# A Study in Disaster

*The Titanic Survivors*

mauveSushi

Skidmore College

# Introduction

Motivation and Purpose and the data itself ...

# 1. Data Exploration

## 1.1 Summary

The dataset contains 130

## 1.2 Variable Selection

## 1.3 Data Preparation

## 1.4 Correlation Investigation

# 2. Methodology

## 2.1 Backward Selection

AN

## 2.2 Logistic Regression

AN

In our project, since our response varialbe is binary variable, we are going to use logistic regression for the model. Logistic regression is a type of generalized linear model. Its equation is in the form of $ln(\frac{p}{1-p}) = \ln(odds) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$, where $p$ is the probability of our outcome, and $x_1, x_2, \ldots, x_n$ are explanatory variables. The right hand side of this equation is the same as a linear model, so it is called generalized linear model. The left-hand side is the natrual log of odds, where odds is a representation of probability, and it eqauls to $\frac{p}{1-p}$. Once we fit a model, we can predict the probability of success $p$ by $p = \frac{1}{1+e^{-RHS}}$, where $RHS = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$.

To perform logistic regression, we can use `glm` function, which stands for generalized linear model, in R, and call`family = binomial(link = "logit")`.

## 2.3 Homser-Lemeshow goodness-of-fit test

Since we used logistic regression in our project, the assumptions we need to verify are not the same as a linear regression model. (For example, we cannot plot a residual plot to see whether the residuals have a pattern because however good or bad a model is, all the residuals will be laid on the curve $\frac{1}{1+e^{-predicted}}$. That is, the residuals will always show a pattern.) To examine how good a logistic regression model is, we can use Homser-Lemeshow goodness-of-fit test. The idea of this test is to divide the sample into several groups according to their predicted values, and compare the expected proportion of success to the observed proportion of success

in each group to see whether there is a significant difference between the expected and the observed probability. The null hypothesis of this test is that there is no difference between the expected and the observed probability. In other words, if the p-value of this test is too low, we will have strong evidence that the fit is not good enough.

To perform Homser-Lemeshow goodness-of-fit test, we can use `hoslem.test` function in `ResourceSelection` package in R.

# 3. Model

## 3.1 Model obtained using backward selection

The first model we obtained is by backward selection. We eliminated `number_of_siblings_and_spouses` and `fares` before we obtained a model whose p-value of each variable is smaller than 0.05.
The model we obtained and its summary are shown below:

```
##
## Call:
## glm(formula = has_survived ~ gender + age + number_of_siblings_and_spouses +
##     passenger_class + embarked_from, family = binomial(link = "logit"),
##     data = titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7755  -0.6791  -0.4560   0.7074   2.5697
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      4.86097    0.47163  10.307  < 2e-16 ***
## gendermale                      -2.60736    0.15798 -16.505  < 2e-16 ***
## age[19,55]                      -2.00712    0.38995  -5.147 2.65e-07 ***
## age[56, above)                  -3.10763    0.53229  -5.838 5.27e-09 ***
## age[6,18]                       -1.76429    0.43159  -4.088 4.35e-05 ***
## agemissing                      -2.21886    0.42028  -5.280 1.30e-07 ***
## number_of_siblings_and_spouses -0.35361    0.09297  -3.803 0.000143 ***
## passenger_classsecond           -0.91904    0.21489  -4.277 1.90e-05 ***
## passenger_classthird            -1.77251    0.19438  -9.119  < 2e-16 ***
## embarked_fromQ                  -0.47350    0.30504  -1.552 0.120601
## embarked_fromS                  -0.67719    0.18791  -3.604 0.000314 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1736.2  on 1305  degrees of freedom
## Residual deviance: 1191.1  on 1295  degrees of freedom
## AIC: 1213.1
##
## Number of Fisher Scoring iterations: 5
```

As can be seen from the summary, each variable has a p-value that is much lower than 0.05 except `emabrked_fromQ`. However, since the p-value of `emabrked_fromS` is low enough, we do not need to eliminate `embarked_from` variable. Also, there is no huge change of coefficients while we eliminate `number_of_siblings_and_spouses` and `fares`, so we do not need to worry about colinearity for these two variables. (Otherwise, we have to study whether the low p-value is caused by colinearity). The formula of this model is

$$\frac{\hat{p}}{1-\hat{p}} = 4.86097 - 2.60736 \times gendermale - 2.00712 \times age[19,55] - 3.10763 \times age[56,above) - 1.76429 \times age[6,18]$$

. Now, we shall evaluate this model using Homser-Lemeshow goodness-of-fit test, and the outcome is as shown below:

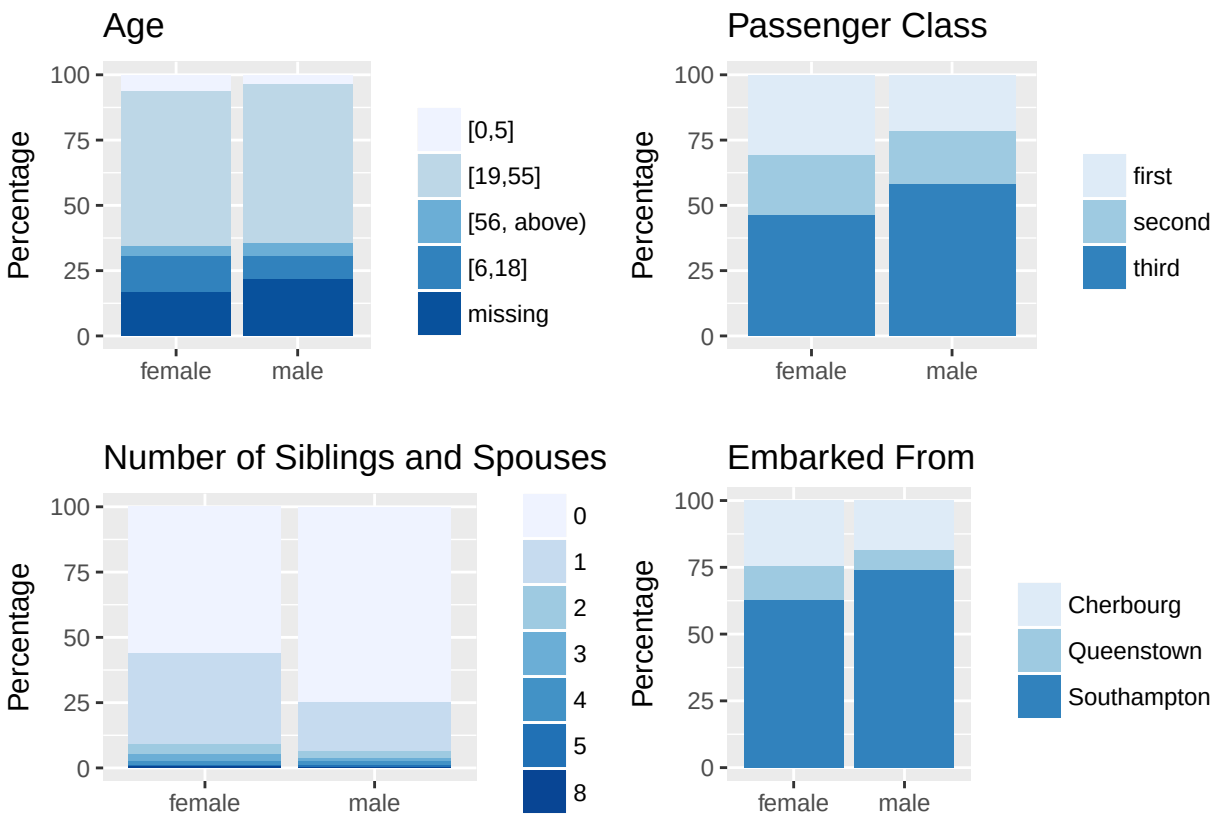
```
hl <- hoslem.test(titanic$has_survived, fitted(m_best), g=8)
hl
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  titanic$has_survived, fitted(m_best)
## X-squared = 22.308, df = 6, p-value = 0.001065
```

The p-value is 0.001065, which is much lower than 0.05, and this suggests that this model is very likely to have a lack of fit. Hence, we need to find a better alternative of this model.

## 3.2 The Final Models

To get our final model, we first verified the skeptical correlations we mentioned in the Data Exploration section. First, we are going to check the correlation of gender with other variables, and we can use the plots below to see it:




It seems that there is not a significant correlation between age and gender, or gender and embarked from, but there

# 4. Conclusion

## 4.1 Findings

## 4.2 Further Study

# Bibliography