Assignment 1 - The Titanic - Discrete Choice Modelling

Marc Sparhuber

Table of contents

Task A

Table 1: Passengers on the Titanic: Descriptive statistics

		Mean	SD	N	Max	Min	Percent
	Age	29.70	14.53	714	80.00	0.42	100.00
	Survived	0.38	0.49	891	1.00	0.00	100.00
	Family	0.32	0.47	891	1.00	0.00	100.00
Pclass	1			216			24.24
	2			184			20.65
	3			491			55.11
Sex	male			577			64.76
	female			314			35.24

Comments: Data from the Titanic R package.

The data set "Titanic" contains information on the 891 recorded individuals on board the eponymous Titanic, the famous passenger and mail carrying ocean liner that sank on April 15th 1912 after colliding with an iceberg. Table 1 displays some of the core variables of importance to later analyses. Survived and Family are both dummy variables, with 1 indicating survival of the sinking in the prior and whether the passenger had family on board in the latter. Aside from the usual demographic information contained in the variables Age and Sex, Pclass displays the passenger class, with 1 being the highest and 3 the lowest.

Contrary to today's cruises perhaps, the mean age of passengers was merely ~30. Importantly, the standard deviation is quite high at 14.53. It is plausible to assume that these values are greatly influenced by entire families being passengers and the mean being easily influenced by outliers such as babies with values such as the minimum at 0.42. Meanwhile, the two dummy variables Survived and Family are to be interpreted carefully, with the maximum and minimum values representing the only two possible values and the standard deviation not being properly interpretable for dichotomous categorical data. It can, however, be seen that 38% of all passengers survived the catastrophe. Regarding the passenger class, it is apparent that the two more prestigious classes had fewer people in them, while more than 50% were in third class. Finally, the distribution of sex shows that about two thirds of all passengers were male.

Task B

Table 2: Survival from Titanic. Linear probability models

	Model 1	Model 2
(Intercept)	0.234***	0.671***
	(0.038)	(0.057)
female	0.547***	0.485***
	(0.032)	(0.031)
Age	−0.001	−0.006***
	(0.001)	(0.001)
Pclass2		−0.211***
		(0.042)
Pclass3		−0.414***
		(0.039)
Family		−0.044
		(0.031)
Num.Obs.	714	714
R2	0.291	0.392
R2 Adj.	0.289	0.388

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Source: Titanic R package.

Comments: Pclass refers to passenger class, first class is the reference category.

Table 2 shows the results of two linear probability models estimating the survival of passengers. While both models use survival as their dependent variable, the first model only inspects the relation of survival with being female and age. Its central question is to answer whether "women and children first" accurately describes the outcomes observed during the sinking of the Titanic. The second model adds to this query as independent variables both passenger class and whether family was present. Notably, only 714 of the initial 891 passengers are inlcuded, which is due to many missing values in the age variable.

In model 1 being female is significantly ($p < 0.001$) associated with a 55 percent point increase in likelihood for survival compared to being a man. Age, with only about -0.1 percentage points, very slightly contributes to an decrease in survival chances the older the passenger, although this should be interpreted with care, not being significant. The intercept is significant ($p < 0.001$) at ~23%, indicating that when all other factors are held constant at 0, the estimated probability of survival is around 23%. The total variance explained by R^2 and adjusted R^2 is around

29%, which considering that only 2 independent variables were used shows their importance in determining survival during the sinking of the titanic.

In model 2 being female and age's beta-coefficients only slightly change, with being female's effect on survival now being slightly lower at around 49 percentage points while still being highly significant ($p < 0.001$) and age's negative effect on survival rising to around -0.6 percentage points while becoming highly significant ($p < 0.001$). Interestingly, the chance of survival holding all else constant at 0 has increased tremendously to 67% while also being significant ($p < 0.001$). This is likely influenced by the passenger class variable, as passengers of class 1 had a much higher likelihood of survival, with passenger's of class 2 and 3 having highly significant ($p < 0.001$) beta coefficients of -21 and -41 percentage points, respectively. Including family as an independent variable in the model produces a non-significant beta-coefficient of -4 percentage points, indicating that interpretations of this result should be careful. Finally, the explained variance rises by around 10 percentage points, confirming that the inclusion of passenger class and family was helpful to explaining survival.

In summary, the question of whether the "women and children first" saying is accurate in the case of the Titanic's sinking cannot be entirely supported. While it is certainly the case that being a woman highly increased the chances of surviving the sinking of the Titanic, being young had only very slight advantages for survival. It is important to note that this may be more closely examined if looking at age as a categorical variable instead, which might show whether children (within a defined age range) may have had a greater chance of survival. Model 2 adding passenger class as an independent variable shows that perhaps the aforementioned phrase should be changed to "the women and rich first", especially considering the moderate negative correlation between fare price and passenger class being -0.55.