Porcentaje",
        xlab = "Uso de cinturón de seguridad",
        main = "Uso de cinturón", ylim = c(0, 90))
box()

# Se crean las proporciones
prop<-sqldf('SELECT exito, ALIENTO, COUNT(exito) as CONTEO
            FROM Ae
            WHERE exito = 1
            GROUP BY exito, ALIENTO
            ORDER BY ALIENTO')
```

```r
# Se crea el gráfico de barras
barplot(height = prop$CONTEO*100/300, cex.names = 0.8,
        names = c("Sí", "No"), col = "#69b3a2",
        horiz = FALSE, las = 1, ylab = "Porcentaje", xlab = "Aliento alcohólico",
        main = "Aliento alcohólico", ylim = c(0, 90))

### Ajuste del modelo ###

# Se realizan los conteos
conteo<-sqldf('SELECT EDO, MES, HORA, DIASEMANA, URBANA, TIPACCID,
CAUSAACCI, CAPAROD, SEXO, ALIENTO, CINTURON, EDAD,
COUNT(exito) AS ensayos, SUM(exito) AS exitos,
SUM(fracaso) AS fracasos, SUM(exito)*1.0/COUNT(exito) AS prop_exitos,
SUM(fracaso)*1.0/COUNT(exito) AS prop_fracasos
FROM entrenamiento
GROUP BY EDO, MES, HORA, DIASEMANA, URBANA, TIPACCID,
CAUSAACCI, CAPAROD, SEXO, ALIENTO, CINTURON, EDAD')

# Se ajusta el modelo logístico
modelo<-glm(cbind(conteo$exitos, conteo$fracasos) ~ EDO + MES + HORA +
              DIASEMANA + URBANA + TIPACCID +
CAUSAACCI + CAPAROD + SEXO + ALIENTO + CINTURON + EDAD, data = conteo,
family = binomial(link = logit))

# ¿Al menos alguna variable está relacionada con la respuesta?
dif_dev<-modelo$null.deviance - modelo$deviance
gl<-modelo$df.null - modelo$df.residual
valor_p<-pchisq(dif_dev, df = gl, lower.tail = FALSE)*2
modelo_r2<-1 - modelo$deviance/modelo$null.deviance
summary(modelo)

# Análisis de multicolinealidad
# Modelo 2 sin la variable HORA
# Se ajusta el modelo logístico
modelo_2<-glm(cbind(conteo$exitos, conteo$fracasos) ~ EDO + MES + DIASEMANA +
                URBANA + TIPACCID + CAUSAACCI + CAPAROD + SEXO + ALIENTO +
                CINTURON + EDAD, data = conteo,
family = binomial(link = logit))

# Modelo 3 sin la variable HORA y EDO
# Se ajusta el modelo logístico
modelo_3<-glm(cbind(conteo$exitos, conteo$fracasos) ~ MES + DIASEMANA +
                URBANA + TIPACCID + CAUSAACCI + CAPAROD + SEXO + ALIENTO +
                CINTURON + EDAD, data = conteo,
family = binomial(link = logit))

# Valores VIF
Valores<-vif(modelo)
Valores_2<-vif(modelo_2)
Valores_3<-vif(modelo_3)

# Variables de interés
Interes<-c("exito", "EDO", "MES", "HORA", "DIASEMANA", "URBANA", "TIPACCID",
```

```
          "CAUSAACCI", "CAPAROD", "SEXO", "ALIENTO", "CINTURON", "EDAD")
ent_mat<-sapply(entrenamiento[, Interes], function(x) as.numeric(x))

# Se crea la matriz de correlaciones y se gráfica
corrplot(cor(ent_mat), method = "number", type = "lower", number.cex = 0.5,
         tl.col = "black", col = COL1('YlOrBr', 200), tl.cex = 0.8,
         title = "Correlaciones lineales",
         mar = c(0, 0, 1, 0))

# Se crea la ventana de 1X3
par(mfrow = c(1, 3))

# Se crea un barplot del modelo inicial
barplot(Valores[, 1], horiz = TRUE, col = "steelblue", axes = TRUE, cex.names = 0.57,
las = 2, xlab = "VIF", ylab = "", main = "Modelo 1")
box()
abline(v = 10, lwd = 2, lty = 2, col = "red")

# Se crea un barplot del modelo 2
barplot(Valores_2[, 1], horiz = TRUE, col = "steelblue", axes = TRUE, cex.names = 0.57,
las = 2, xlab = "VIF", ylab = "", main = "Modelo 2")
box()
abline(v = 10, lwd = 2, lty = 2, col = "red")

# Se crea un barplot del modelo 3
barplot(Valores_3[, 1], horiz = TRUE, col = "steelblue", axes = TRUE, cex.names = 0.57,
las = 2, xlab = "VIF", ylab = "", xlim = c(0, 12), main = "Modelo 3")
box()
abline(v = 10, lwd = 2, lty = 2, col = "red")

# Seleccionar el mejor subconjunto de variables
x<-entrenamiento %>% select(c(MES, DIASEMANA, URBANA, TIPACCID,
                              CAUSAACCI, CAPAROD, SEXO, ALIENTO, CINTURON, EDAD))
# La variable respuesta debe ir al final
y<-entrenamiento %>% select(exito)
xy<-cbind(x, y)

# Seleccionar los 5 mejores modelos
# Se usa la función bestglm (usando AIC), mejor poder predictivo.
glm_com<-bestglm(Xy = xy, family = binomial, IC = "AIC", method = "exhaustive")
# Se listan los 5 mejores modelos
mod_5<-glm_com$BestModels

# Se realizan los conteos
conteo_mej<-sqldf('SELECT URBANA, TIPACCID, EDAD,
COUNT(exito) AS ensayos, SUM(exito) AS exitos,
SUM(fracaso) AS fracasos, SUM(exito)*1.0/COUNT(exito) AS prop_exitos,
SUM(fracaso)*1.0/COUNT(exito) AS prop_fracasos
FROM entrenamiento
GROUP BY URBANA, TIPACCID, EDAD')

# Se ajusta el modelo logístico
modelo_mej<-glm(cbind(conteo_mej$exitos, conteo_mej$fracasos) ~ URBANA +
```

```
                  TIPACCID + EDAD, data = conteo_mej,
family = binomial(link = logit))
summary(modelo_mej)
modelo_r2_mej<-1 - modelo_mej$deviance/modelo_mej$null.deviance

# Datos atípicos con el gráfico de valores ajustados vs residuales estandarizados
r_est<-residuals(modelo_mej, type = "pearson")/sqrt(1 - hatvalues(modelo_mej))
plot(predict(modelo_mej, type = "response"), r_est, xlab = "Valores ajustados",
     ylab = "Residuales estandarizados de Pearson",
     main = "Análisis de residuales", ylim = c(min(r_est) - 1, max(r_est) + 1))
abline(h = c(-3, 3), col = "red", lty = 2)

# Valores mayores o menores a la cota 3 o -3
Atipico<-which(r_est >= 3 | r_est <= -3)

# Datos con alta palanca
hii<-hatvalues(modelo_mej)

# Cota 4(p+1)/n
n_ent<-nrow(entrenamiento)
var_coef<-length(modelo_mej$coefficients)
cota<-4*(var_coef)/n_ent

# ¿Cuáles son mayores a la cota?
alta_palanca<-which(hii>cota)

# Datos sin atípicos ni alto apalancamiento
entrenamiento_f<-entrenamiento[-c(Atipico, alta_palanca), ]

# Se realizan los conteos
conteo_mej_f<-sqldf('SELECT URBANA, TIPACCID, EDAD,
COUNT(exito) AS ensayos, SUM(exito) AS exitos,
SUM(fracaso) AS fracasos, SUM(exito)*1.0/COUNT(exito) AS prop_exitos,
SUM(fracaso)*1.0/COUNT(exito) AS prop_fracasos
FROM entrenamiento_f
GROUP BY URBANA, TIPACCID, EDAD')

# Se ajusta el modelo logístico
modelo_mej_f<-glm(cbind(conteo_mej_f$exitos, conteo_mej_f$fracasos) ~ URBANA +
                    TIPACCID + EDAD, data = conteo_mej_f,
family = binomial(link = logit))
summary(modelo_mej_f)
modelo_r2_mej_f<-1 - modelo_mej_f$deviance/modelo_mej_f$null.deviance

# Análisis de residuales
par(mfrow = c(2, 2))
plot(modelo_mej_f)

# Se generar algunas predicciones
X0<-data.frame(URBANA = "1", TIPACCID = "1", EDAD = 50)
add_ci(X0, modelo_mej_f, names = c("lwr", "upr"), alpha = 0.05)
X0<-data.frame(URBANA = "0", TIPACCID = "1", EDAD = 20)
add_ci(X0, modelo_mej_f, names = c("lwr", "upr"), alpha = 0.05)
```

```r
# Modelo con regularización LASSO
# Se cargan las librerías de interés
library(glmnet)

# Se preparan las variables
X<-model.matrix(exito ~ MES + DIASEMANA + URBANA + TIPACCID + CAUSAACCI + CAPAROD +
                  SEXO + ALIENTO + CINTURON + EDAD, entrenamiento)[,-1]
Y<-entrenamiento$exito

#Fijamos una semilla y encontrar mejor lambda
set.seed(12345)
cv.mlog<-cv.glmnet(X, Y, alpha = 1, family = "binomial")

# Se obtiene el mejor lambda y los coeficientes del modelo
mejor_lambda_lasso<-cv.mlog$lambda.min
mlog_final_lasso<-glmnet(X, Y, alpha = 1, lambda = mejor_lambda_lasso)
coef(mlog_final_lasso)

# Se preparan las variables
X_prueba<-model.matrix(exito ~ MES + DIASEMANA + URBANA + TIPACCID + CAUSAACCI +
                         CAPAROD + SEXO + ALIENTO + CINTURON + EDAD, prueba)[,-1]
Y_prueba<-prueba$exito

# Obtenemos el error de predicción del modelo de regresión inicial y LASSO
normal_pred<-predict(modelo_mej_f, newx = X_prueba)
MSE_mean_n<-mean((normal_pred - Y_prueba)^2)
lasso_pred<-predict(mlog_final_lasso, s = mejor_lambda_lasso, newx = X_prueba)
MSE_mean_lasso<-mean((lasso_pred - Y_prueba)^2)
```