Estimating the effect of luxury cabin on the chances to survive Titanic

March 27, 2019

Calendar reads 15 April Anno 1912. The whole world is listening to the Radio, waiting to hear the precise numbers of the disaster. The RMS Titanic, a British passenger liner, is sinking in the Atlantic Ocean after colliding with an iceberg.

In the next several days, it becomes clear, that more than 1500 of its 2224 passengers have died. The whole cruiser industry is facing hard times.

You are a statistician working for the Labour party. You are interested on the **effect of the economic status (class) of the cabine** on the chances for survival as a passenger on a liner. You obtain from the British Board of Trade, a governmental body investigating the case, a unique dataset with information for each passenger on the board, including whether he died or not.

1. Use the R package "datasets". It contains the dataset Titanic.

2. Checkout the variables in the dataset using help(Titanic). Read the description of this dataset.

3. Save the dataset under a name "mydata" in your working session, so I can follow all your steps.

4. Write a section "Data and descriptives" in which you describe your dataset: which variables, summary statistics for your variables, some nice associations (correlations, correlation matrices, densities, etc.), maybe some nice plots. Use your skills obtained in the lecture to provide an interesting description of your data. Which associations do you find interesting (on a premature, i.e. non-causal level)? Which assertions do you have after exploring the dataset? This section should be around a page, a page and a half. You might want to check some empirical paper (e.g. the one that I send to you) to see how such a section looks like.

5. Write a section "Empirical strategy and results" to describe well... your empirical results and strategy.

   (a) Describe the model with which you would like to answer the question. Start with the simplest model, e.g. simple linear regression.

   (b) Define the object of interest and say what is its interpretation.

   (c) What is actually the independent variable measuring? Can the effect be interpreted in a different way?

   (d) Can you imagine a randomized experiment, in which the main independent variable is assigned per random? When you describe it, you should disregard cost, ethical or legal aspects.

   (e) Which are the main assumptions behind this simple model?

   (f) What do you think is contained in the error term of the equation?

   (g) Can you estimate a more complex model? What is the added value of the higher complexity? Which assumption is more credible?

   (h) Estimate your simple model. What does your result say?

   (i) Is your result a result of a random choice of your sample?

   (j) Suppose you had the hypothesis that the economic class does have an effect on the chances to survive. Write the corresponding Null hypothesis.

   (k) How can you test this hypothesis within your simple model? Describe the result. What could be a meaningful interpretation (if the result is correct)?

   (l) Now perform a thorough model diagnostics of your simple model. Describe your conclusions (and add pictures if possible).

   (m) Repeat your estimations for the complex model. Compare the results.

   (n) Suppose that your more complex model is correct. What does it tell you about the estimator in your simple model? Can you quantify the bias? Or at least tell in which direction it is likely to be?

6. Summarize your conclusions in a way the policy maker can understand them.