TREBALL DE FI DE GRAU

Grau en Estadística

Títol: Creació d'una aplicació amb Shiny per a il·lustrar el funcionament del model lineal i model de regressió logística en la docència

Autor: Marina Barber Andreo

Director: Isaac Subirana Cachinero

Departament: Genètica, Microbiologia i Estadística

Convocatòria: Juny 2023

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Resum

L'objectiu d'aquest treball és il·lustrar el funcionament d'uns tipus d'anàlisis de dades utilitzat en la docència. Per això, s'ha creat una interfície interactiva usant eines del paquet d'R Shiny per dur a terme aquest estudi amb models lineals i models logístics. Per últim, s'han analitzat i modelitzat unes dades de COVID. En aquesta base de dades es troba el nombre de casos confirmats, de morts i recuperats cada dia a tot el món. Amb l'ajut de la plataforma web, es presenten les dades i els resultats dels models lineals o logístics aplicats a aquestes dades d'una manera didàctica.

Paraules clau: Model lineal, model logístic, shiny, COVID, docència, anàlisi.

Summary

The objective of this work is to illustrate the operation of some types of data analysis used in teaching. Therefore, an interactive interface has been created using tools from the R Shiny package to carry out this study with linear models and logistic models. Finally, some COVID data have been analyzed and modeled. This database contains the number of confirmed cases, deaths and recoveries per day worldwide. With the help of the web platform, the data and the results of the linear or logistic models applied to these data are presented in a didactic way.

Key words: linear model, logistic model, shiny, COVID, teaching, analysis.

Classificació AMS:

MATHEMATICS EDUCATION:

- *97K40 Descriptive statistics*
- *97K80 Applied statistics*

STATISTICS:

- *62-07 Data análisis*
- *62-09 Graphical methods*

Agraïments

En primer lloc, voldria donar les gràcies al meu tutor per accedir a portar-me el TFG i guiar-me durant el procés. He tingut llibertat per experimentar amb el tema i provar a programar amb un nou format que no havia vist abans.
En segon lloc, m'agradaria agrair tant a familiars com als amics que han estat al meu costat durant aquesta etapa. Gràcies a aquestes persones he pogut viure al màxim l'experiència del grau.

Finalment, em sento contenta i agraïda a les Universitats i sobretot al professorat que m'han guiat fins a descobrir què és el que m'agrada i voldré dedicar-me en un futur.

Índex

Índex Figures

Índexs Taules

Índex *Listings*

Capítol 1

Introducció

Un paper fonamental per a l'estadística és la teoria i l'aplicació de models lineals. Dintre d'aquests models trobem la regressió simple, múltiple i polinòmica junt amb l'anàlisi de la variància, el disseny d'experiments, els estudis de corbes de creixement i els models log-lineals.
Normalment es tendeix a creure que el tractament del model lineal és més fàcil però en realitat és tot el contrari. Aquest model s'adapta tan bé a la natura que exigeix un estudi rigorós i amb totes les seves dimensions constitueix una especialitat de l'estadística.

D'altra banda, el mètode de regressió logística [6] és una eina matemàtica per a relacionar una variable binària (variable resposta) amb una o més variables (variables independents). A continuació, prediu el valor d'un d'aquests factors utilitzant la relació basada en l'altre. Normalment, la predicció només té dos resultats possibles, com ara sí o no.

A l'àrea de la intel·ligència artificial i el "Machine Learning"(ML), la regressió logística és una tècnica clau. Els models ML són programes de software que poden gestionar tasques difícils de processament de dades sense necessitat d'un humà. L'anàlisi predictiu pot utilitzar aquestes dades per reduir els costos operatius i augmentar la productivitat.

Finalment, aquests models lineal i logístic són més "transparents", és a dir, els resultats que proporcionen són més fàcils d'interpretar i també són més fàcils de validar. També es simplifica la detecció i correcció d'errors.

La qüestió que es planteja és: com podem implementar aquests models per facilitar la seva docència. És a dir, es busca crear una eina visual per a poder explicar millor els models de regressió lineal i logística.

La base de dades que s'utilitza en la interfície web tracta sobre la COVID. Aquest va ser un virus de nova aparició que afecta a la funció respiratòria de l'individu. Va provocar un estat d'alarma mundial a causa de les seves conseqüències, al no poder controlar-ho les autoritats van decidir fer un tancament de la majoria d'establiments junt amb una quarantena de tres mesos i altres mesures per protegir a la població. Degut a la importància que va tenir, es van dur a terme molts registres i es van recol·lectar moltes dades sobre aquest virus. Una d'aquestes bases és la que s'ha fet servir per dur a terme aquest treball.

Per dur a terme aquest estudi s'explicarà com modelitzar i analitzar les dades, de la forma més idònia; s'estudiaran les escollides a través dels models esmentats abans: els models lineals i els models de regressió logística. Aquestes es presentaran de manera més extensa dins de l'apartat de metodologies.

Aquest treball s'estructura en les següents seccions o capítols:

1. Descriptiva de les dades: aquí s'explica i es visualitzen les variables que es troben en la base de dades COVID que són amb les que es treballa en la interfície.

2. Metodologia: s'explicarà de manera detallada les dades amb les que s'ha treballat, junt amb el motiu de l'elecció d'aquestes. A més s'explicarà cada un dels models aplicats a l'anàlisi i incorporats a la interfície per poder fer-ho de manera automàtica.
3. Funcionament de la interfície interactiva web: aquí es trobarà explicat com s'utilitza l'aplicació per les dades executades en ella i les seves possibles opcions i diferents funcionalitats.
4. Resultats d'un exemple amb la base de dades COVID: s'explicarà pas a pas l'anàlisi que s'ha fet de la base i els resultats obtinguts de la modelització d'aquestes dades COVID.
5. Conclusions: per últim es farà una valoració del treball tenint en compte els resultats obtinguts. També es trobaran suggeriments per a la seva aplicació en diversos contextos i futurs canvis/millores que es puguin realitzar.

L'objectiu d'aquest treball és crear una interfície gràfica per a il·lustrar la regressió lineal i la regressió logística a partir d'una base de dades real com a exemple.

Capítol 2

2. Dades

La base de dades utilitzada s'ha extret de la pàgina web "*Kaggle*" [9] (https://www.kaggle.com/datasets/imdevskp/corona-virus-report). El repertori de dades COVID ha estat creat pel Centre de ciències i enginyeria de sistemes de la universitat Johns Hopkins (JHU CSSE), amb l'ajut del equip *"Living Atlas"* de ESRI i el laboratori de física aplicada de la universitat Johns Hopkins (JHU APL). Les fonts que han fet servir per crear aquesta base de dades ha estat la organització mundial de la salut (OMS) i els centres pel control i la prevenció de malalties de EEUU (CDC).

Aquesta base esta formada per 15 columnes i 187 files, cada una de les quals representa un país sent així. S'observa que la informació que ens dona és sobre la COVID, aquesta es un virus que afecta a la funció respiratòria. Les dades que trobem a la base són de entre els anys desembre de 2019 i març de 2023.

En aquesta s'observen variables com poden ser el nombre de casos confirmats, morts, recuperats i actius del inici de la pandèmia. Després trobem el nombre de casos, morts i recuperats de anys més propers a les dates actuals. Sobre les dades del inici podem veure diferents percentatges en forma de variables en la base, aquests són: el nombre de morts per cada 100 casos, el nombre de recuperats per cada 100 casos i per últim el nombre de morts per cada 100 recuperats. Per últim, hi ha tres variables sobre l'ultima setmana registrada, es a dir, la setmana del tres al deu de març. Per aquestes dates existeixen les variables nombre de confirmats en l'última setmana, canvis en una setmana i percentatge d'increment en una setmana.

2.1 . Anàlisi descriptiva de les dades

En aquesta secció es presenten de manera breu les principals descriptives estadístiques per així tenir un major coneixement de les dades. Es separa l'anàlisi en variables categòriques i numèriques ja que els gràfics utilitzats són diferents, en les primeres s'utilitzen gràfics de barres i en la segona histogrames.

2.1.1. Variables categòriques

En tota la base només s'observà dos variables categòriques, una anomenada "*Country Region*" i l'altra "*Who Region*". La primera parla sobre els països, dona els noms de 187 països diferents. La segona comenta de quina regió són les dades.

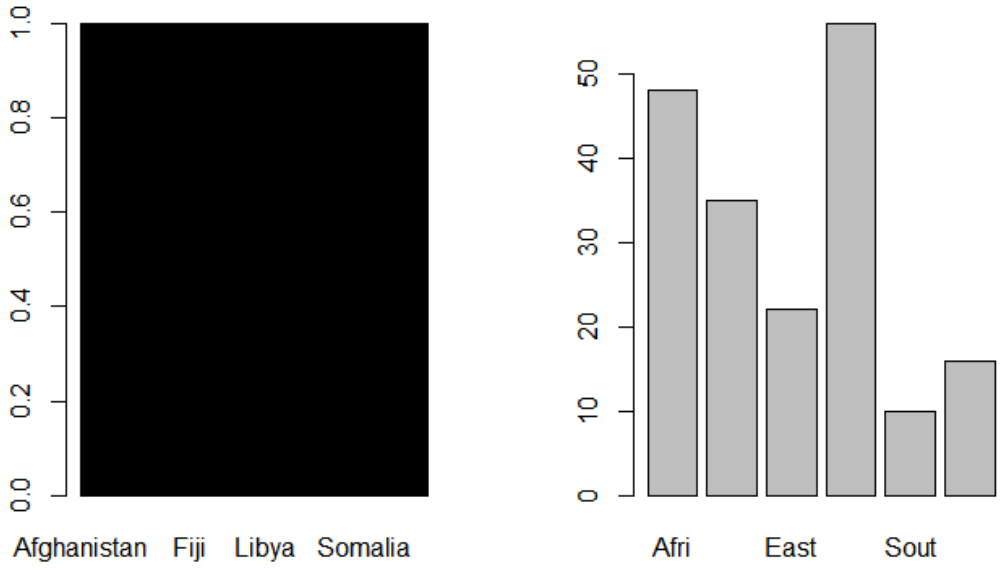

Figura 2.1: Barplots variables "Country Region" i "Who Region"

Es veu com la descriptiva sobre la variable "*Country Region*" no aporta cap informació, ja que esta composta per 187 dades diferents. Per la següent variable es veu com esta formada per 6 categories diferents, les quals indiquen la regió a la que pertany el país. Aquestes categories són en ordre*: Africa (Afri), Americas (Amer), Eastern Mediterranean (East), Europe (Euro), South-East Asia (Sout) i Western Pacific (West).* En parèntesis esta indicat el nom que trobem a la base de dades per una millor visualització i maneig de les dades. La sintaxi utilitzada per dur a terme aquest procés es la següent:

```
Who.Region<-substr(data$WHO.Region, start = 1, stop = 4)
```

Listing 2.1: Sintaxi aplicada en R per fer una abreviatura

2.1.2. Variables numèriques

Les altres 13 columnes restants son variables numèriques, les quals han estat analitzades amb gràfics de tipus histograma. Es comença amb la descriptiva de les primeres quatre variables numèriques de la base. Aquestes són: *"Confirmed", "Death", "Recovered" i "Active*", respectivament cada una parla sobre casos confirmats, morts, recuperats i actius.

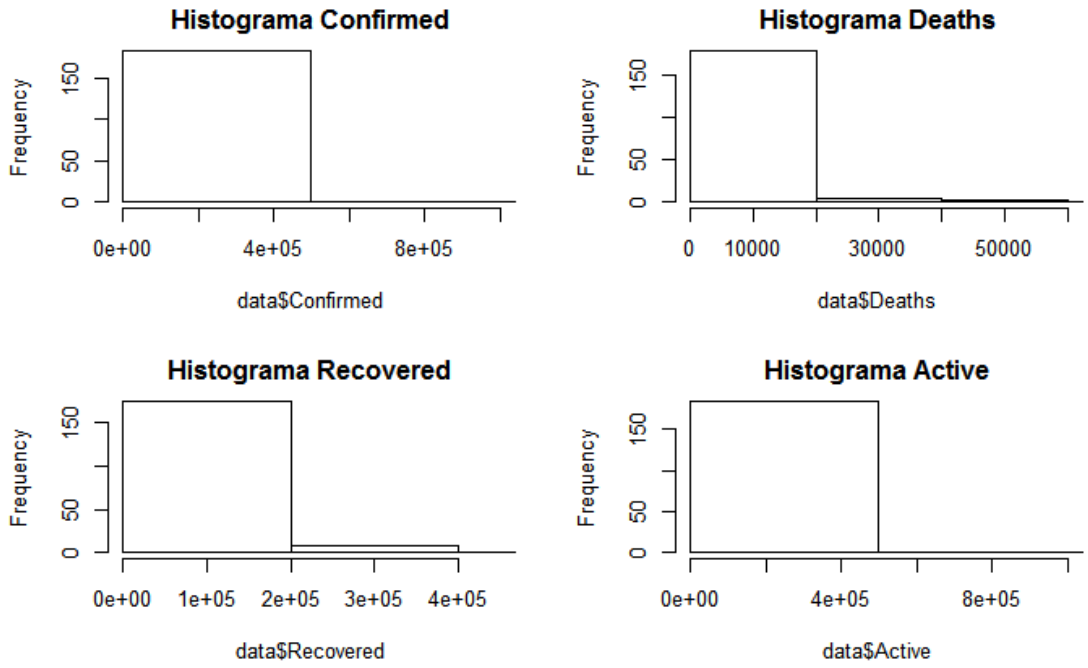

Figura 2.2: Histogrames variables *"Confirmed", "Deaths", "Recovered" i "Active"*

Es veu com la xifra de confirmats arriba a un màxim de 450000 casos amb una freqüència continua. El nombre de morts té més canvis, es troba un gran volum de dades entre les els 0 i els 20000 casos i unes freqüències més petita entre els valors 20000 i 40000. Encara que es pot veure com la freqüència més petita en aquesta variable esta entre els 40000 i 60000 casos. Amb això ens vol dir que pels diferents països que composen la base el nombre de morts ha estat elevat però ell valors més repetits són entre els 0 i els 20000 casos.

Per la variable "*Recovered"* que parla sobre els recuperats s'observa dos freqüències, la primera la més abundant durant el conjunt de dades és la que es troba entre els valors 0 i els 200000. Per últim, la variable "*Active*" parla sobre els casos actius, aquests arriben a ser un màxim de 500000 casos.

Les pròximes tres variables analitzades ens parles sobre els nous caos, les noves morts i les noves recuperacions.

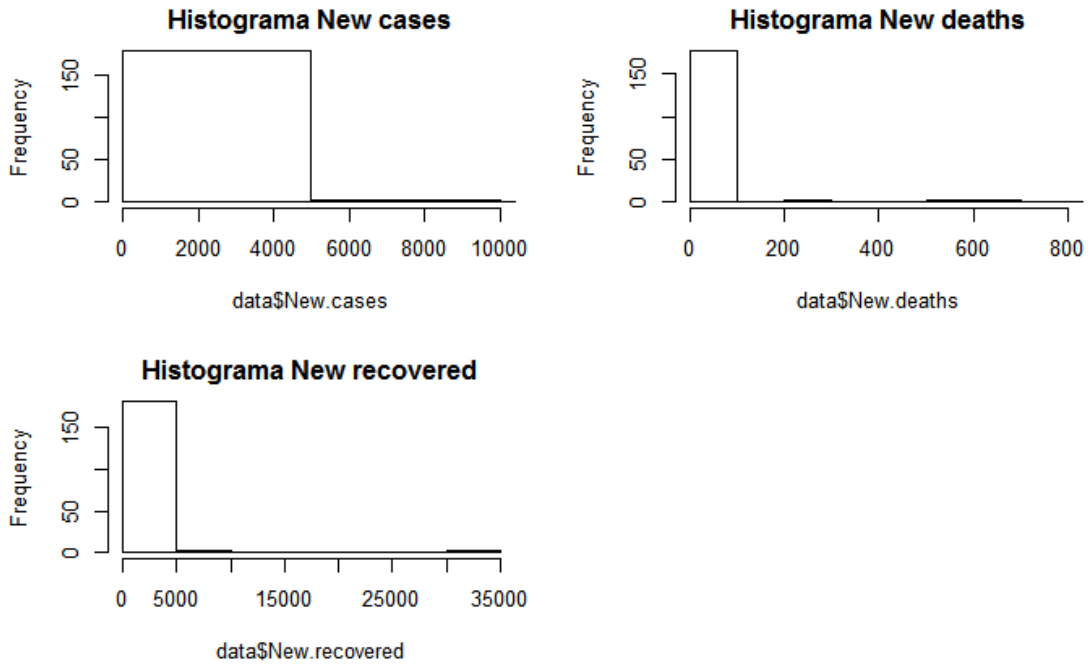

Figura 2.3: Histogrames variables *"New cases", "New deaths" i "New recovered"*

Es pot veure com en tots tres casos hi ha un conjunt de les dades que es troben en un mateix interval. En el primer cas l'interval més freqüent es entre els 0 i els 5000 casos nous, en el segon cas aquest interval es entre els 0 i els 100 morts nous, uns valors molt menors als vists anteriorment. Per últim, en els nous recuperats s'observa que casi tots els països estan entre els 0 i els 5000 casos. Cal destacar que hi ha una petita franja entre els 30000 i els 35000 casos que ens donen bones noticies. Això indica que la vacunació i les mesures imposades per les màximes autoritats en cada cas han sorgit efecte.

Ara es veuran els percentatges per cada 100000 casos de les variables principals, és a dir, morts per cada 100000 casos i recuperacions per cada 100000 casos. Per últim, hi ha la variable morts per cada 100000 recuperacions. Al contrari que per les altres variables amb aquestes si es pot fer una comparació entre els països ja que tenen en compte la grandària del país per calcular els valors.

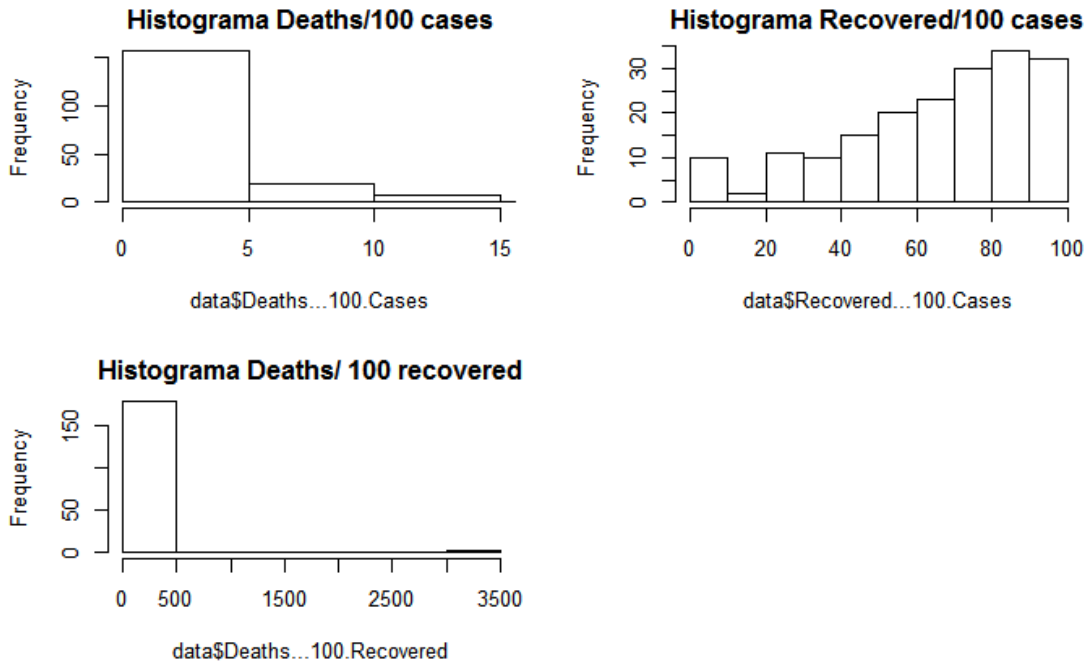

Figura 2.4: Histogrames variables *"Deaths/100 cases", "Recovered"/100 cases i "Deaths/100 recovered"*

Si es compara el nombre de morts per cada 100000 casos i el nombre de recuperats per cada 100000 casos es veu com hi ha millors valors en els recuperats. Arribant a tenir en algun cas un valor de 100000 mentre que de morts el valor màxim és del 15. Analitzant més exhaustivament el nombre de morts s'observa que són xifres no molt elevades però encara això el total és un valor molt gran. Una cosa a destacar és que la freqüència de dades més gran es troba entre el 0 i el 5. Els recuperats per cada 100000 casos té el seu màxim en el 100, encara això s'observa que els dígits més repetits són entre els 80-90, el que indica que casi tota la població es recupera.

Per últim, es troba el histograma que parla sobre el nombre de morts per cada 100000 recuperats. A l'anàlisi de les dades es veu com aquest valors és menor de 25 en casi tots els casos, els altres deu restant superen el 40 arribant a un valor màxim per aquesta variable de 3259.26.

Les tres variables restants parlen sobre els nombre de confirmats en l'última setmana, els canvis en una setmana i el percentatge d'increment en una setmana. A la base de dades es reconeixen amb els següents noms respectivament, "*Confirmed last week", "1 week change" i "1 week %increase".*

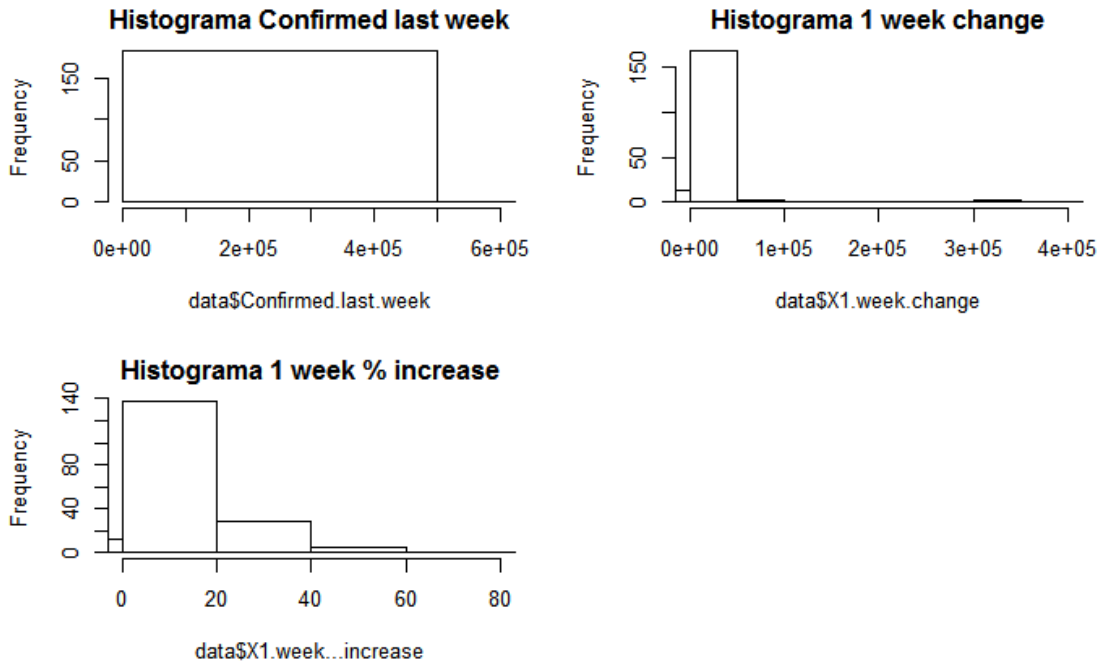

Figura 2.5: Histogrames variables *"Confirmed last week", "1-week change" i "1-week % increase"*

Es veu com per la primera variable que parla sobre el nombre de casos confirmats en l'última setmana té un valor màxim de 50000 casos. Per la següent variable trobem una petita porció dels països té un canvi negatiu, mentre que els valors més comuns són entre 0 i 50000. Per últim, la variable que calcula el % d'increment mostra com l'increment més freqüent és entre el 0 i el 20%. Es pot veure com el màxim per aquesta variable es de 60%.

Per últim, s'ha creat una nova variable binaria per realitzar la regressió logística més endavant. Aquesta nova variable s'anomena "*Death5per*" i s'ha creat a partir de la variable *"Deaths/100cases".* Aquesta indica quin d'aquests percentatges es més gran o igual de cinc amb un 1 i els que son més petits amb un 0. La seva representació gràfica seria la següent:

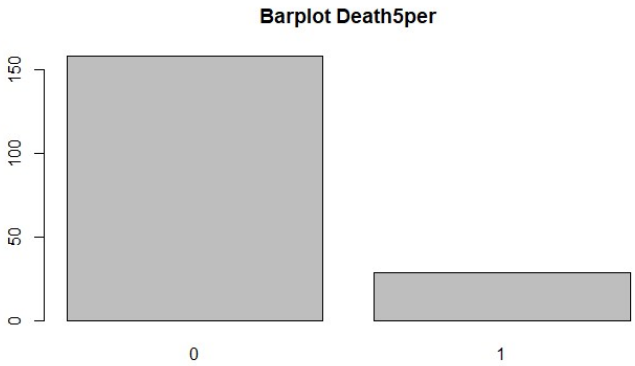

Figura 2.6: *Barplot* nova variable anomenada "*Death5per*"

Es pot afirmar que el percentatge és menor a cinc en molts dels països registrats en aquesta base. La sintaxi d'R utilitzada per crear aquesta nova variable ha estat la següent:

```
data$Death5per <- ifelse(data$Deaths...100.Cases >= 5, 1, 0)
```

Listing 2.2: Sintaxi aplicada en R crear una nova variable binaria

2.2. Resum de les dades

En aquesta secció es presenta un resum de la base de dades junt amb els principals estadístics per a cada variable. Obtingut a partir de la funció *summary* d'R el qual dona una descriptiva de cada variable segons la seva naturalesa (numèrica, factor,...).

Country.Region	Confirmed	Deaths	Recovered	Active	New.cases
Length:187	Min. : 10	Min. : 0.0	Min. : 0.0	Min. : 0.0	Min. : 0.0
Class :character	1st Qu.: 1114	1st Qu.: 18.5	1st Qu.: 626.5	1st Qu.: 141.5	1st Qu.: 4.0
Mode :character	Median : 5059	Median : 108.0	Median : 2815.0	Median : 1600.0	Median : 49.0
	Mean : 88131	Mean : 3497.5	Mean : 50631.5	Mean : 34001.9	Mean : 1223.0
	3rd Qu.: 40460	3rd Qu.: 734.0	3rd Qu.: 22606.0	3rd Qu.: 9149.0	3rd Qu.: 419.5
	Max. :4290259	Max. :148011.0	Max. :1846641.0	Max. :2816444.0	Max. :56336.0

New.deaths	New.recovered	Deaths...100.Cases	Recovered...100.Cases	Deaths...100.Recovered
Min. : 0.00	Min. : 0.0	Min. : 0.000	Min. : 0.00	Min. :0.00
1st Qu.: 0.00	1st Qu.: 0.0	1st Qu.: 0.945	1st Qu.: 48.77	1st Qu.:1.45
Median : 1.00	Median : 22.0	Median : 2.150	Median : 71.32	Median :3.62
Mean : 28.96	Mean : 933.8	Mean : 3.020	Mean : 64.82	Mean : Inf
3rd Qu.: 6.00	3rd Qu.: 221.0	3rd Qu.: 3.875	3rd Qu.: 86.89	3rd Qu.:6.44
Max. :1076.00	Max. :33728.0	Max. :28.560	Max. :100.00	Max. : Inf

Confirmed.last.week	X1.week.change	X1.week...increase	Who.Region	Death5per
Min. : 10	Min. : -47	Min. : -3.840	Length:187	Length:187
1st Qu.: 1052	1st Qu.: 49	1st Qu.: 2.775	Class :character	Class :character
Median : 5020	Median : 432	Median : 6.890	Mode :character	Mode :character
Mean : 78682	Mean : 9448	Mean : 13.606		
3rd Qu.: 37080	3rd Qu.: 3172	3rd Qu.: 16.855		
Max. :3834677	Max. :455582	Max. :226.320		

Taula 2.1: Taula resum dels estadístics principals

Fent una ullada a la base, podem veure com els valors són molt elevats en moltes de les variables, arribant en algunes d'elles a tenir un màxim infinit com és el cas de *"Deaths/100Recovered".* Aquesta ens diu que el nombre de morts es infinit en comparació de 100 casos de recuperacions. Un altre cosa a destacar sobre la variable "*Confirmed last week",* que són els casos confirmats en l'ultima setmana és que els valors són molt elevats pel moment que estem vivint junt amb les mesures que s'han posat per a que aquest virus desaparegui. Això ens indica que encara es present la COVID i hi ha gent que la pateix. Per últim, comenta que la variable "*New Deaths*" que ens indica el nou nombre de morts té un màxim de 1076 encara que es un valor molt elevat podem veure que a comparació amb les morts inicials aquest ha baixat un 137%.

Capítol 3

3. Model lineal i model de regressió logística

3.1. Definició del model lineal

3.1. 1. Condicions d'ús

Els models de regressió utilitzen l'equació següent: "observació"= "model" + "error aleatori" fixant així el model com una funció lineal d'uns paràmetres.
L'equació que defineix el model de regressió simple és la següent:

$$y_i = \beta_0 + \beta_1 x_1 + \varepsilon_i \qquad i = 1, \dots, n$$

Un model es considera lineal si ho és pels seus paràmetres.
Com a expressió general es suposa que es té una variable aleatòria Y que és igual a un valor fix η que representa la part determinista de l'experiment, més una desviació aleatòria ε que representa l'error. La seva expressió es:

$$y = \eta + \varepsilon$$

On ε és l'error i η un valor desconegut que pot representar el valor real mentre que Y seria el valor observat. Segons el model lineal s'imagina que η pren valors diferents d'acord a situacions experimentals diverses. L'error ε converteix la relació matemàtica $y = \eta$ en la relació estadística $y = \eta + \varepsilon$.
Els models amb k > 1 variables independents, predictores o regressores, s'identifiquen com models de regressió múltiple. Identifiquem aquests models amb l'expressió següent:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_k x_{ik} + \varepsilon_i \qquad i = 1, \dots, n$$

El model lineal té la anomenada variable resposta o dependent que es representa com y_i. També trobem els paràmetres β_i , aquests son desconeguts i s'estimaran a partir de les dades (mostra). Les variables que es troben al model son valors coneguts, aquestes , x_i es fixaran com variables amb valor constant.
L'expressió en forma matricial del model lineal parlat anteriorment seria:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Trobarem la seva forma resumida com: $y = X\beta + \varepsilon$. Els elements que formen el model lineal són:

-Vector d'observacions $y = (y_1, y_2, \cdots, y_n)^{`}$.
-Vector de paràmetres $\beta = (\beta_1, \beta_2, \cdots, \beta_m)^{`}$.
-Matriu del model **X** = $\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix}$
-Vector de desviacions aleatòries o errors $\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)^{`}$.

3.1.2. Ajust del model lineal: estimació dels paràmetres pel mètode del mínims quadrats

Sabem que el millor ajust es té quan menors son els residus. Aquest mètode de mínims quadrats escull de tots els possibles valors de β_j aquells que minimitzant

$$S = \sum_{i=1}^{n} \varepsilon^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}))^2$$

La regressió lineal simple [5] en forma expressió seria la següent:

$$S = \sum_{i=1}^{n} \varepsilon^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Amb aquesta equació s'aconsegueix tenir els estimadors per mínims quadrats al derivar o igualar a 0. Aquest criteri coincideix amb el de màxima versemblança si es compleixen les premisses del model que es llistaran a continuació.

3.1.3. Premisses del model lineal: Les condicions de Gauss-Markov

El model lineal definit en l'apartat anterior suposa uns errors ε, aquests es comporten com variables aleatòries sent les corresponents desviacions. Es vol que aquests errors aleatoris verifiquin les següents condicions:

1. $E(\varepsilon_i) = 0 \qquad i = 1, \dots, n$

   Amb aquesta condició ens assegurem que el model lineal no està esbiaixat.

2. $Var(\varepsilon_i) = \sigma^2 \qquad i = 1, \dots, n$

   Aquesta propietat es la d'homoscedasticitat. El paràmetre desconegut definit com σ^2 es el corresponent a la variància del errors del model definit abans. Podem trobar punts influents o atípics que son aquells que es troben separats del grup i poden fer que no es compleixin aquestes condicions.

3. $E(\varepsilon_i \cdot \varepsilon_i) = 0 \qquad \forall i \neq j$

   Aquesta condició determina l'estat de incorrelació entre les n desviacions, aquesta ens indica que no hi ha dependència lineal.

En síntesi, es pot dir que el model és lineal normal si cada ε_i es $N(0, \sigma)$ i $\varepsilon_1, \dots, \varepsilon_n$ són estocàsticament independents. Així es té $Y \sim N_n(X\beta, \sigma^2 I_n)$ on Y segueix una distribució normal multivariant amb vector de mitges $X\beta$ i matriu de covariàncies $\sigma^2 I_n$.
S'anomena rang del disseny al rang de la matriu X

$$r = rango\ X.$$

A partir del valor de r es sap que aquest es el nombre de paràmetres del disseny, en el sentit de que si r<n es possible reparametritzar el model perquè r sigui igual al nombre de paràmetres. En molts dissenys es veu com verifiquen directament que r=n i diem que es rang màxim. Si el model lineal verifica les condicions esmentades anteriorment es pot dir que esta sota les condicions de Gauss-Markov.

3.1.4. Mesures de qualitat del model de regressió lineal: La R^2

Per tenir una mesura sobre l'ajust de la regressió del model es pot pensar en la suma de quadrats $\sum_{i=1}^{n} e_i^2$ però aquesta es fixa en les unitats de y_i al cuadrat. Per mirar si el model esta ben ajustat s'utilitza el coeficient de determinació però aquest s'expressa diferent segons les condicions del model. El coeficient de determinació (R^2) es una mesura estadística que mesura la força de la relació lineal entre dos variables i es molt utilitzat pels investigadors quan es realitzen anàlisis de tendència.

En un primer cas si el nostre model tenim que $\beta_0 \neq 0$ la mesura es:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Aquesta mesura es troba entre els valors $0 \leq R_2 \leq 1$. Aquesta correlació coneguda como bondat de l'ajust es representada entre els valors 0 i 1. El valor 1 indica un ajust perfecte i un model molt fible per les prediccions futures, en canvi un valor de 0 ens indicaria que no arriba a modelar les dades amb precisió.
En un segon cas en el nostre model tenim que $\beta_0 = 0$, el coeficient de determinació s'expressaria:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} {y_i}^2}$$

Amb aquestes dos expressions sobre el coeficient de determinació es treu com a conclusió que no es pot comparar models que no tinguin terme independent amb els que si el tenen.
Des de el punt de vista geomètric, es veu com la bondat ens mesura la distancia entre la línia ajustada i tos els punts de dades que estan dispersos pel diagrama. El conjunt de dades ajustades tindrà una línia de regressió propera als punts, això ens indica que la distancia es petita. Aquesta dada deixa a l'usuari avaluar per si mateix el significat de la correlació i com pot ser aplicat en un futur per altres anàlisis de tendència. Quan es té més d'una variable independent en la regressió lineal múltiple, la interpretació és la mateixa, però enlloc de una recta de regressió tindrem un hiperplà.

3.2. Model de Regressió Logística

Aquest model de regressió logística explica el comportament d' Y en funció d'una o més variables independents qualitatives o quantitatives. En el nostre cas se sap que Y es una variable dependent discreta i dicotòmica, els valors que pot tenir la nostra Y son 0 i 1.
L'objectiu és fer un model que descrigui els efectes dels canvis en les variables explicatives sobre la probabilitat que Y valgui 1, o equivalentment, la probabilitat d'èxit o esdeveniment d'interès representat pel valor 1. La probabilitat d'èxit del nostre esdeveniment està determinada per P(Y=1)=p. En canvi, la probabilitat de no èxit o fracàs de l'esdeveniment es determina com P(Y=0)=1-p.

El model de regressió logística simple amb només una variable independent x és
on els lògits representen funcions lineals de les diferents variables explicatives.
Es representen de la següent forma:

$$\log it = \ln \frac{p}{1-p} = \beta_0 + \beta_1 x$$

Es parla d'*Odds* quan denominem la raó d'una probabilitat al ser valor complementari, amb l'expressió següent:

$$Odds = \frac{p(Y=1)}{p(Y=0)} = \frac{p(Y=1)}{1-p(Y=1)}$$

Si volem formular un model a partir dels termes *Odds* ho hauríem de treure de la següent forma:

$$\frac{p}{1-p} = e^{\beta_0+\beta_1 x}$$

Per poder tenir una idea de quan es més o menys probable un succés al passar de x a x+1, és a dir, al augmentar d'unitat una variable independent s'utilitza la raó d'*Odds* expressada de la següent forma:

$$OR\left(\frac{x+1}{x}\right) = \frac{Odd(x+1)}{Odd(x)} = e^{\beta_1}$$

L'exponencial del coeficient que acompanya a la variable x, es pot interpretar com l'augment relatiu de *l'odds* en incrementar una unitat la variable x.
Com s'ha comentat abans amb la raó *d'Odds* es pot comparar dos quocients *d'Odds* de la variable resposta durat dos situacions caracteritzades per valors adaptats per la variable independent x.
Si ens volem fixar en el model de regressió logística múltiple amb k variables explicatives ho expressaríem així:

$$log\ it = \beta_0 + \sum_{j=1}^{k} \beta_j x_j$$

A partir dels contrastos d'hipòtesis sobre els coeficients del model es pot decidir quins factors o variables independents son importants per descriure la probabilitat d'èxit del succés d'interès.

3.2.1 Ajust del model de regressió logística

Una vegada obtinguda la relació lineal entre logaritme dels *ODDs* i la variable predictora x, s'han d'estimar els valors dels paràmetres $\beta_0\ i\ \beta_1$. La estimació d'aquests paràmetres no es pot fer pel mètode ordinari de mínims quadrats per dos motius:

1. l'equació del model no es una funció lineal en els paràmetres, per això obtenim els estimadors β_k que minimitzen la suma de quadrats de les diferencies entre els valors observats Y_j i els seus valors mitjos $H(\beta_1, \dots, \beta_k)$ no porten a un sistema d'equacions lineals fàcils de resoldre.
2. la distribució de les Y_j no es normal, sinó de Bernouilli. Això ens diu que si avaluem la minimització de la suma de quadrats no ens porta a estimadors estadísticament òptims.

L'estimació dels paràmetres s'obtindrà per màxima versemblança. La funció de versemblança del vector $(x_1, \dots, x_n)$ és:

$$L(\theta) = f_{x_1,\ldots,x_n}(x_1, \ldots, x_n; \theta).$$

El logaritme de la versemblança conjunta dels N valors observats Y_j ve donar per l'equació següent:

$$\ln V = \ln\left(\prod_{j=1}^{j=N}(p_j^{Y_j}(1-p_j)^{(1-Y_j)}\right) = \sum_{j=1}^{j=n}[Y_j \ln p_j + (1-Y_j)\ln(1-p_j)]$$

Substituïm las p_j per l'expressió:

$$\left(\frac{p}{X_1 = x_{1j}, \ldots, X_l = x_{lj}}\right) = \frac{e^{\sum_{k=1}^{k=p}\beta_k Z_{kj}}}{1 + e^{\sum_{k=1}^{k=p}\beta_k z_{kj}}}$$

i les Y_j i x_{ij} pels valors observats en cada individu, acaba sent una funció $\ln V = H(\beta_1, \ldots, \beta_k)$ dels k paràmetres desconeguts del model. Els estimadors màxim versemblants dels paràmetres s'obtenen calculant amb tècniques d'optimització els valors que maximitzen l'expressió anterior. A diferència del model lineal, per a obtenir els valors màxim versemblants és necessari un mètode iteratiu.

S'obté una estimació a través de la regressió logística mitjançant el mètode de mínims quadrats iteratius (IRLS). En aquest es calculen les diferències entre les observacions reals i les estimacions d'una funció lineal, s'itera el procés fins que els resultats de les observacions i els resultats de les estimacions siguin els més simples possibles.

Aquest mètode es considera eficient per obtenir estimacions a partir de regressió logística. Encara que hi ha altres vies d'estimació aquesta continua essent la més àmpliament utilitzada.

3.2.2. Mesures de qualitat del model de regressió logística: Corba ROC i àrea sota la corba

Per poder descriure i comparar la precisió de les prediccions s'ha acceptat àmpliament l'anàlisi de la corba ROC. La corba ROC representa per cada llindar la tassa positiva verdadera contra la tassa de falsos negatius.

La tassa positiva verdadera és coneguda també com sensibilitat [10]. Es defineix com la probabilitat de que els resultats d'una proba siguin positiu si realment es dona l'esdeveniment, és a dir, una persona amb una malaltia sigui positiu si realment té aquesta malaltia. En aquest cas es veu que el nombre de casos de falsos negatius (persones amb la malaltia que donen un resultat negatiu a la proba) anirà disminuint a mesura que augmenti el valor de la sensibilitat de la proba.

En cas contrari trobem la especificitat, aquesta ens detecta el percentatge de persones sense la malaltia que van donar un resultat negatiu en una proba. Si la especificitat augmenta ens esta indicant que disminuirà el nombre de casos de persones sanes que donen un resultat negatiu a la prova.

	Malaltia present	Malaltia absent	
Proba positiva	a	b	a+b
Proba negativa	c	d	c+d
	a+c	b+d	

Taula 3.1: Taula sensibilitat i especificitat

Fórmules per al càlcul de paràmetres d'una prova diagnòstica
Sensibilitat= a/(a+c) ó Sensibilitat= VP/(VP+FN)
Especificitat= d/(b+d) ó Especificitat= VN/(FP+VN)
Valor predictiu positiu= a/(a+b) ó Valor predictiu específic=VP/(VP+FP)
Valor predictiu negatiu= d/(c+d) ó Valor predictiu específic=FN/(FN+VN)
% de falsos negatius= 100-sensibilidad
% de falsos positius= 100-especificitat
Exactitud=(a+d)/n ó Exactitud=(VP+VN)/n

Taula 3.2: Taula fórmules per al càlculs de paràmetres d'una prova diagnòstica

Si la corba ROC puja ràpidament cap a la cantonada del gràfic el valor de l'àrea sota la corba (AUC) es gran i per tant, ens indica que el nostre model funciona bé. Si pel contrari es un valor petit ens indica que alguna cosa no funciona.

$-AUC \approx 1 \rightarrow el\ model\ es\ excelent$

$-AUC \approx 0.5 \rightarrow el\ model\ no\ val$

Si es mira el gràfic de la corba ROC del model, es pot veure que el punt (0,1) correspon al classificador perfecte, això vol dir que es classifica de manera correcta tot positiu i negatiu, la taxa de veritables positius és 1 i la de falsos positius és 0. El corresponent punt (1,1) representa un classificador que prediu tots els casos positius i en el contrari el punt (0,0) prediu tots els casos negatius.

L'AUC parla sobre dos individus un amb resposta negativa i un altre amb una positiva. Aquest s'interpreta com la probabilitat que l'individu amb resposta positiva tingui una probabilitat prevista més alta.

Parlant sobre l'AUC d'una corba ROC podem interpretar-ho segons la guia següent:

0.9-1 =excel·lent

0.8-0.9= molt bo

0.7-0.8= bo

0.6-0.7= dolent

0.5-0.6= molt dolent

Per ajustar diferents models de regressió logístics es pot utilitzar la funció glm() en R, de les sigles "*Generalized Linear Model"* [8] *,* a banda de la regressió logística pot ajustar models de *Poisson* i d'altres de la família exponencial. La sintaxis bàsica per poder obtenir un model lineal es la següent:

```
glm(dependiente~independiente1+independiente2, family=binomial(),
                              data=datos")
```

Listing 3.1: Sintaxi aplicada en R per crear un model logístic

Existeix un paquet en R anomenat pROC que permet l'emmagatzematge i l'anàlisi de models de regressió logística. Es pot examinar la informació utilitzada en models de regressió logística i seguir els resultats a través de visualitzacions gràfiques. A més, té funcions per estimar la sensibilitat i l'especificitat mitjançant paràmetres indicadors com l'àrea sota la corba, que es poden utilitzar per mesurar la precisió dels models, més concretament, la capacitat de discriminació.

3.4. Recursos informàtics

Per tancar aquest capítol cal anomenar els programes informàtics que s'han utilitzat per dur a terme aquest projecte.
Primerament parlar del llenguatge informàtic utilitzat en aquest estudi ha estat RStudio. Aquest disposa d'una gran varietat de funcions i paquets amb els quals es pot dur terme un bon anàlisi de les dades.
Després comentar que per la creació de l'aplicació web s'ha utilitzat la llibreria shiny [1] [2] [3]. Aquesta llibreria permet crear la interfície o s'ha desenvolupat l'anàlisi dels models lineals i logístics de la base de dades COVID determinada. Per aprendre a utilitzar-la i descobrir les diferents funcionalitat que proporciona es va consultar el material del curs de la *Summer School* [4] del professor Isaac Subirana. Per fer els models s'ha utilitzat les funcions lm i glm de la llibreria *stats*. Per la corba ROC hem utilitzat la llibreria pROC, la funció de la qual s'ha explicat anteriorment.
Per últim, es vol recordar que la sintaxis d'R per a l'aplicació es pot trobar a l'apèndix d'aquest treball.

Capítol 4

4. Funcionament de la interfície interactiva web

En aquesta secció es mostrarà les diferents parts i opcions que composen la interfície interactiva web dissenyada per il·lustrar el funcionament del model lineal i el model de regressió logística en la docència.
Aquesta aplicació web ha estat desenvolupada utilitzant llenguatge R, més concretament, s'ha utilitzat la plataforma RStudio juntament amb les extensions Shiny per crear una pàgina web interactiva per aconseguir un disseny intuïtiu per fer més fàcil la visualització dels gràfics, taules i les diverses funcions disponibles.

4.1. Guia d'ús

Al entrar a aquesta interfície web, es podrà observar la primera pantalla tal i com es pot veure a la il·lustració 1. En aquesta primera plana es veu com a la dreta del títol trobem les diferents pestanyes de les quals esta formada aquesta aplicació per el anàlisi de les dades. Per defecte ens apareix la plana "Introducció", on es troba una petita explicació sobre el projecte fet per aquesta web.

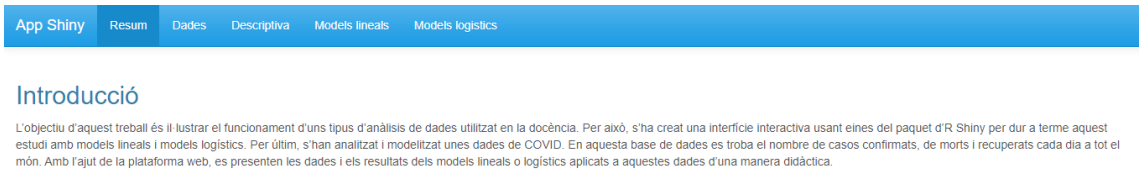

Figura 4.1: Primera pàgina de l'aplicació web

Si es vol treballar amb les dades ja carregades a la web no es necessari canviar cap configuració ja que s'ha treballat amb una base que parla sobre el COVID.
D'altra banda, si es vol utilitzar altres dades en aquesta interfície caldrà carregar-les en el codi d'R com un nou arxiu .csv amb el nom de data, per així no haver de fer cap altre canvi en el codi. De totes maneres, això no ho podrà fer l'usuari, sinó que cal fer-ho des del servidor. Això sí, no caldria canviar el codi empleat per a confeccionar l'aplicatiu.
En la següent plana, anomenada "Dades" es troba un resum sobre les bases de dades carregades. A la dreta hi ha tres possibles opcions, les quals donen informació sobre la base. Es pot observar un petit menú a l'esquerra d'aquesta primera plana que esta relacionat amb la primera finestra d'aquesta plana. Aquest menú deixa escollir el nombre d'observacions que es vol mirar fins a un màxim de tota la base de dades. Per defecte el valor seleccionat es de 10 dades.

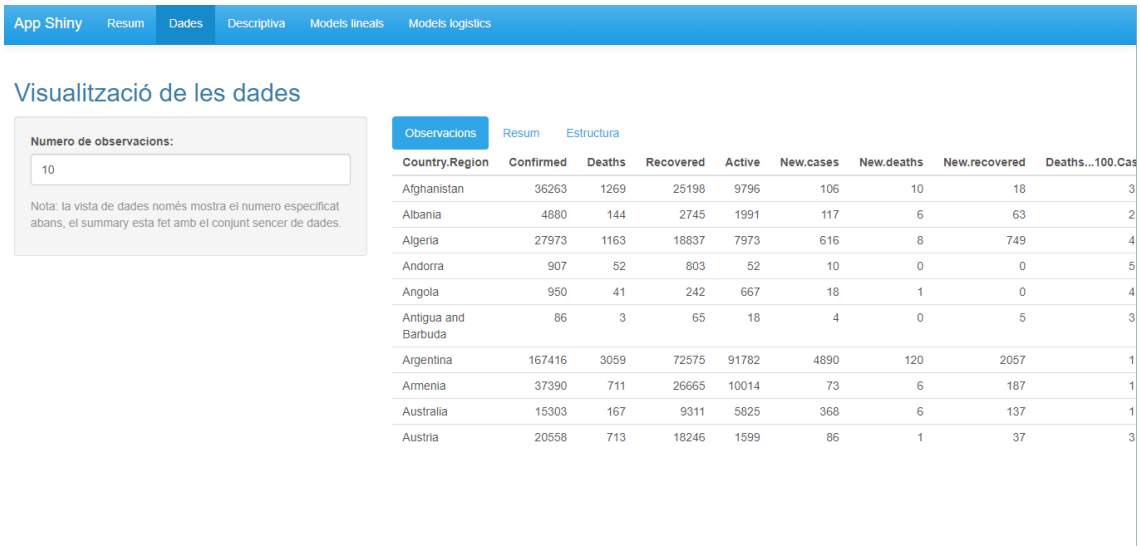

Figura 4.2: Visualització de les dades en l'aplicació web

En la segona finestra hi ha un resum de les dades. Per a cada una de les variables es pot observar un petit resum amb alguns dels estadístics principals com són el min., màx, mitjana,...

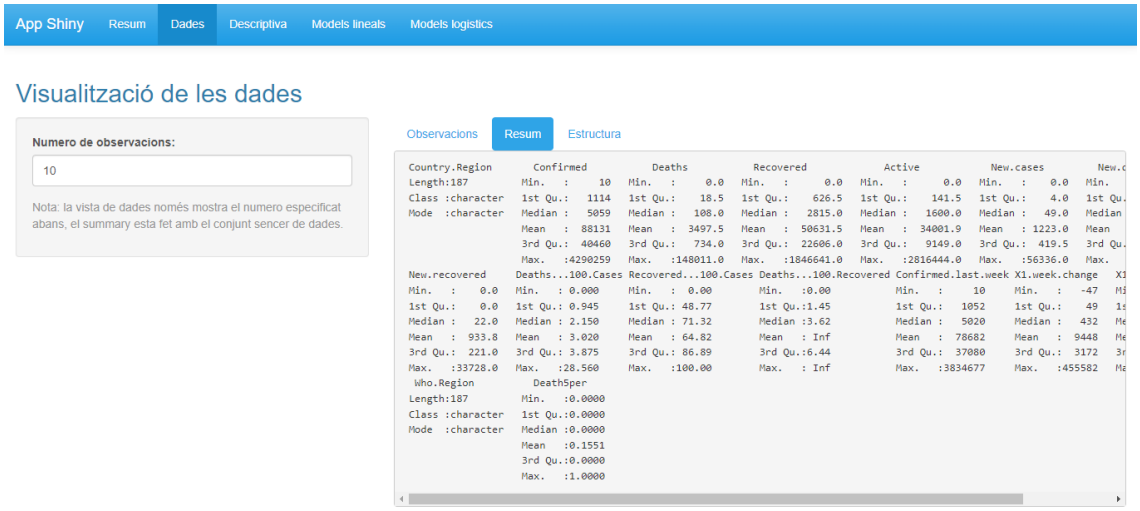

Figura 4.3: Resum de les dades

En l'ultima finestra trobem un llistat de totes les variables on dona informació sobre la seva estructura i el tipus de variable que és.

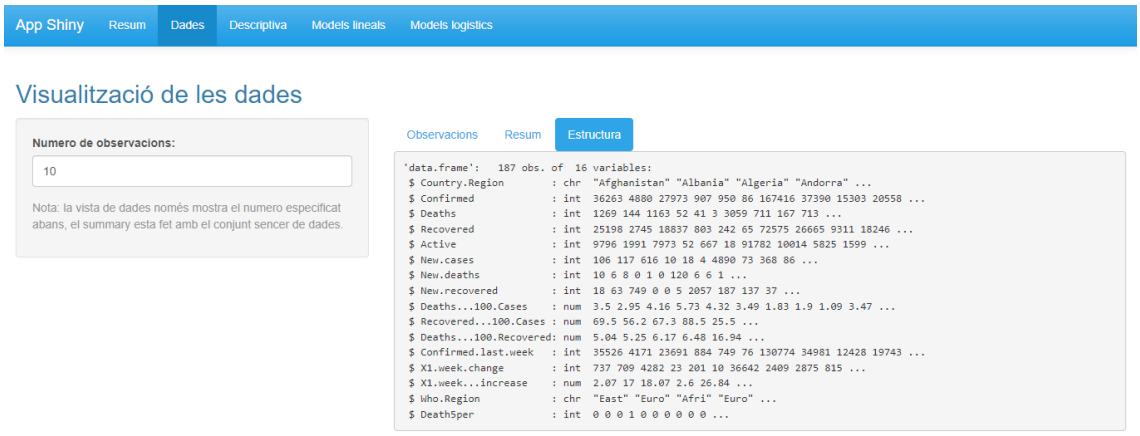

Figura 4.4: Estructura de les dades

4.2. Descriptiva

La següent plana de la web es l'anomenada "Descriptiva". Es pot veure un menú a la part esquerra que permet escollir la variable que vols representar gràficament. En aquest menú trobem una nota que ens indica la informació donada sobre els gràfics anteriorment.

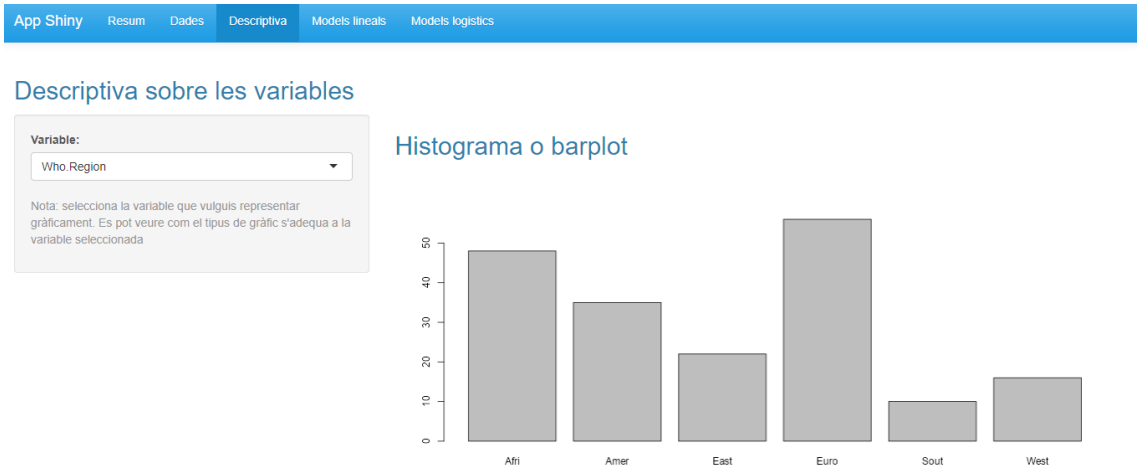

Figura 4.5: Descriptiva de les dades (Gràfic de barres de la variable *"Who.Region")*

Per a les variables categòriques s'ha optat per fer un gràfic de barres i per les numèriques un histograma, aquesta opció no es torba com a possible elecció ja que s'ha programat així per una major eficiència. En la pagina actual es troba una descriptiva més visual de les dades que junt amb els estadístics i l'observació de las dades a la primera plana ens aporta un coneixement total de la base de dades amb la que es treballa durant el projecte.

4.3 Model lineal

En aquesta pàgina es presenten les implantacions de la web per poder iniciar el procés de modelització lineal. Amb la informació obtinguda en les pagines anteriors (Dades i descriptiva) es pot tenir una idea de la regressió lineal que es vol plantejar en aquesta secció. En el panell que es troba a l'esquerra com s'indica s'ha de seleccionar una variable dependent i tantes variables independents com es vulgui segons el model lineal que es vulgui plantejar.

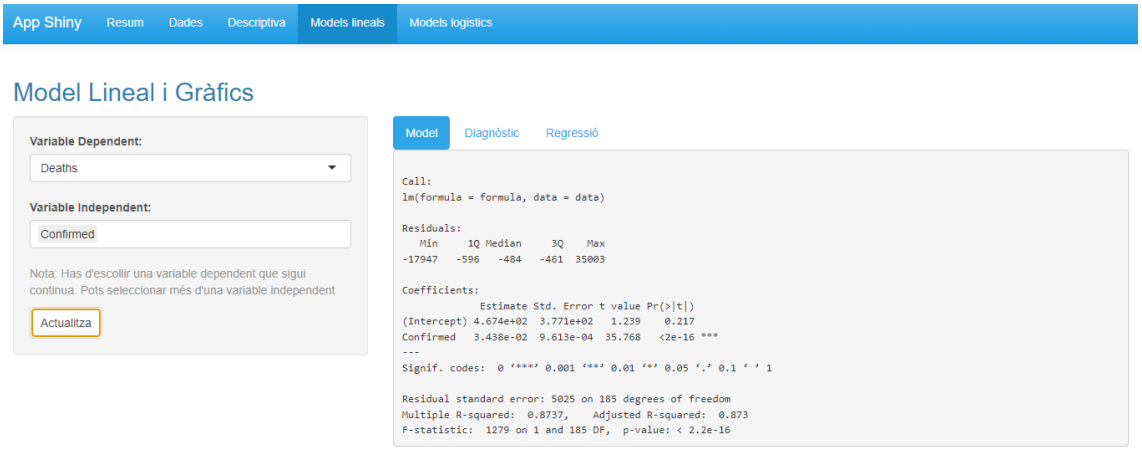

Figura 4.6: Model lineal

En la primera pestanya trobem el model lineal amb les variables escollides en un primer moment. En aquesta dona el valor dels coeficients, del p-valors de cada variable per saber si aporten informació o no al model escollit, el R^2 que diu què tan ajustat està el model i que quant més a prop de l'1 millor ajusta, i altres estadístics per estudiar el model. En la següent finestra es troben els gràfics de diagnòstics que ens indiquen si es compleixen les premisses del model (normalitat, homocedesticitat) així com ens identifiquen possibles valors *"outliers"* i/o influents.

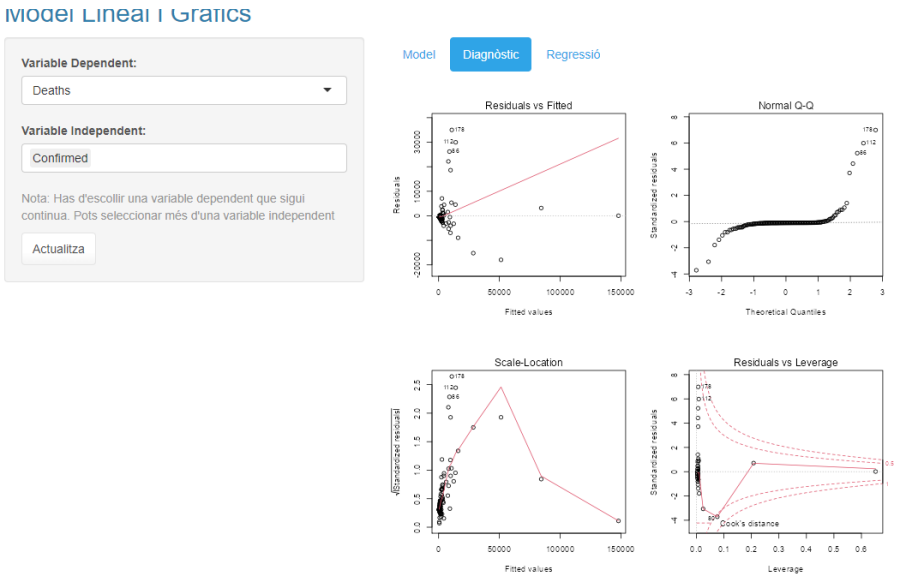

Figura 4.7: Gràfics de diagnòstic model lineal

Per últim en aquesta plana trobem una finestra que fa un gràfic de regressió lineal entre la variable dependent i una de les variables independent escollides. Per poder obtenir aquesta representació no es poden escollir més d'una variable independent, o sigui, que no apareixerà el gràfic si no es tracta d'una regressió lineal simple.

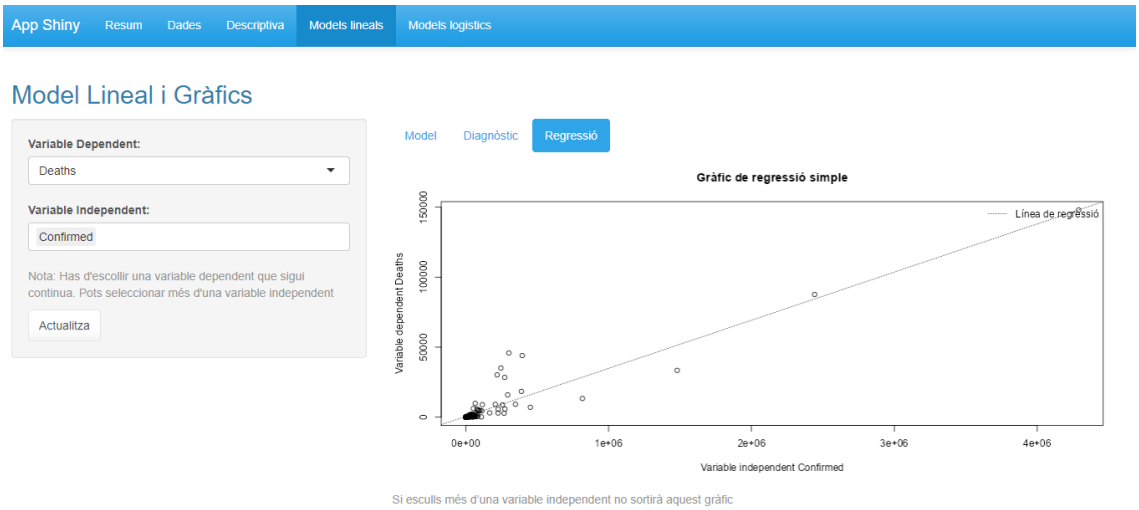

Figura 4.8: Regressió lineal entre variable dependent i una de les variables independents

S'observa una nota sota de la regressió que ens indica que només es pot escollir una variable, per donar una solució en cas de que surti l'error.

4.4. Model logístic

En l'ultima plana es presenta la implantació en la web del model logístic per les dades escollides. Mirant la informació que tenim de les dades i sabent que la variable dependent ha de ser binària per la seva modelització s'escull la més adequada. En el planell de l'esquerra s'indica que s'ha de seleccionar una variable dependent binària i tantes independents com es vulgui per fer el model desitjat.
Si l'elecció de la variable dependent no es adequada el model dona un missatge de error (4.8) que indica que l'has de canviar la variable dependent i s'ha d'escollir una variable binària.

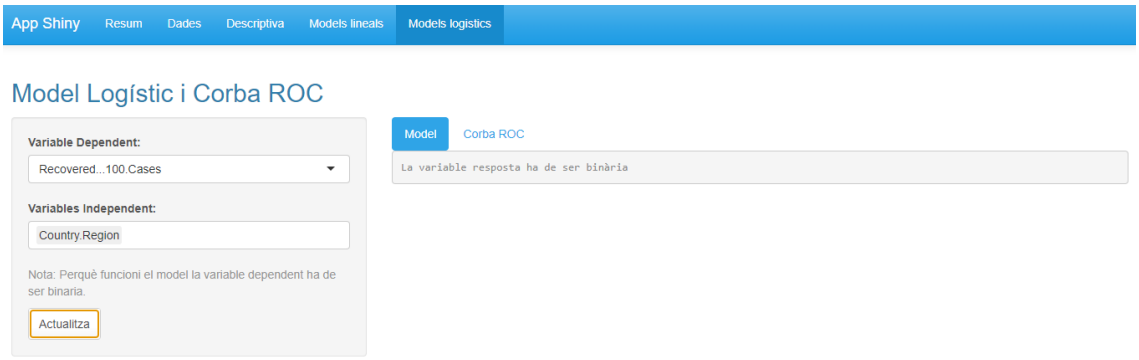

Figura 4.9: Regressió logística, missatge d'error

Figura 4.10: Regressió logística

En aquesta pestanya trobem la modelització logística de les variables escollides pe l'usuari. S'observa el valors del coeficients, els p-valors , l'AIC, entre altres dades. Aquestes tres son les necessàries per saber quin valors tenen cada una de les variables en el model en el cas dels coeficients, per saber si les variables donen informació al model en el cas dels p-valors. L'AIC serveix per a comparar models amb diferents variables independents. Per exemple, si en un model una variable no es necessària i el tornem a fer sense aquesta veurem que el valor del AIC es major, esta millor ajustat.

A partir de les dades que proporciona aquest apartat podem calcular *l'ODDS ratio* junt amb el valor de la nostra variable dependent. A la metodologia tenim les formules a obtenir aquests valors.

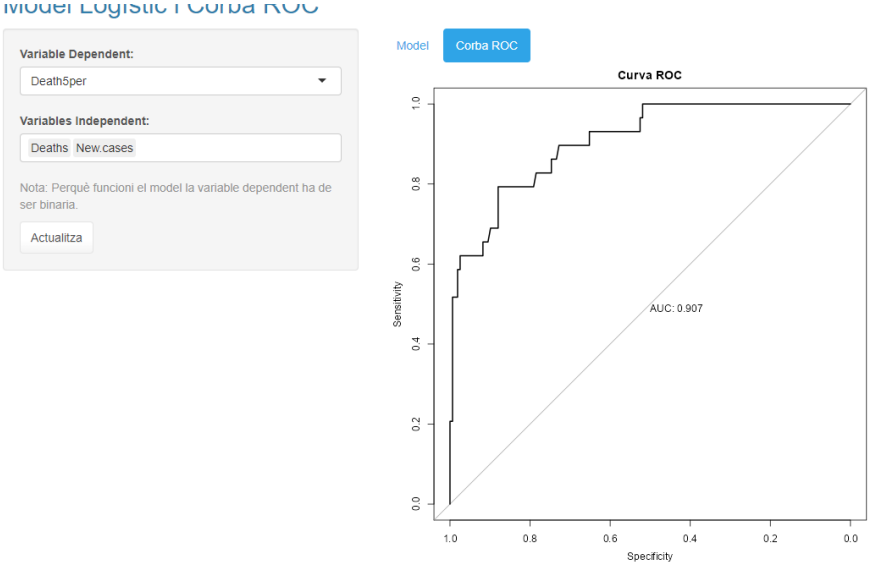

Figura 4.11: Corba ROC

Per últim, es veu la corba ROC on estan representades la sensibilitat i l'especificitat del model creat junt amb l'AUC.

Capítol 5

5. Anàlisi de la base COVID

En aquest capítol s'analitzarà pas a pas, segons els models explicats i implementats dins de la interfície web creada pel projecte, la base de dades COVID. L'objectiu es fer un bon anàlisi per la seva ensenyança en un futur pròxim. Retornant a la petita descriptiva duta a terme en el capítol de la descripció de les dades, hi havia dos variables categòriques i catorze numèriques, de les quals només una és binària. Aquesta ha estat recodificada a partir d'una variable existent a la base de dades tal i com s'ha descrit a la secció de mètodes.

5.1 Model lineal

L'aplicació ha estat creada en aquest projecte per modelitzar les dades de dos formes diferents i que pugui ser usada en la docència per a explicar aquests models. Es comença amb la modelització lineal. En aquest cas s'ha decidit utilitzar com a variable dependent *"Deaths"* i com a variables independents *"Confirmed"* i *"New Cases*". La variable dependent parla del nombre de morts que hi ha. Les variables independents parlen sobre el nombre de casos confirmats i els nous casos respectivament. Per tant, en aquest model s'intentarà predir en nombre de morts a partir del nombre de asos confirmats i el nombre de casos emergents o nous.

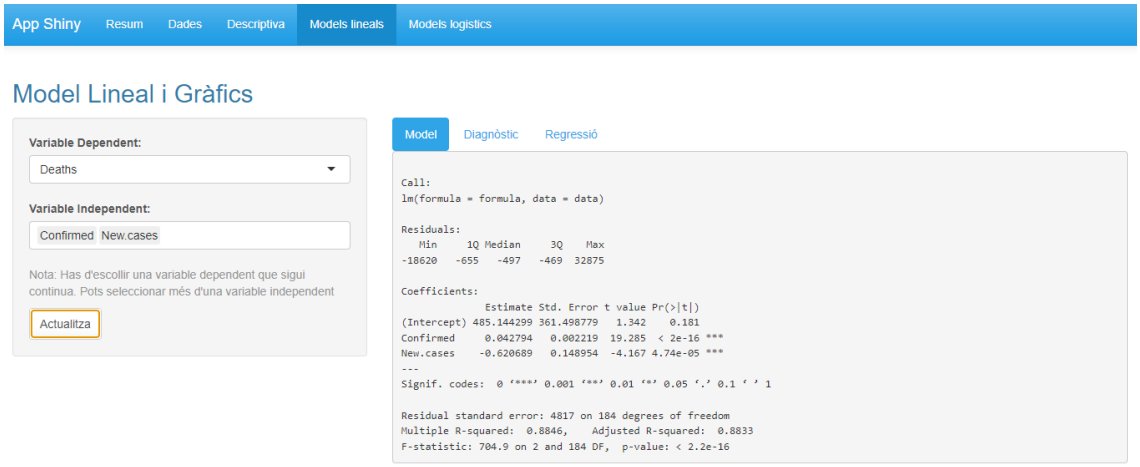

Figura 5.1: Model lineal *Deaths ~ Confirmed + New Cases*

En la modelització (Figura 5.1) es pot veure com les variables del model son significatives, amb això esta dient que són necessàries. El valor "*intercept*" és bastant elevat parlant de morts, ja que si no hi ha cap cas confirmat o el nombre de nous casos és zero el nombre de morts serà de 485. Encara tenir aquest resultat no es pot rebutjar la hipòtesi de que aquest tingui un valor de 0 ja que el seu p-valor no es significatiu i aquesta opció tindria més lògica que l'obtinguda ara. L'estimació dels paràmetres ens indica que per cada cas de COVID confirmat el valor de

les morts augmenta en 0.042794, mentre que en el cas de confirmar un nou cas de COVID el nombre de morts disminueix en 0.620689. Això que el nombre de morts baixi en augmentar el nombre de casos nous és una mica contradictori. Encara que es podria explicar per l'efecte retard en les morts, que de mitjana entre la infecció o inici dels símptomes i la mort poden passar dos o tres setmanes.

La equació resultat del model és la següent:

$$Nº\ de\ morts\ estimat = 485.144299 + 0.042794 * \text{Confirmed} - 0.620689 * \text{New.cases}$$

Per tant si en un país aleatori el nombre de casos confirmat es de 100 i hi ha 50 casos nous el nombre de morts serà de 458.

Observem com el valor de R^2 és de 0.8846 i com que és proper a 1 direm que el model està ben ajustat. El model lineal creat per les variables *"Deaths", "Confirmed" i "New-cases"* es considera significatiu i per tant, direm que són variables necessàries per explicar el nombre de morts.

La següent pestanya ens mostra els gràfics de diagnòstic que ens ajudaran a verificar les premisses del model: homocedesticitat i normalitat dels residus, presència *d'outliers* i valors influents.

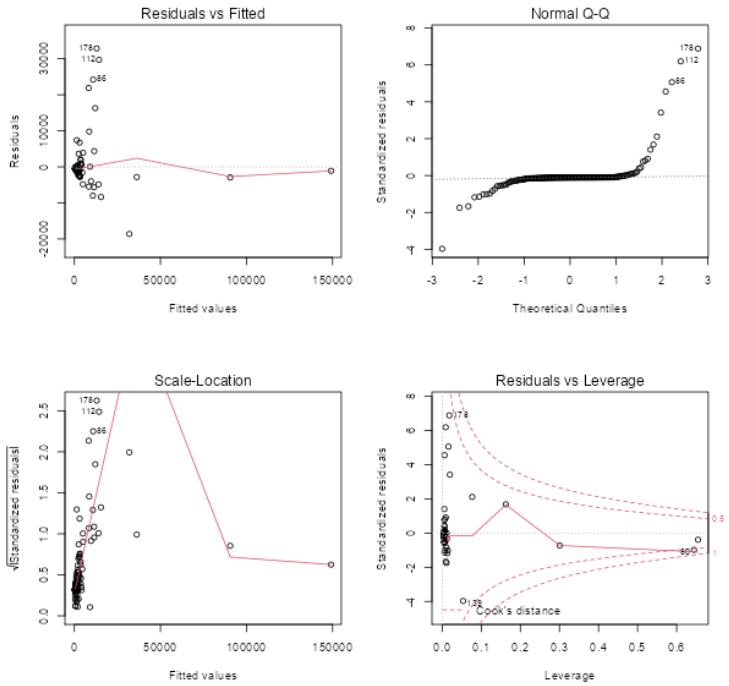

Figura 5.2: Gràfics de diagnòstic del model *Deaths ~ Confirmed + New Cases*

S'analitzarà un a un els gràfics de diagnòstic (Figura 5.2) del nostre model. El primer fa una representació gràfica dels valors dels residus. La informació que es pot treure sobre aquest és que la mitja no és 0 encara que s'apropa però al no ser constant no es pot verificar aquesta opció. Un altra dada important que s'obté d'aquest és que la seva variància no és constant ja que tenim un munt de punts en la mateixa zona i altres molt separats. El gràfic de normalitat

s'observa que hi ha un munt de dades en la mitja però les cues són grans i sobretot a la dreta trobem bastants valors *outliers*. Per tant es pot considerar que no segueixen una distribució normal. Amb els altres dos gràfics es pot intuir la presencia *d'outliers* ja que hi ha molts punts molt lluny de la resta de dades. Aquests *outliers* s'han vist durant l'anàlisi de les dades on hi havia un petit percentatge dels països que tenien un nombre de morts molt elevat.

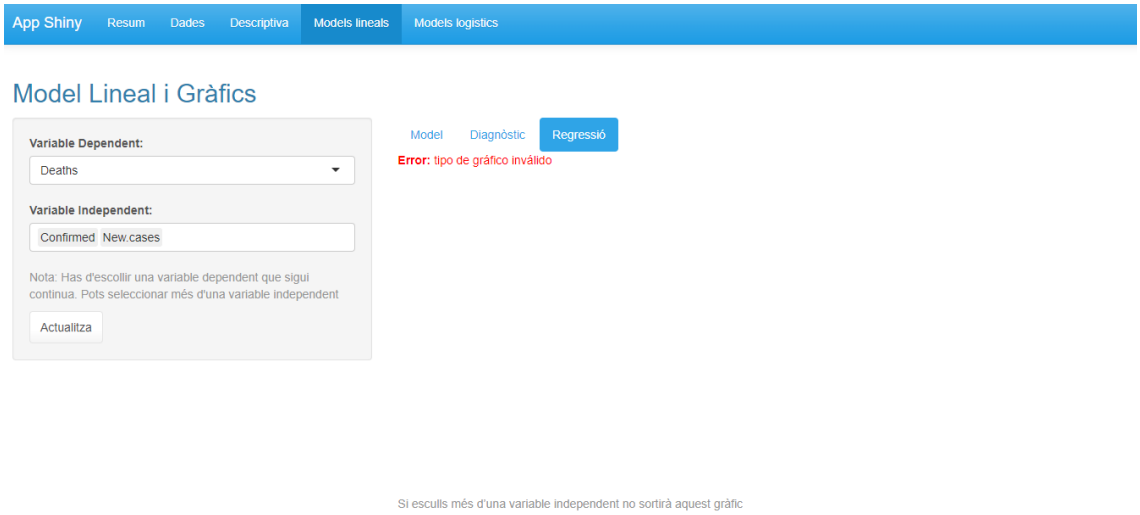

Figura 5.3: Gràfic regressió lineal model *Deaths ~ Confirmed + New Cases*

Si se selecciona més d'una variable independent apareix un missatge d'error enlloc del gràfic. I és que l'aplicatiu està pensat que només aparegui el gràfic per a una regressió lineal simple, o sigui, quan escollim una sola variable independent. A sota del error es pot veure un missatge que diu: "Si esculls més d'una variable independent no sortirà aquest gràfic".

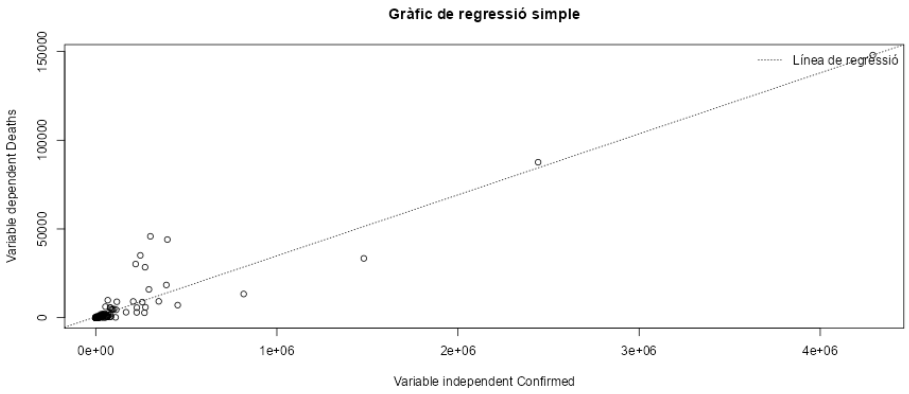

Figura 5.4: Gràfic regressió lineal model *Deaths ~ Confirmed*

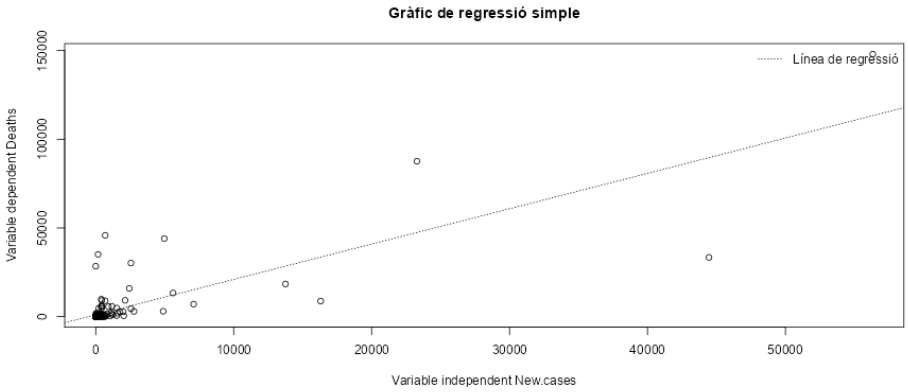

Figura 5.5: Gràfic regressió lineal model *Deaths ~ New Cases*

Tant en la Figura 5.4 com la Figura 5.5 podem veure com les dades es troben acumulades a prop del valor 0. Mentre que hi ha casos on els punts es troben dispersos aquests són els determinats *outliers*. S'observa que per tots dos casos el nombre de morts es major quant més casos hi ha confirmats i quan més nous casos hi ha.

5.2 Model logístic

En aquest apartat s'analitzarà les dades pel model logístic [7]. En aquest cas l'única variable a utilitzar com variable dependent és la creada per nosaltres amb el nom de *"Death5per"* que val 1 si el nombre de casos per a cada 100,000 habitants és major de cinc i 0 en cas contrari. Aquesta és binària i com se sap la variable dependent o y ha de ser binaria per fer aquest anàlisi. Les variables escollides per la part independent seran *"Deaths" i "New Cases".*

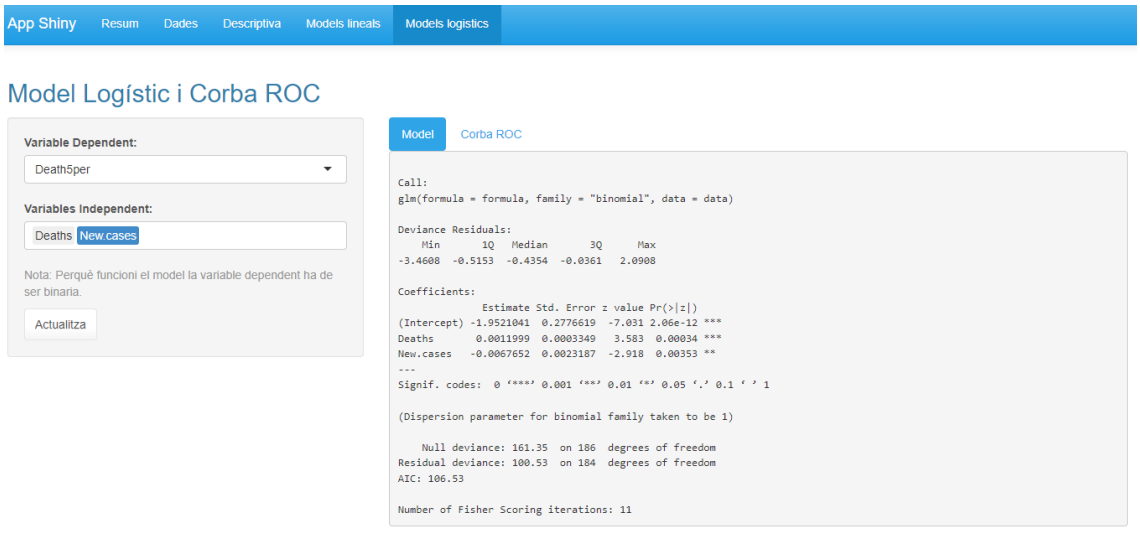

Figura 5.6: Model logístic *Death5per ~ Deaths + New Cases*

En primer lloc, es veu que les variables escollides pel model són significatives, per tant esta dient que son necessàries per treure la probabilitat de que el percentatge de morts per 100 casos sigui major a 5. El model obtingut a partir d'aquest anàlisi tindria l'expressió següent:

$$\text{logit}(\text{P}(> \%5) = -1.9521041 + 0.0011999 * Deaths - 0.0067652 * New.cases$$

Per tant la probabilitat d'un país amb 100 morts i 50 casos nous és del 14,49%.Pel contrari, se sap que la probabilitat de que el percentatge de morts sigui menor a 5% es de 85,51%. Per arribar a aquests resultats s'ha utilitzat la següent formula:

$$P(>\%5) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 * Deaths + \beta_2 * New\ cases)}}$$

Un altre valor necessari per saber si l'ajust del model és adequat es l'AIC, aquest s'utilitza per a comparar models amb diferents variables independents. El millor model i per tant el que explica millor les dades i s'ajusta millor a elles és aquell que tingui un menor AIC. Un terme del que s'ha parlat a la metodologia és *l'Odds Ratio*, per aquest model l'*OR* corresponent seria el observat en la Taula 5.1.

	OR	2.5 %	97.5 %
(Intercept)	0.1419750	0.07924036	0.2371009
data$Deaths	1.0012006	1.00066380	1.0020071
data$New.cases	0.9932576	0.98786479	0.9969815

Taula 5.1: *Odds ratio* model logístic *Death5per ~ Deaths + New Cases*

En l'última finestra anomenada "Corba ROC", treu el gràfic de la corba ROC junt amb el seu valor de AUC, el quan sabem que com més proper sigui a 1 millor és el model

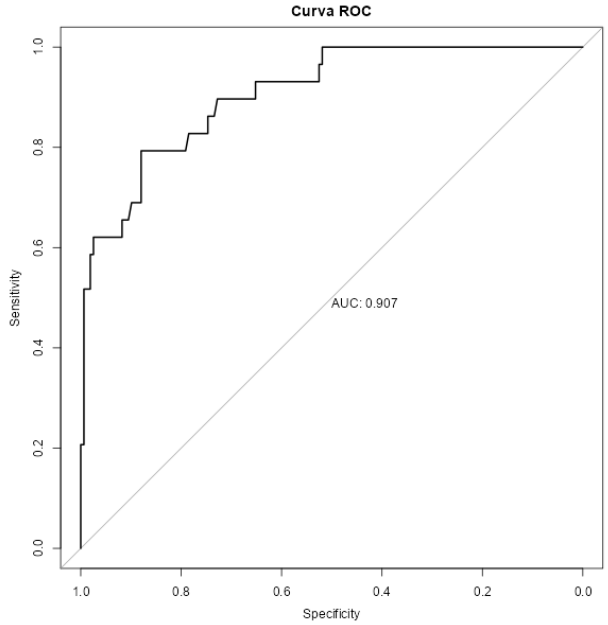

Figura 5.7: Corba ROC Model logístic *Death5per ~ Deaths + New Cases*

Si s'observa la forma de la corba veiem com puja ràpidament cap a la cantonada amb això indica que el valor de àrea sota la corba es gran i per tant el nostre model funciona bé. El model fet en aquest exemple té un AUC de 0.907 això indica que és un model excel·lent.

Capítol 6

Conclusions

Amb l'elaboració d'aquest projecte, en primer lloc poden afirmar que s'ha complert l'objectiu de dissenyar i crear una plataforma interactiva per fer l'estudi d'una base de dades per models lineals i models logístics. I per altra banda, és una web que estarà disponible online pel seu ús a la docència, per explicar els dos models. També s'afegirà el codi de Shiny en l'apèndix per si qualsevol ho vol utilitzar o millorar.
Amb aquest dos punts, es pot afirmar que s'han complert els objectius definits a l'inici del projecte, dissenyant una interfície web que facilita la modelització i explicació del model lineal i logístic per poder ser usat a la docència d'una manera fàcil i visual.
Dins de la aplicació s'incorporen els principals gràfics, taules i funcions necessàries per l'anàlisi dels models. A més a cada una de les planes, podem veure notes i indicacions que confeccionen l'aplicatiu.
L'estudi fet sobre la base de dades es podria estendre mirant totes les variables, quines donen més informació i ens serveix o pel contrari no son significatives i no són necessàries. Amb aquesta base de dades es podria analitzar centenars de models i fer un estudi sobre cadascun d'ells fixant-se en els estadístics principals dels que s'ha parlat durant l'anàlisi de la base de dades. En l'exemple mostrat en aquesta memòria s'ha vist que els models estaven ven ajustat i que les variables escollides eren necessàries per l'explicació del resultat o de la probabilitat segons el model. Aquesta condició de bon ajust s'ha pogut verificar més tard amb gràfics com es la corba ROC en el cas de la modelització logística. Pel cas del model lineal s'ha vist la R^2 i s'han utilitzat els gràfics de diagnòstic per a validar el model.
Finalment, les planificacions de futurs per aquest projecte és la creació d'actualitzacions on es calculin altres paràmetres com podria ser *l'odds ratio* de cada variable de la regressió logística, o el tractament *d'outliers* o transformació de variables (com per exemple el logaritme) per a normalitzar-les. Una altra opció interessant podria ser afegir que l'usuari carregués la seva pròpia base de dades en el moment de fer servir l'aplicatiu.

A través d'aquest treball no només s'ha investigat sobre la COVID i els seus afectes en els diferents països, sinó que també s'ha programat una web interactiva des de zero, cosa que no s'explica al grau a diferència dels models lineals i logístics. Ha estat un camí d'aprenentatge constant i amb moltes novetats però sempre amb la base teòrica obtinguda en el grau d'estadística.

Bibliografia

[1] Shiny, [s. d.] online[cons. 2023-06-23]. Disp. a: https://shiny.rstudio.com/.
[2] Shiny (s.f.) online [cons. 2023-06-23]. Disp. a: https://shiny.posit.co/
[3] Shiny (s.f.) online [cons. 2023-02-06]. Disp a: https://shiny.posit.co/r/gallery/widgets/widget-gallery/

[4] De Catalunya, U.U.P. (s.f.) Creacion de aplicación web con Shiny. *Master's degre in Statictics ans Operation Research.* Online [cons. 2023-02-25]. Disp a:https://mesioupcub.masters.upc.edu/en/xvi-summer-school-2023/courses/creacion-de-aplicaciones-web-con-shiny

[5] AMAT Joaquín, Regresión logística simple i múltiple. Online. [cons. 2023-04-16]. Disp. a: https://www.cienciadedatos.net/documentos/27_regresion_logistica_simple_y_multiple#Concepto_de_ODDS_o_raz%C3%B3n_de_probabilidad,_ratio_de_ODDS_y_logaritmo_de_ODDS

[6] LARRANAGA Pedro, INZA Iñaki, MOUJAHID Abdelmalik. Regresón Logística. Online. Ciencias de la computación e Inteligencia Artificial. Universidad del País Vasco-Euskal Herriko Unibertsitatea. [cons. 2023-04-16]. Disp. a: http://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/t7logistica

[7] BARÓN Francisco Javier. Regresión logística binaria. Online. [cons. 2023-04-17]. Disp. a: https://www.bioestadistica.uma.es/apuntesMaster/regresi%C3%B3n-log%C3%ADstica-binaria.html

[8] PALADINO, Martín. Modelos logit con R. Online. [cons. 2023-05-27]. Disp. a: https://www.institutomora.edu.mx/testU/SitePages/martinpaladino/modelos_logit_con_R.html#ajuste-de-modelos-logit-con-glm

[9] *Kaggle* (s.f.). Online. [cons. 2022-10-20]. Disp. a : https://www.kaggle.com/datasets/imdevskp/corona-virus-report

[10] VÍZCAINO-SALAZAR, Gilberto. Importancia del cálculo de la sensibilidad, la especificidad y otros parámetros estadísticos en el uso de las pruebas de diagnóstico clínico y de laboratorio. Online. 2017 [cons. 2023-05-07]; Volumen 23, número 7-8. Disponible en: https://docs.bvsalud.org/biblioref/2018/05/883697/importancia-calculo-sensibilidad-y-especifidad.pdf

Apèndix

Codi Shiny

```
library(shiny)

library(shinythemes)

library(magrittr)

library(dplyr)

library(pROC)


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

data<- data[,-1] #Eliminem la primera fila, ja que es una llista creada per excel i no ens dona cap informació


# Define UI for application that draws a histogram

ui <- navbarPage(title = "App Shiny",

                 theme = shinytheme("cerulean"),

                 tabPanel("Resum", titlePanel("Introducció"),

                          p("En aquest treball es busca il·lustrar el funcionament d'uns tipus d'anàlisis de dades utilitzat en la docència." ,

                            "Per això, s'ha creat una interfície interactiva amb l'ajut de Shiny per dur a terme aquest estudi amb models lineals i models logístics.",

                            "Per últim, s'ha modelitzat i creat l'anàlisi de les dades COVID carregades en aquesta web.",

                            "En la base es troba el nombre de casos confirmats, de morts i recuperats cada dia a tot el món. ",

                            "Amb l'ajut de la plataforma web, s'han representat les dades amb els models esmentats abans i és més il·lustratiu per a la docència d'aquests.")),

                 tabPanel("Dades", titlePanel("Visualització de les dades"),

                          sidebarLayout(

                            sidebarPanel(

                              numericInput("obs",

                                           "Numero de observacions:", 10),

                              helpText("Nota: la vista de dades només mostra el numero especificat abans,",
```

```
                                    "el summary esta fet amb el conjunt
sencer de dades.")
                            ),
                            mainPanel(
                              tabsetPanel(type="pills",
                                          tabPanel("Observacions",
tableOutput("view")),
                                          tabPanel("Resum",
verbatimTextOutput("result")),
                                          tabPanel("Estructura",
verbatimTextOutput("files"))
                              )
                            )
                          )
                 ),
                 tabPanel("Descriptiva",
                          titlePanel("Descriptiva sobre les variables"),
                          sidebarLayout(
                            sidebarPanel(
                              selectInput("variables", "Variable:", choice=
names(data)),
                              helpText("Nota: selecciona la variable que
vulguis representar gràficament.
                                       Es pot veure com el tipus de gràfic
s'adequa a la variable seleccionada")
                              ),
                            mainPanel(
                              titlePanel("Histograma o barplot"),
                              plotOutput("res"))
                          )),
                 tabPanel("Models lineals",
                          titlePanel("Model Lineal i Gràfics"),
                          sidebarLayout(
                            sidebarPanel(
                              selectInput("var_deplineal","Variable
Dependent:", choice=colnames(data)),
                              selectInput("var_indeplineal","Variable
Independent:", choice=colnames(data), multiple = TRUE),
```

```
                              helpText("Nota: Has d'escollir una variable
dependent que sigui continua.
                                       Pots seleccionar més d'una variable
independent"),
                              actionButton("fit_lineal", "Actualitza")
                            ),
                            mainPanel(
                              tabsetPanel(type="pills",
                                          tabPanel("Model",
verbatimTextOutput("mod")),
                                          tabPanel("Diagnòstic",
plotOutput("plot", width=600, height=600)),
                                          tabPanel("Regressió",
plotOutput("plot_simple"),
                                                   helpText("Si esculls més
d'una variable independent no sortirà aquest gràfic"))
                              )
                            )
                          )),
                 tabPanel("Models logistics",
                          titlePanel("Model Logístic i Corba ROC"),
                          sidebarLayout(
                            sidebarPanel(
                              selectInput("var_deplogi",          "Variable
Dependent:", choices =colnames(data)),
                              selectInput("var_indeplogi",       "Variables
Independent:", choices = colnames(data), multiple = TRUE),
                              helpText("Nota: Perquè funcioni el model la
variable dependent ha de ser binaria."),
                              actionButton("fit_logistic", "Actualitza")
                            ),
                            mainPanel(
                              tabsetPanel(type="pills",
                                          tabPanel("Model",
verbatimTextOutput("mod2")),
                                          tabPanel("Corba             ROC",
plotOutput("roc_plot", width=600, height=600))
                              )
                            )
```

```
                            )
                  )
)
##############################################################################
##############################################################################
##################################################################
server <- function(input, output) {

output$view <-renderTable({
    head(data,
         n=input$obs)
})

output$result <-renderPrint({
    summary(data)
})



  output$files <- renderPrint({
    str(data)
})



output$res <-renderPlot({
    variable<-data[[input$variables]]
    if(is.numeric(variable))
      hist(variable)
    else
      barplot(table(variable))
})



modelLineal<- reactive({
  if (input$fit_lineal==0) return(NULL)
  isolate({
```

```
    var_deplineal <- input$var_deplineal

    var_indeplineal <- input$var_indeplineal

    validate(need(is.numeric(data[[var_deplineal]]), "La variable resposta
ha de ser numèrica"))

    validate(need(length(var_indeplineal)>0,"Has de seleccionar com a mínim
una variable independent"))

    formula <- as.formula(paste(var_deplineal, "~", paste(var_indeplineal,
collapse = "+")))

    lm(formula, data = data)

  })

})




output$mod<-renderPrint({

  req(modelLineal())

  summary(modelLineal())

})



output$plot <- renderPlot({

  req(modelLineal())

  par(mfrow=c(2,2))

  plot(modelLineal())

})


output$plot_simple <- renderPlot({

  var_indeplineal<-input$var_inpeplineal

  validate(need(length(var_indeplineal)==1,"Només es pot seleccionar una
variable independent per aquest gràfic"))


  plot(data[, input$var_indeplineal],

       data[, input$var_deplineal],

       main = "Gràfic de regressió simple",

       xlab = paste("Variable independent", input$var_indeplineal),

       ylab = paste("Variable dependent", input$var_deplineal))

  abline(modelLineal(), lty=3)
```

```
  legend("topright","Línea de regressió", lty=3, bty="n")
})

modelLogistic<- reactive({
  if(input$fit_logistic==0) return(NULL)
    isolate({
      var_deplogi <- input$var_deplogi
      var_indeplogi <- input$var_indeplogi
      validate(need(length(unique(data[[var_deplogi]]))==2,   "La   variable
resposta ha de ser binària"))
      validate(need(length(var_indeplogi)>0, "Has de seleccionar com a mínim
una variable independent"))

      #Ajustar el model logístic
      formula  <-  as.formula(paste(var_deplogi,  "~",  paste(var_indeplogi,
collapse = "+")))
      glm(formula, data = data, family = "binomial")
    })
})

output$mod2<-renderPrint({
  req(modelLogistic())
  summary(modelLogistic())
})
output$roc_plot <- renderPlot({
    req(modelLogistic())
    modellogi<- modelLogistic()
    req(modellogi)
    prob <- predict(modellogi, type = "response")
    roc_obj <- roc(modellogi$y, prob)
    plot(roc_obj, main = "Curva ROC", print.auc = TRUE)
})
}

shinyApp(ui = ui, server = server)
```