LUIS FELIPE VILLOTA

LUND UNIVERSITY
QDA USING R - SIMM61

LAB ASSIGNMENT 1 - LOGISTIC REGRESSION: TITANIC DATABASE

Introduction

This report is presented as an expert analysis required by the Court on the claim of Mrs. Sue against her father Mr. Leornardo, on the occasion of the Titanic catastrophe. According to Mrs. Sue, the fact that Mr. Leonardo was not present on the ship had a negative affectation on her and her mother's chances of survival. In this sense, the Court has requested a rigorous analysis of the database to determine statistically if *Leonardo's decision of not accompanying Sue and Kate on the trip decreased their chances of survival.*

Methods:

Final model description. After an extensive descriptive and exploratory data analysis, the technical party has first cleaned the dataset in which 179 passengers are removed due to missing values in the variables "Age" (177) and "Embarked" (2). We didn't remove the rest of the missing cases from "Cabin" (529) since it would have reduced a great amount of passengers from our database, - and we did not intend to use this variable in the present exercise. This left us with a database with a total of 712 "complete" observations/rows with 12 columns. Then, the variables were properly transformed into a correct type (*PassengerId, Survived, Pclass, Sex, Ticket, Cabin* and *embarked* were converted to factors). Finally, we have built a final model in which the outcome is the dichotomous variable of "Survived" with two levels (*Yes* or *No)*. The five predictors chosen were: *Age*, *Sex*, *Pclass*, *SibSp* and *Parch*. The **regression equation** is, *Y (Survival) = 4.34 + (-0.04) *Age + (-2.63) *Sex + (-1.40) *Pclass2nd + (-2.64) *Pclass3rd + (-0.36) *SibSp + (-0.03) *Parch*. We note that *Pclass1st* is the reference level for predictor *Pclass* and *Female* is the reference in *Sex*. (We see the complete estimates in *Table i*).

Results:

Goodness of fit and effectiveness of the model. As seen in *Table ii*, our final model and the null model (without predictors) are significantly different from each other in terms of prediction accuracy. The model with predictors is significantly better (Chi^2 = 324.96, df = 6, $p < 0.001$, AIC of the model = 650, -2LL of the model = 636, AIC of null model = 963, -2LL of the null model = 961) for explaining variation for the outcome variable of interest (*Survived*). The model with predictors has an AIC of more than 2 points lower than the null model, which makes it significantly better at the prediction accuracy. Our model, after accounting for all the variance

(explained by the predictors included) has a much smaller total error left (sums of differences in probability between the predicted outcome and the actual outcome for each observation) as noted. Since R^2 is not an appropriate index for logistic regression, we use pseudo R-squared methods to assess the the proportion of explained variance. In this case, we will use the McFadden R^2 index, - which gives us ~34% for our model and 0 % for the null. In the dataset provided only 288 (40%) passengers survived and 424 didn't (60%). Our final model has an overall prediction rate of 80% (573 of 712 correct cases in total), for those who actually survived it has an 73% (210 out of 288 of actual survivors) and an 86% (363 out of 424 actual deaths) for those who actually died, as shown in *Table iii*.

Influence of predictors. Through dominance analysis *Age*, *Sex*, *Pclass* and *SibSp* show significant p-values and according to our confidence intervals and these are different from zero so they add predictive value to the model. *Parch* was not significant. Nevertheless, *Sex* was found to be the most influential predictor with an average R^2m of 0.205 as seen in *Figure 1*.

As for the **probabilities for Sue and Kate**, after introducing these passengers' characteristics (personal, way of travel and accompaniment), the results of the equation presented were then exponentiated to have the log odds for each individual in the investigation and then transformed in to a probability scale with (exp(Y) / (1 + exp(Y))) *100 where Y is the equation prediction. This revealed the following results: in Leonardo's absence 82% of survival chances for Sue and 70% for Kate. With Leonardo´s presence 81% for Sue and 62% for Kate, as shown in *Table iv*.

Discussion

With the information given by the court, a quantitative data analysis approach was used and by means of the logistic regression method, **the technical party has gathered evidence that rejects Mrs. Sue's hypothesis**. What was found is that Mr. Leonardo's absence during trip did not likely worsened the chances of survival of Mrs. Sue and her mother (as claimed); - on the contrary, his presence would have diminished them even more. The most influential predictor for survival was "Sex" as mentioned before; - to exemplify the difference between levels (women and men), if "Sex" was the only predictor for "Survival" it can be said for all passengers that if you were female you had ~75% of probabilities of surviving and ~20% if you were a male. In combination with the other significant predictors such as Pclass and Age (in a minor contribution) we find the main fluctuations in chances of survival. The presence of a spouse, yet, had very little additive value of prediction to the final model. The fact of being

accompanied by a parent or traveling with children had no significance in order to explain variance in chances of survival, as demonstrated in this report.

R code: https://github.com/FelipeVillota/SIMM61_QDA-with-R/blob/main/titanic.R

Appendix

Table i. Regression coefficients

	Survived					
Predictors	*Odds Ratios*	*std. Beta*	*CI*	*standardized CI*	*Statistic*	*p*
(Intercept)	77.23 ***	16.58	32.48 – 194.55	9.88 – 28.81	9.53	**<0.001**
Age	0.96 ***	0.52	0.94 – 0.97	0.41 – 0.66	-5.46	**<0.001**
Sex [male]	0.07 ***	0.07	0.05 – 0.11	0.05 – 0.11	-11.97	**<0.001**
Pclass [2nd]	0.24 ***	0.24	0.14 – 0.42	0.14 – 0.42	-4.94	**<0.001**
Pclass [3rd]	0.07 ***	0.07	0.04 – 0.12	0.04 – 0.12	-9.25	**<0.001**
SibSp	0.69 **	0.71	0.54 – 0.88	0.56 – 0.89	-2.90	**0.004**
Parch	0.96	0.97	0.76 – 1.21	0.79 – 1.18	-0.31	0.757
Observations	712					
R^2 Tjur	0.410					
Deviance	635.943					
AIC	649.943					
log-Likelihood	-317.972					

** p<0.05 ** p<0.01 *** p<0.001*

Table ii. Model comparison with performance indices (rounded up)

Model	Log likelihood ratio test	-2 Log likelihood	AIC
Final with predictors	-318	636	650
Without predictors (null model)	-480	961	963

Table iii. Percentages of prediction accuracy

Total prediction % of the final model	Prediction % for those who actually survived	Prediction % for those who actually died
80% (573 of 712 cases in total)	73% (210 out of 288 of actual survivors)	86% (363 out of 424 actual deaths)

Table iv. Probabilities of survival for Sue and Kate crossed with Leonardo's presence and absence

Presence of Leonardo	Sue	Kate
Absent	82%	70%
Present	81%	62%

Figure 1. Dominance Analysis