ASSIGNMENT 5

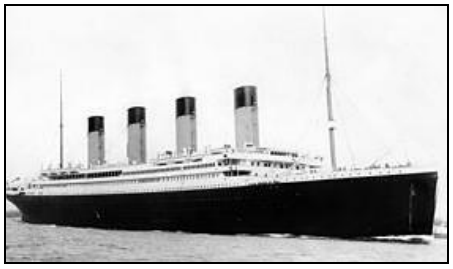
BINARY LOGISTIC REGRESSION

Binary logistic regression describes the relationship between a categorical dependent variable with two possible outcomes (0 or 1) and one or more independent variables. As the mean of a binary variable is a probability, the logistic regression model expresses the probability as a function of explanatory variables.

In this lab assignment you will use a binary logistic regression model and SPSS to analyze and interpret data related to the passengers of the British ocean liner *Titanic* that sank in 1912 after colliding with an iceberg. In particular, you will explore the impact of sex, ticket class, and age on a passenger's likelihood of surviving the shipwreck.

The Titanic Disaster

On April 15, 1912, during her maiden voyage, the British ocean liner *Titanic*, the largest ship afloat at the time, sank in the North Atlantic Ocean after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew.



In this lab assignment you will use dataset that describes the survival status of 1309 passengers on the Titanic. The data does not contain information for the crew, but it does contain actual and estimated ages for almost 80% of the passengers.

This dataset is based on the *Titanic Passenger List* edited by Michael A. Findlay, originally published in Eaton & Haas (1994) *Titanic: Triumph and Tragedy*, Patrick Stephens Ltd, and expanded with the help of the internet community.

The data are available in SPSS file *lab5.txt* located on *STAT 337* Laboratories web site at http://www.stat.ualberta.ca/statslabs/stat337/index.htm (click *Stat 337* link, and *Data* for *Lab 5*). The data are not to be printed in your submission. The following is a description of the variables in the data file:

Variable Name	**Description of Variable**
name	full name of the passenger,
pclass	passenger class, (1=1st; 2=2nd; 3=3rd); proxy for socio-economic status (SES) 1st ~ Upper; 2nd ~ Middle; 3rd ~ Lower;
survived	survival (0=No; 1=Yes);
sex	gender (female or male);
age	age (in years); fractional if age is less than 1.

Use the data to answer Questions 1-5.

1. Is the study an example of an observational study or a randomized experiment? Can the study be used to conclude that one of the genders was more apt to survive the shipwreck? Can the results of the study be generalized to a broader population? Provide brief explanations.

2. Now you will examine the relationship between survival, gender and passenger class with bar chart and cross-tabulation tool in SPSS.

   (a) Obtain a stacked bar chart displaying the percentages of males and females who survived or perished in each passenger class. The variable sex should be placed on categorical axis, stacks defined by survival status and columns used for passenger class. Paste the chart with the title, the names of the axes, and a legend explaining the meaning of the each colour in the chart into your report. Comment briefly about association between survival, gender and passenger class.

   (b) Use *Crosstabs* tool under *Descriptive Statistics* in the *Analyze* menu to obtain the survival summaries for each gender and passenger class. Paste the summaries into your report. Also obtain the chi-square statistics for each passenger class. Report the overall Pearson's chi-square statistic for the three passenger classes combined and corresponding p-value. Comment about association between survival and gender.

   (c) Using hand calculation and your cross-tabulation in part (b), obtain the odds of survival for each gender (ignore passenger class). What is the odds ratio? Interpret this odds ratio.

3. You will now use a binary logistic regression model to estimate the odds of survival for passengers of each gender. Paste relevant SPSS output into your report (relevant means all outputs necessary to answer the following questions).

   (a) Define a logistic regression model with survival as the response variable and gender as the covariate.

   (b) Run the logistic model in part (a) for the Titanic data. What is the estimated logistic regression model?

   (c) Does gender have any association with survival? State relevant hypotheses in terms of the odds ratio (OR) of the association between gender and survival. Report the test-statistic, the distribution of the test statistic under the null hypothesis and p-value of the test. State your conclusion.

   (d) Report the 95% confidence interval for the odds ratio for survival between females and males. Is the inference about the association between gender and survival from the confidence interval consistent with the inference of the hypotheses test in part (c)? Explain briefly.

   (e) Interpret the estimated intercept and the slope obtained in part (b). What are the odds of survival for males and females? What is the estimated odds ratio of association between gender and survival? Is the odds ratio consistent with your hand calculation from the cross-tabulation in part (c) of Question 2?

4. Now you will expand the model of part (b) of Question 3 to include passenger class *pclass* as another predictor. Use the binary regression tool in SPSS to fit the model with survival as dependent variable and gender and passenger class as independent variables. Paste relevant SPSS output into your report.

   (a) Define the logistic model with survival as dependent variable and gender, passenger class as independent variables.

   (b) Run the binary logistic regression tool for the model specified in part (a). Note that passenger class is a categorical variable with three categories (1, 2 and 3) so the last category (third class) is automatically the reference category (odds for survival for all the other categories will be compared to the reference in the output). Report the estimated logistic regression model.

(c) Test the utility of the present logistic model defined in part (a) by stating the null and alternative hypotheses, reporting the value of the G statistic, state the distribution of the statistic under the null hypothesis, and the p-value. Refer to *Omnibus Tests of Model Coefficients* table to report the required statistics. State briefly your conclusions.

(d) Assess the fit of the present logistic model with the Hosmer-Lemeshow goodness-of-fit test. In particular, define the null and alternative hypotheses; report the value of the test statistic, the distribution of the test statistic under the null hypothesis, and the p-value of the test. State your conclusions briefly.

(e) Use the drop-in-deviance test and the SPSS output in Question 3 to determine whether or not the explanatory variable *pclass* is adding significantly to the predictive ability of the model. In particular, define the null and alternative hypotheses; hand-calculate the test statistic, state the distribution of the test statistic under the null hypothesis, and the p-value of the test. State your conclusions. Compare the outcome with the Wald's test for the variable in part (b).

(f) Use the estimated regression model in part (b) to compare the odds of survival of passengers in first class with the odds of survival of passengers in third class. Also compare the odds of survival of females to males. What were the odds of survival for females in first class?

(g) Are the log odds of survival associated with passenger class different for males than for females? Define a logistic model to answer the above question and report the estimated coefficients for the model.

(h) Use the estimated regression model to compare the odds of survival of first class male passengers with the odds of survival of third class male passengers. Also compare the odds of survival of female passengers in first class with the odds of survival of female passengers in third class. Finally, compare the odds of survival for female and male passengers in third class.

5. Now you will expand the model of Question 4 to include one more predictor: *age*. Use the binary regression tool to fit the model with survival as dependent variable and age, gender and passenger class as independent variables. Paste relevant SPSS output into your report.

(a) Define the logistic model with survival as dependent variable and gender, passenger class, interaction of passenger class and gender and age as independent variables.

(b) Report the estimated logistic regression model.

(c) Assess the fit of the present logistic model with the Hosmer-Lemeshow goodness-of-fit test. In particular, define the null and alternative hypotheses; report the value of the test statistic, the distribution of the test statistic under the null hypothesis, and the p-value of the test. State your conclusions briefly.

(d) Use the drop-in-deviance test and the SPSS output in Question 4 part (g) to determine whether or not the explanatory variable *age* is adding significantly to the predictive ability of the model. In particular, define the null and alternative hypotheses; hand-calculate the test statistics, state the distribution of the test statistic under the null hypothesis, and the p-value of the test. State your conclusions.

(e) Use the estimated regression model in part (b) to obtain the odds-ratio for survival between a 20-year-old woman and a 40-year-old woman in third class. Also obtain the odds of survival of a 20-year old man in first class and the odds of survival a 20-year old man in third class.

(f) Estimate the change in the odds of survival for each one year increase in age (holding all other predictors constant) with a 95% confidence interval.

LAB 5 ASSIGNMENT: MARKING SCHEMA

Question 1

Observational study or experiment: 2 point
Causational relationship between survival and gender: 2 points

Question 2

(a) Stacked bar chart: 3 points
    Association between survival, gender and passenger class: 3 points
(b) Survival summaries for each gender and passenger class: 3 points
    Overall Pearson's chi-square statistic: 2 points
    Comments about association between gender and survival: 2 points
(c) Odds of survival for females: 2 points
    Odds of survival for males: 2 points
    Odds ratio: 1 point

Question 3

(a) Logistic regression model: 2 points
(b) Estimated logistic model: 2 points (output included)
(c) Test about the association of gender and survival:

    Null and alternative hypotheses: 2 points
    Test statistic: 1 point
    Distribution of test statistic: 1 point
    P-value: 1 point
    Conclusion: 1 point

(d) 95% confidence interval: 3 points
    Consistency of the interval and the outcome of the test: 2 points
(e) Intercept interpretation: 2 points
    Slope interpretation: 2 points
    Odds of survival for males (based on the estimated regression model): 2 points
    Odds of survival for females: 2 points

Question 4

(a) Logistic model: 2 points
(b) Estimated logistic regression model: 2 points
(c) Omnibus tests:

    Null and alternative hypotheses: 2 points
    Test statistic: 1 point
    Distribution of test statistic: 1 point
    P-value: 1 point
    Conclusion: 1 point

(d) Hosmer-Lemeshow goodness-of-fit test:

    Null and alternative hypotheses: 2 points
    Test statistic: 1 point
    Distribution of test statistic: 1 point
    P-value: 1 point
    Conclusion: 1 point
(e) Drop-in-deviance test:

Null and alternative hypotheses: 2 points
Test statistic: 3 points
Distribution of test statistic: 1 point
Estimate of p-value: 1 point
Conclusion: 1 point

(f) Comparison of odds of survival for first class passengers with those of third class: 2 points
Comparison of odds of survival for females and males: 2 points
Odds of survival for females in first class: 2 points

(g) Logistic model: 2 points
Estimated logistic model: 2 points

(h) Comparison of odds of survival for first class male passengers with those of third class: 2 points
Comparison of odds of survival for first class male passengers with those of third class: 2 points
Comparison of odds of survival for females and males in third class: 2 points

Question 5

(a) Logistic model: 2 points
(b) Estimated logistic regression model: 2 points
(c) Hosmer-Lemeshow goodness-of-fit test:

Null and alternative hypotheses: 2 points
Test statistic: 1 point
Distribution of test statistic: 1 point
P-value: 1 point
Conclusion: 1 point

(d) Drop-in-deviance test:

Null and alternative hypotheses: 2 points
Test statistic: 3 points
Distribution of test statistic: 1 point
Estimate of p-value: 1 point
Conclusion: 1 point

(e) Odds-ratio for survival between a 20-year-old and a 40-year-old woman in third class: 2 points
Odds of survival of a 20-year old man in first class: 2 points
Odds of survival a 20-year old man in third class: 2 points

(f) 95% confidence interval for age: 3 points

TOTAL= 112

Henryk Kolacz
University of Alberta