Universitat Autònoma de Barcelona

Grau de Matemàtiques

Models Lineals

Predicció de la Supervivència al Titànic

Autors:
Xavier Civit Escoté
Xavier López Español

Supervisors:
Pedro Puig Casado
Manuel Higueras Hernáez

Aquesta fotografia es va fer des de el *Carpathia*, el vaixell que va rebre la senyal d'emergència del Titànic i va anar a rescatar els supervivents. `http://research.archives.gov/description/278338`

Índex

Introducció

L'accident del RMS Titànic és un dels naufragis més coneguts de la història, el 15 d'Abril de 1912 el RMS Titànic va xocar amb un iceberg i es va enfonsar en el seu viatge inaugural.
L'accident va causar la mort de més de 1.500 persones, és un dels desastres marítims més mortals en temps de pau a la història moderna, entre els passatgers hi havia algunes de les persones més riques del món, i a la vegada, centenars d'imigrants que buscaven una nova vida a Amèrica. [1]

Una de les causes que va portar a la mort de molts passatgers va ser que no hi havia suficients bots salvavides per a tots els passatgers. Encara que la sort juga un paper important sobre la supervivència dels passatgers, alguns grups de persones són més propensos a sobreviure, com les dones, els nens i la primera classe.

En aquest treball, proposem un model per a predir la supervivència basat sobre les dades d'alguns dels passatgers a bord del RMS Titànic.
Les dades de les que disposem i basem el model es poden descarregar aquí:
`https://www.kaggle.com/c/titanic-gettingStarted/download/train.csv` [2]

Usarem un **model lineal generalizat**, sabem que un model d'inferència estadística pot no resultar satisfactori per a modelar un fet que no es tornarà a repetir, no obstant, usant croos-validation hem comprovat que és efectiu.

[1] `http://en.wikipedia.org/wiki/RMS_Titanic`
[2] Kaggle és una plataforma per a concursos de data science, un dels concursos que tenen oberts per als principiants és realitzar un model que predigui la supervivència del RMS Titànic `https://www.kaggle.com/c/titanic-gettingStarted`

Informació de les dades

Al fitxer *train.csv* hi ha tota la informació que utilitzem per al model, en aquest fitxer hi ha 891 observacions de la supervivència de passatgers del Titànic, on les variables descriptives són les següents:

- survival: Supervivència, (0=No sobreviu, 1= Sobreviu).
- pclass: Classe del passatger, pot ser 1,2 o 3.
- name: Nom del passatger.
- sex: sexe del passatger.
- sibsp: Nombre de germans, germanastres, esposa o marit a bord.
- parch: Nombre de parents o fills o fillastres a bord.
- ticket: Número del ticket.
- fare: Tarifa cobrada al passatger.
- cabin: Número de cabina del passatger (pot no tenir cabina)
- embarked: Port d'embarcació, C=Cherbourg, Q=Queenstown, S=Southampton.

Procediment

Del fitxer *train.csv* usarem només els passatgers dels quals coneguem totes les variables 714 de 891, d'aquests basarem el nostre model amb el 75% d'aquestes dades, i el 25% restant l'usarem per a testejar el model fent prediccions i comprovant si encertem o no (proces corss-validation) [3].

Backwards Elimination

Pot passar que les dades de les que disposem en tinguem de no significatives, per saber si les hem de considerar o no per al nostre model usem el mètode backwards elimination.
Formalment no estem usant backwards elimination, doncs tenim variables categòriques i quan ens surti que una classe d'una variable té el p-valor més gran i sigui més gran que el α_{crit} i es compleixi que la majoria de les altres categories també tenen un p-valor més gran que el α_{crit}, aleshores considerarem que la variable no és explicativa.

Usant aquest procediment elminimem les variables port, fare i parch del model en aquest ordre.

Fent el summary del model sense aquestes variables ens trobem que la classe siblings1 (que representa tenir 1 germà) té un p-valor més gran que el α_{cirt}, pren un valor estimat molt proper a zero[4], però les altres variables de la mateixa categoria tenen un p-valor més petit que α_{cirt} amb un valor estimat semblant entre elles.
En aquest cas hem decidit fusionar classes, la variable siblings ara pendrà valors 0 o 2, on 0 indica que tens un germà o no en tens cap, i 2 indica que tens 2 o més germans.

Així doncs, de moment els coeficients del model són:

[3] `http://en.wikipedia.org/wiki/Cross-validation_(statistics)`
[4] És a dir, no canvien gaire la supervivència tenint un germà o no tenint-ne cap.

Variable	Coeficient	Std. Error	$P(Z > \|z\|)$
(Intercept)	3.524377	0.600190	< 0.0001
clase(2)	-0.713990	0.438813	0.10372
clase(3)	-1.902402	0.438813	< 0.0001
sexe(home)	-2.494997	0.214879	< 0.0001
edad	-0.050982	0.008599	< 0.0001
sibsp(2 o més)	-1.474878	0.402703	0.00025
cabin(tenen)	0.758821	0.397934	0.05653

Interacció vs no interacció

Pot ser interessant mirar si hi ha interacció entre aquestes variables explicatives. Hem considerat tots els tipus d'interacció possible i hem procedit de manera anàloga a quan hem decidit quines variables eren significatives per detectar quines interaccions eren significatives i quines no.
Hem arribat a la conclusió de què l'única interacció significativa és entre la classe i el sexe, obtenim el model amb els coeficients següents:

Variable	Coeficient	Std. Error	$P(Z > \|z\|)$
(Intercept)	5.318553	0.947965	< 0.0001
clase(2)	-1.184098	0.904101	0.190299
clase(3)	-4.214379	0.869209	< 0.0001
sexe(home)	-4.325316	0.768058	< 0.0001
edad	-0.059430	0.009603	< 0.0001
sibsp(2 o més)	-1.561253	0.427047	0.00025
cabin(tenen)	0.784309	0.435186	0.071508
clase(2):sexe(home)	-0.033751	0.918349	0.970683
clase(3):sexe(home)	3.053703	0.815910	0.000182

Com podem veure la interacció *clase(2):sexe(home)* no és gens significativa tot i així hem optat per mantenir la iteracció ja que entre *clase(3):sexe(home)* la interacció és molt significativa.

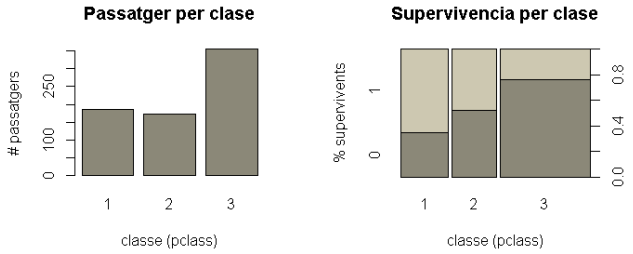
A més, el AIC del model sense interacció és més gran que el AIC del model amb interacció (el AIC és un indicador de mesura de qualitat relativa (més informació sobre AIC))

Fins i tot, comprovant amb cross-validation tenim que el model amb interacció prediu 84.83146% de les dades i el model sense interacció prediu 83.14607% de les dades, per tant, tot indica que considerar la interacció és una millora. No obstant en els dos casos els resultats prediuen notablement la supervivència dels passatgers.

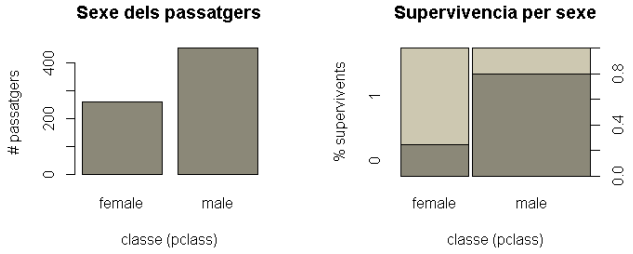
Visualització de les variables explicatives

És interessant, per tal de crear una idea intuïtiva, visualitzar les variables explicatives.

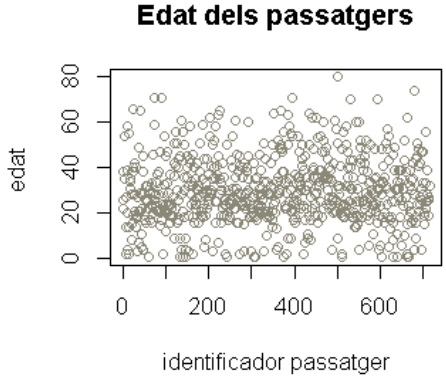
Comencem per la classe, es pot veure gràficament que la majoria dels passatgers pertanyen a la tercera classe, a més, es pot veure que els de la primera classe sobreviuen en un percentatge més elevat que els de la tercera classe. És a dir, sembla evident que la classe juga un paper important a la supervivència.



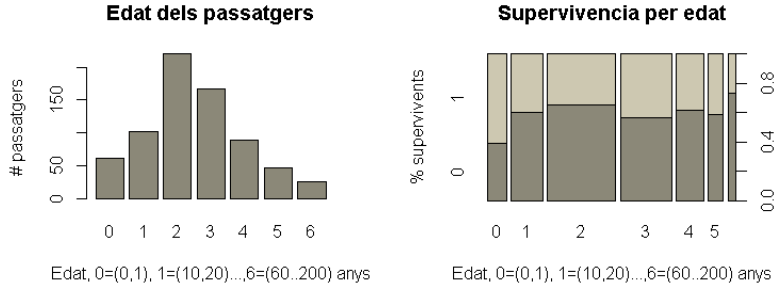
De la mateixa forma, podem veure que la majoria dels passatgers són homes, i que el percentatge de supervivència és molt més elevat en les dones que en els homes, per tant el sexe també ha de jugar un paper important.



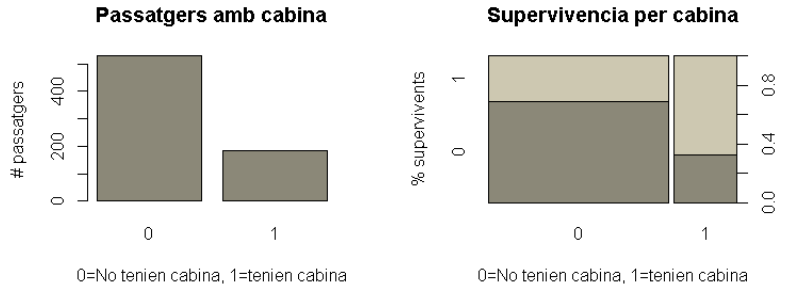
Si volem fer un gràfic anàleg amb l'edat, ens trobem que és difícil d'interpretar al no ser una variable categòrica:



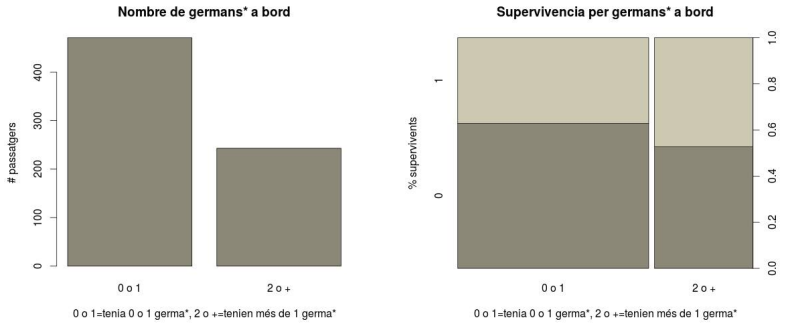
Per tant, dividirem l'edat en 7 categories: $(0, 10), (10, 20), (20, 30), \ldots, (50, 60), (60, 200)$ anys



Es pot veure doncs que el grup de passatgers més comú són els que tenen entre 20 i 40 anys, també podem observar que els que tenen una edat entre 0 i 10 anys tenen més probabilitat de sobreviure que els que tenen una edat de més de 60 anys.



Podem observar que els passatgers que disposen de cabina privada són un grup minoritari que, a més, té més probabilitats de sobreviure.



Similarment, en la comparativa entre tenir un o cap germà i tenir-ne dos o més, s'observa que, la majoria de passatgers no té cap germà a bord o en té un, a més, la probabilitat de sobreviure tenint 2 germans o més és més gran que la de no tenir germans (o tenir-ne un).

El recompte sobre les dades de les variables explicatives és el següent:

Edat	Sexe	Classe	Cavina	Germans	Mor	Viu	Total
Adult (+16)	Home	1	Si	≥2	2	0	2
				(0,1)	47	32	79
			No	≥2	0	1	1
				(0,1)	12	4	16
		2	Si	≥2	0	0	0
				(0,1)	2	1	3
			No	≥2	4	0	4
				(0,1)	76	5	81
		3	Si	≥2	0	0	0
				(0,1)	3	1	4
			No	≥2	5	0	5
				(0,1)	180	27	207
	Dona	1	Si	≥2	0	5	5
				(0,1)	2	63	65
			No	≥2	0	0	0
				(0,1)	5	48	53
		2	Si	≥2	0	0	0
				(0,1)	1	7	8
			No	≥2	0	3	3
				(0,1)	5	48	53
		3	Si	≥2	0	0	0
				(0,1)	1	2	3
			No	≥2	2	2	2
				(0,1)	37	25	62
Joves (-16)	Home	1	Si	≥2	0	0	0
				(0,1)	0	3	3
			No	≥2	0	0	0
				(0,1)	2	6	8
		2	Si	≥2	0	1	1
				(0,1)	0	2	2
			No	≥2	0	0	0
				(0,1)	2	6	8
		3	Si	≥2	0	0	0
				(0,1)	0	1	1
			No	≥2	19	1	20
				(0,1)	8	8	16
	Dona	1	Si	≥2	0	0	0
				(0,1)	1	5	6
			No	≥2	0	0	0
				(0,1)	0	9	9
		2	Si	≥2	0	1	1
				(0,1)	0	0	0
			No	≥2	0	0	0
				(0,1)	0	9	9
		3	Si	≥2	0	0	0
				(0,1)	1	1	2
			No	≥2	10	4	14
				(0,1)	4	13	17

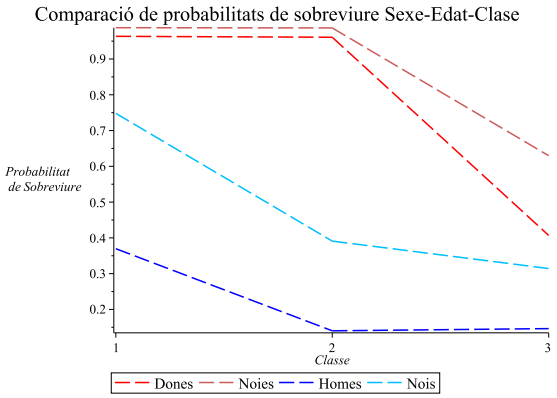
Resultats

Cross-validation

Al paràgraf anterior esmentem els resultats obtinguts usant cross-validatoin per comparar si la interacció millora la predicció o no, no obstant això, és interessant remarcar que el model final (amb interacció) prediu un 84.83146% de les dades.

Comparativa edat-sexe-classe

Els grups que destaquen més són el sexe, l'edat i la classe, és interessant comparar i quantificar si realment les dones, els nens i la primera classe tenen més probabilitats de sobreviure, i quanta probabilitat més, en el següent gràfic es pot veure la comparativa:



Per fer el gràfic hem considerat l'edat de les dones/homes com la mitjana de l'edat de les dones/homes majors de 16 anys, de la mateix amanera hem considerat l'edat de les noies(/nois) com la mitjana d'edat de les noies/nois menors de 16 anys, per les altres vairables hem usat un procediment anàleg, a vegades agafant la mediana en lloc de la mitjana.

En el gràfic s'observa que les dones són el grup amb més supervivència, els joves també són un grup especialment privilegiat (especialment joves homes i joves noies de tercera classe) i la classe també juga un paper rellevant (especialment en el cas dels homes entre primera i segona classe, i les dones entre segona i tercera classe).
Podem concloure doncs que el concepte: *les dones i els nens primer* és va practicar, més concretament les noies, les dones i els nois primer (i si ets ric, millor).

Debilitats del model

Usant el cross-validation hem guardat unes dades que no usem per construir el model, i les usem per a comprovar si encertem en les prediccions o no.
És interessant veure quins perfils de passatgers el nostre model acostuma a no predir correctament, per tal de millorar-lo en futures edicions.
En el cross-validation podem observar que el model falla algunes prediccions de passatgers de classe 2 o 3, que no tenien cabina i que tenien 2 o més germans a bord.

La següent taula mostra els errors del cross-validation, on (R:Mor P:Viu) significa que realment es mor, però el model preveu que viuen, i (R:Viu, P:Mor) significa que realment viu, però el model preveu que es moriren.

Edat	Sexe	Classe	Cavina	Germans	R: Mor, P: Viu	R: Viu, P: Mor	T. erros de predicció
Adult (+16)	Home	1	Si	≥2	0	**1**	**1**
				(0,1)	0	0	0
			No	≥2	0	0	0
				(0,1)	0	0	0
		2	Si	≥2	0	0	0
				(0,1)	0	0	0
			No	≥2	**1**	**6**	**7**
				(0,1)	0	0	0
		3	Si	≥2	0	0	0
				(0,1)	0	0	0
			No	≥2	**1**	**6**	**7**
				(0,1)	0	0	0
	Dona	1	Si	≥2	0	0	0
				(0,1)	0	0	0
			No	≥2	0	0	0
				(0,1)	0	0	0
		2	Si	≥2	0	0	0
				(0,1)	0	0	0
			No	≥2	**1**	**1**	**2**
				(0,1)	0	0	0
		3	Si	≥2	0	0	0
				(0,1)	0	0	0
			No	≥2	**1**	**1**	**2**
				(0,1)	0	0	0
Joves (-16)	Home	1	Si	≥2	0	0	0
				(0,1)	0	0	0
			No	≥2	0	0	0
				(0,1)	0	0	0
		2	Si	≥2	0	0	0
				(0,1)	0	0	0
			No	≥2	0	**1**	**1**
				(0,1)	0	0	0
		3	Si	≥2	0	0	0
				(0,1)	0	0	0
			No	≥2	0	**1**	**1**
				(0,1)	0	0	0
	Dona	1	Si	≥2	0	0	0
				(0,1)	0	0	0
			No	≥2	0	0	0
				(0,1)	0	0	0
		2	Si	≥2	0	0	0
				(0,1)	0	0	0
			No	≥2	0	0	0
				(0,1)	0	0	0
		3	Si	≥2	0	0	0
				(0,1)	0	0	0
			No	≥2	0	0	0
				(0,1)	0	0	0

Titànic: la pel·lícula

La pel·lícula del Titànic és molt coneguda, és una pel·lícula on es relata la història de Jack Dawson i Rose DeWitt Bukater, dos joves de classes diferents que es coneixen i s'enamoren al RMS Titànic.

En aquesta secció farem la predicció del final d'aquesta història d'amor. Suposant que la noia era un passatgera de 17 anys de la primera classe, amb 1 promès a bord i amb cabina pròpia, la probabilitat que tenia de sobreviure segons el nostre model és de 99.03899%, en canvi Jack Dawson era un noi de 20 anys de tercera classe, sense esposa a bord ni germans i sense cabina pròpia, per tant la probabilitat que tenia de sobreviure segons el nostre model és de 20.48467% .

Així doncs, no cal veure la pel·lícula per estimar que acabarà malament.