IMT573 Problem Set 7: Logistic Regression

Deadline: Sat, Dec 2nd, 5pm

Instructions

This problem set revolves around logistic regression, in particular interpretation of logistic regression results. It contains two parts: Titanic data, and World Value Survey data. Both parts are fairly similar in terms of the tasks you have to do.

1 Titanic: What Happened During Her Last Hours? (40pt)

Titanic and her maiden voyage

Titanic was a huge luxurious ocean liner that sank in night to April 15, 1912, after hitting an iceberg at full speed during her maiden voyage.

By Willy Stöwer - Magazine Die Gartenlaube, Public Domain,
https://commons.wikimedia.org/w/index.php?curid=97646

In those years, the maritime safety protocols were very different from what they are now. First, and according the standard practices of the time, Titanic did not carry enough boats to fit all her passengers. There were boats for 1,178 people only, for only about half of the approximately 2,200 passengers and crew onboard. The reason was that the boats were designed to transfer passengers to another ship nearby, and not as the sole floating devices for the all passengers for many hours. Second, the crew was untrained for such emergencies. There were little understanding about how many people one can fit in the boats, and how to launch collapsible boats. As a result, only about 700 people made they way into boats, almost all others died in the icy waters. The ship did not have a public address system, so stewards had to go door-to-door and wake up passengers. This may have had implications for third class passengers who may have had little idea about what was happening.

Our information about the last hours of the ship mostly originates from the survivors' accounts (but also from radio messages and analysis of the wreckage). Briefly, the story is as follows: approximately half an hour after the collision, the captain initiated evacuation. According to the habits of the time, women and children were first to get to boats. There are also accounts of officers lowering half-full lifeboats as there were not more women and children around, and leaving men on board. In practice, "women and children" meant 1st and 2nd class women and children as for the third class passengers it was much harder to reach up to the boat deck. They also received much less direction, and many of them did not understand explanations in English.

But are these accounts correct? Are survivors of an unimaginable maritime disasters reliable sources about the last chaotic hours on the ship, the technical side of which they had very little idea? Maybe the order on the ship broke early, and instead of women and children, these were strong young men who made it? Your task is to find it out!

1.1 Titanic Data

First, get to know your data. Each row contains one passenger and there is data about 1309 individuals. The dataset contains the following variables:

survived : Survived (1) or died (0)

pclass : Passengers class

name : Passengers name

sex : Passengers sex

age : Passengers age

sibsp : Number of siblings/spouses aboard

parch : Number of parents/children aboard

ticket : Ticket number

fare : Fare

cabin : Cabin

embarked : Port of embarkation

boat number of lifeboat the person was found in

body number of body if body identified

home.dest the final destination

1. (2pt) load file *titanic.csv*, and do quick sanity checks.
2. (3pt) find the number of missings in the important variables. You are definitely going to use variables *survived*, *pclass*, *sex*, *age*, and you may use more (see below).
3. (4pt) Are there implausible values that are technically not missing?

1.2 Logistic Regression

Now it is time to analyze the survival.

1. (4pt) Based on the survivors accounts, described above, which variables do you think are the most important ones to describe titanic survival? How should those be related to the survival? (should they increase or decrease the probability of survival)
2. (2pt) Create a new variable *child*, that is 1 if the passenger was youger than 14 years old.
3. (4pt) Explain why do we have to treat *pclass* as categorical. Convert it to categorical using `factor(pclass)`.

4. (4pt) Estimate a multiple logistic regression model where you explain survival by these variables. Show the results.

5. (6pt) Interpret the results. Did men or women, old or young have larger chances of survival? What about different passenger classes? How big were the effects?

6. (4pt) Experiment with other variables you see fit for this task. For instance, you may create a variable "young man" and see if they survived more likely than others.

   Do other variables change your results in any major way?

7. (7pt) Based on the results above, explain what can you tell about the last hours on Titanic. Are the survivors' accounts broadly accurate? Did the order break down? Can you tell anything else interesting?

2 Religiousness and Attitude toward Women

The next task is technically similar in a sense we work with logistic regression and try to understand what do the results tell us. Your task is to find out if those who have more conservative attitude toward women are also more religious. The analysis is based on World Value Survey, a global survey of various personal attitudes and habits, including politics, gender roles, environment, work and much more. Most of the questions ask whether the respondents agree or disagree with various claims (e.g. *religion is more important than science*). We have prepared a small subset of the questions for this analysis.

First some preparatory stuff:

1. (3pt) Load the data *wvs-logit.csv*. Consult the corresponding readme file for the exact definition of the variables. Perform basic consistency checks. Ensure you know the data types (numeric/something else) of the variables.

2. (4pt) Remove the missing values. These may be in two forms: a) missing in the sense of NA values, and b) negative or empty values.

   Report the size of the final cleaned dataset.

3. (4pt) Ensure you know coding of all variables. Convert the variable coding into a form where larger number corresponds to more affirmative answer to the question the variable answers. For instance, currently:

   - *believeInGod* is coded as 1-yes, 2-no. Recode it in a way that 1-yes, 0-no.

     Hint: this recoding can be done as `2 - believeInGod`.
   - *environmentImportant* is coded as 1-very much ... 6-not at all. Recode this in a way that 0=not at all ... 5=very much.

     Hint: you can recode this variable in a similar way.
   - Should education be categorical or not? Explain, and convert if you think it should be categorical.

4. (4pt) In the following analysis we use variable *believeInGod* as a measure of "religiosity". Do you think it is a good question for this purpose? Do you see any issues here? (Consult the exact questions in the readme file).

5. (4pt) In a similar fashion, as a proxy for "attitude toward women" we are using agreement to the statement "Having a job is the best way for a woman to be an independent person". Do you think it is a good variable? If you were to design a new survey, would you come up with a different question?

Now we are done with the preparations. Time for some modeling. But before we get there, a few words about how to include the variables. Some of the variables are clearly of nominal measure (like *continent*), while others are ordered (e.g. *womanJob*). I recommend to include the ordered ones as numbers (essentially assuming these are interval measures). Although not quite correct, it substantially simplifies the analysis.

6. (7pt) Run a logistic regression model where you explain religiosity with the attitude toward women.

   Interpret the coefficients: are those who think that job is the best way for a woman to be an independent person more or less religious than those who do not think so? By how much? Is the effect statistically significant?

7. (7pt) Add *age* to the model, in addition to *womanJob*. What do you find–are younger or older people more religions?

8. (7pt) Add also education, importance of environment, and continent to the model (i.e. include all variables in data). Discuss the results:
   (a) re-visit the question of gender role as you answered in 6
   (b) are people who care more about environment more religious or not? By how much?
   (c) how does religiosity associate with education?
   (d) which is the least and the most religious continent?

Finally tell us how many hours did you spend on this PS.

References