Logistic Regression I

In this tutorial, we will learn how to fit a (binary) logistic regression model in R. Logistic regression is used when the response variable is binary. We model the log odds of "success" as a linear combination of coefficients and predictors:

$$\log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_1 + \cdots \beta_k x_k.$$

Before proceeding, we usually have to distinguish if we have *ungrouped* or *grouped* data. *Ungrouped* data refers to data recorded at the individual level; *grouped* data refers to data recorded at a group level.

1) Ungrouped data

The first example will be based on ungrouped data. The dataset `titanic.txt` consists of data on some of the passengers on the Titanic passenger liner that sank on April 15, 1912. The variables are

- `Survived`: 0 for did not survive, 1 for survived
- `Sex`: gender of the passenger
- `Age`: age of the passenger
- `Fare`: fare paid by the passenger

We will use `Survived` as the response variable, and see if the odds of survival on the Titanic is associated with the other predictors.

Read the data in, and load the `tidyverse` package for some of the data visualizations that we will create prior to fitting a logistic regression

```
library(tidyverse)
Data<-read.table("titanic.txt", header=TRUE)
```

a) Exploratory data analysis

Given that our response variable is categorical, we can create a simple table to see how many people survived and did not survive the Titanic, as well as the proportions

```
##frequency table
table(Data$Survived)
```

```
##
##   0   1
## 424 283
```

```
##table with proportions
prop.table(table(Data$Survived))
```

```
##
##         0         1
## 0.5997171 0.4002829
```

We have 424 who did not survive, and 283 who survived. Or about 60% of passengers did not survive and 40% survived.

To see how survival rate differs by gender, we can create a two-way table as well as a table showing survival rate by gender.

```
##2 way table
mytab<-table(Data$Sex, Data$Survived)
mytab
```

```
##
##            0   1
##   female  64 195
##   male   360  88
```

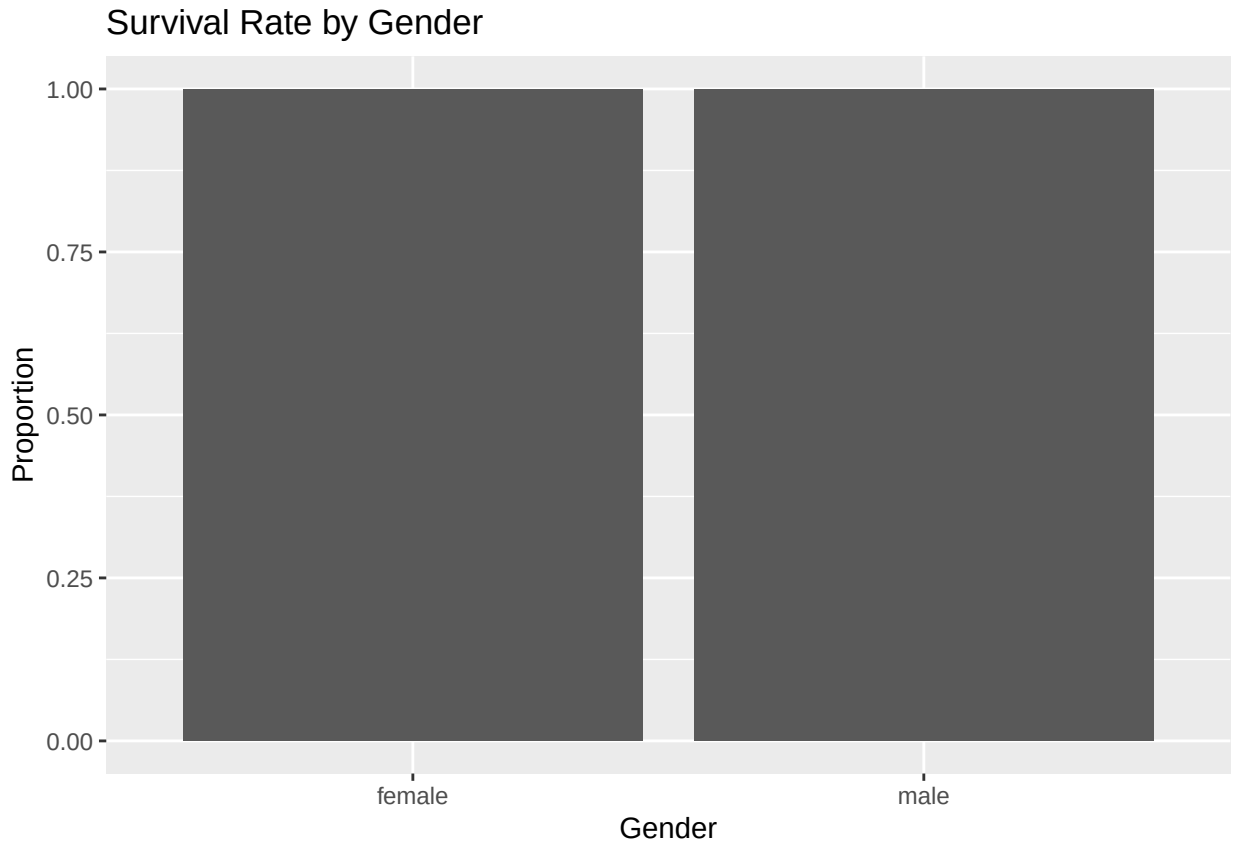
```
##survival rate for each gender
prop.table(mytab, 1)
```

```
##
##                   0         1
##   female 0.2471042 0.7528958
##   male   0.8035714 0.1964286
```

We can see there were more males in this data set. The survival rate for females was about 75%, which is a lot higher than the survival rate for males, which was about 20%. Based on this table, gender is probably going to a significant factor for survival on the Titanic.

We can create a bar chart to display the survival rates across genders, since both variables are categorical

```
ggplot(Data, aes(x=Sex, fill=Survived))+
  geom_bar(position = "fill")+
  labs(x="Gender", y="Proportion",
       title="Survival Rate by Gender")
```



Notice this bar chart is not displaying the information correctly. This is because the variable `Survived` was coded using 0-1 indicators. While such numeric coding is fine when using `lm()` or `glm()`, variables coded numerically may need to be converted to a factor when creating data visualizations. So we convert `Survived`, give descriptive names to the levels, and recreate the bar chart

```
is.numeric(Data$Survived)
```

```
## [1] TRUE
```

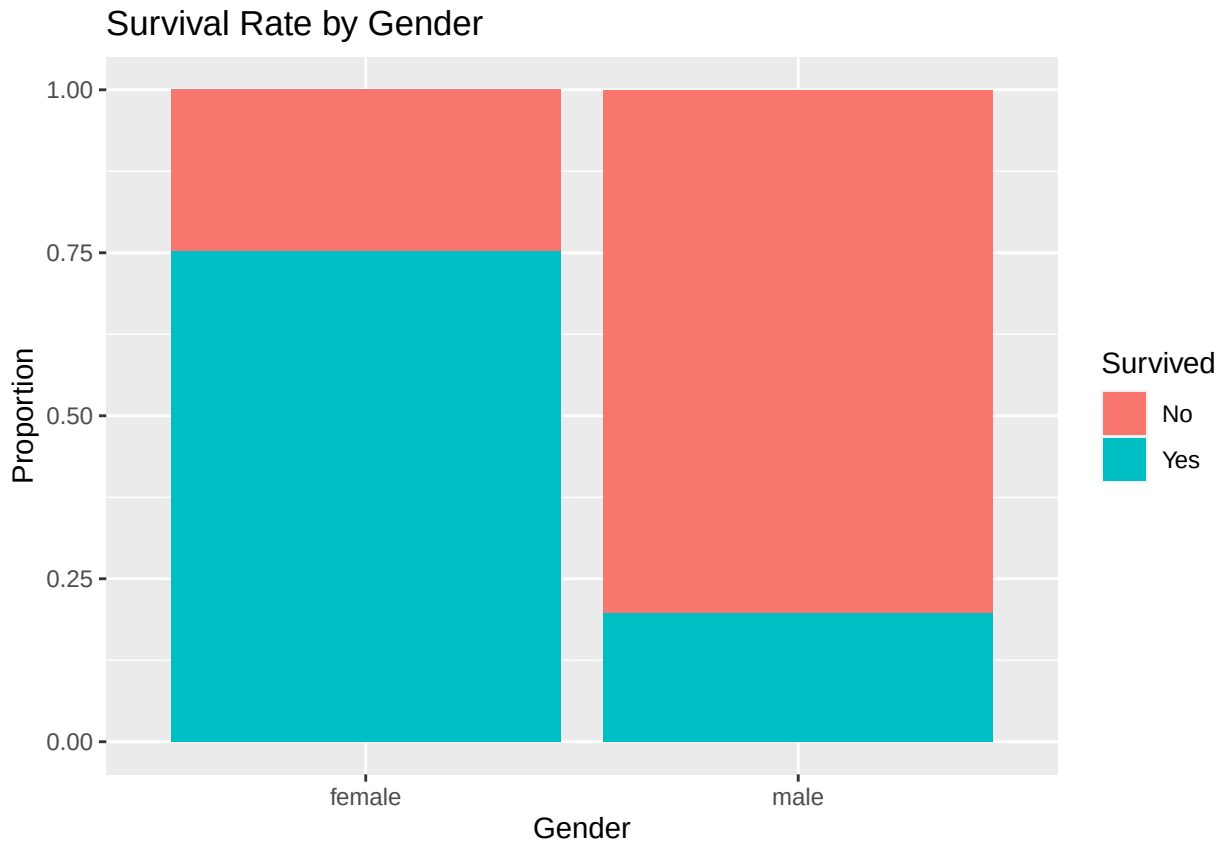
```
##change Survived to factor for visuals
Data$Survived<-factor(Data$Survived)
levels(Data$Survived)
```

```
## [1] "0" "1"
```

```
##notice 0 first, then 1
##give descriptive labels
levels(Data$Survived) <- c("No","Yes")
levels(Data$Survived)
```

```
## [1] "No"  "Yes"
```

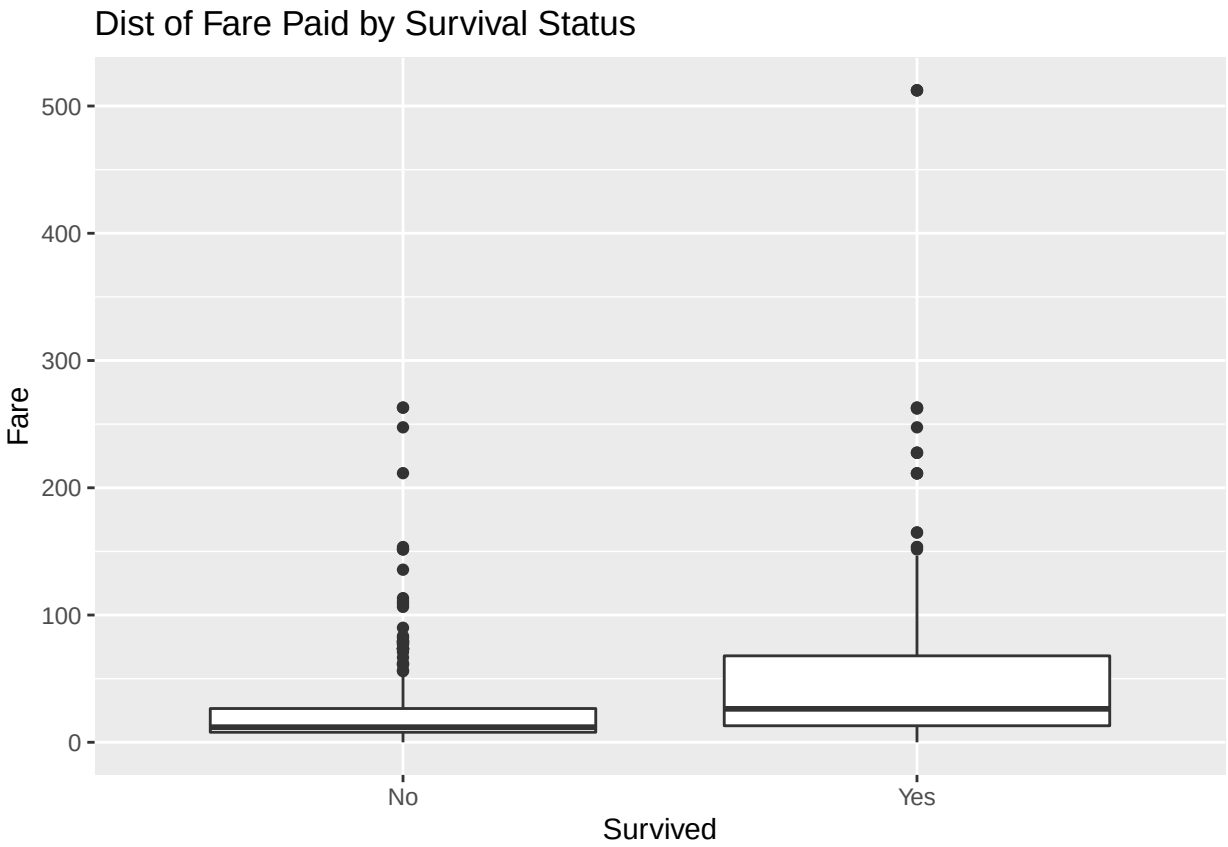
```
##Recreate bar chart
ggplot(Data, aes(x=Sex, fill=Survived))+
  geom_bar(position = "fill")+
  labs(x="Gender", y="Proportion",
       title="Survival Rate by Gender")
```



The bar chart is a visual representation of the survival rates by gender.

Next, we can create some visuals involving `Survived` and one of the quantitative predictors, `Fare`. We can use a side by side boxplot

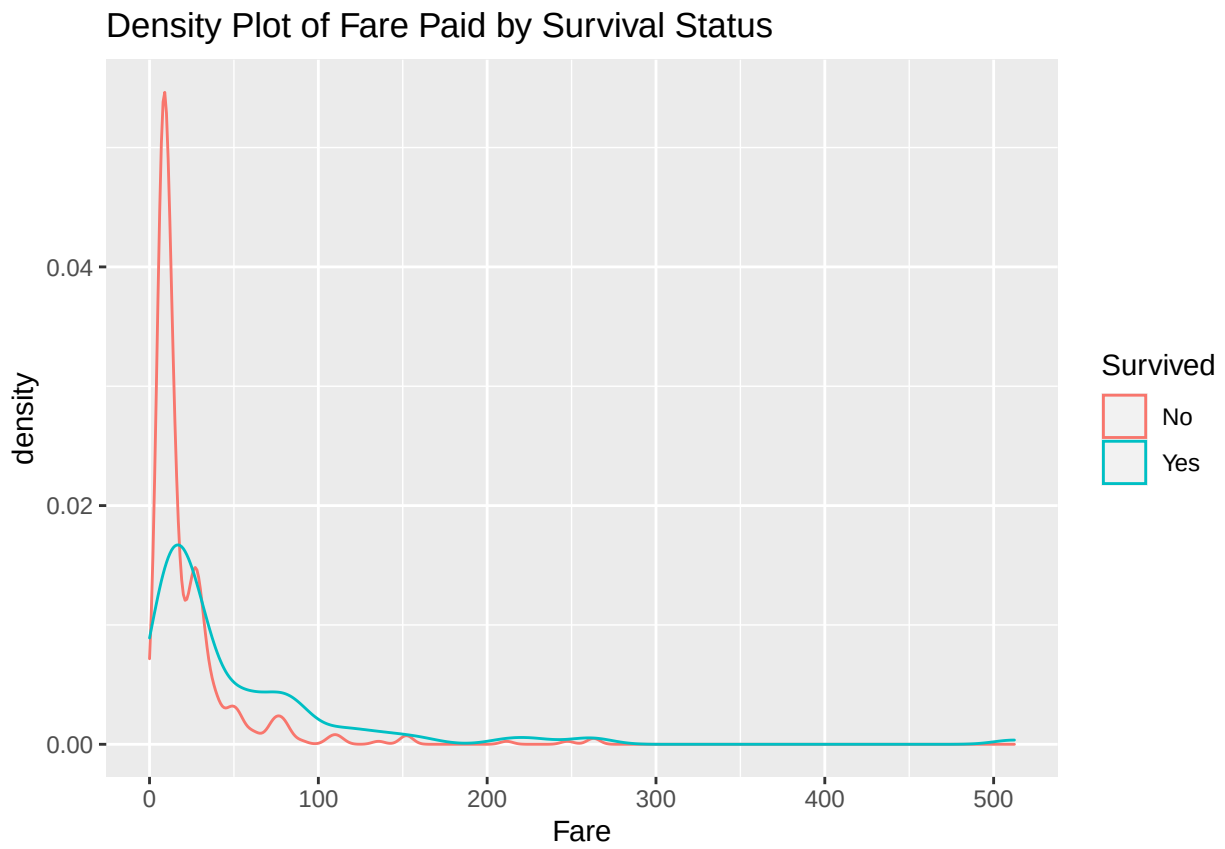
```
##side by side boxplots
ggplot(Data, aes(x=Survived, y=Fare))+
  geom_boxplot()+
  labs(title="Dist of Fare Paid by Survival Status")
```



We can see that those who did not survived generally paid lower fares. So higher paid fares may be associate with higher survival rates.

We can also create density plots to compare the distribution of fares paid between survivors and non survivors

```
##density plots
ggplot(Data,aes(x=Fare, color=Survived))+
  geom_density()+
  labs(title="Density Plot of Fare Paid by Survival Status")
```



We see that among non survivors, the distribution is highly right skewed; a huge proportion of them paid low fares (around less than $20). While the distribution is also right skewed among survivors, a smaller proportion of them paid low fares, compared to non survivors.

As practice, create your own visuals to see how age is associated with survival on the Titanic.

b) Fitting logistic regression

Based on the EDA, we suspect that gender is going to be a strong predictor of survival. We start by fitting this 1-predictor logistic regression, using `glm()`

```
result<-glm(Survived ~ Sex, family = "binomial", data=Data)
```

Notice we specify the argument `family = "binomial"`. This has to be specified for a logistic regression. If this is not specified, a linear regression is fitted instead. The function `glm()` uses maximum likelihood estimation whereas `lm()` uses ordinary least squares.

```
summary(result)
```

```
##
## Call:
```

```
## glm(formula = Survived ~ Sex, family = "binomial", data = Data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6721  -0.6613  -0.6613   0.7534   1.8041
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.1141     0.1441   7.734 1.04e-14 ***
## Sexmale      -2.5229     0.1868 -13.506  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 951.80  on 706  degrees of freedom
## Residual deviance: 733.52  on 705  degrees of freedom
## AIC: 737.52
##
## Number of Fisher Scoring iterations: 4
```

From `summary()`, we can write our logistic regression equation as

$$\log(\frac{\hat{\pi}}{1-\hat{\pi}}) = 1.1141 - 2.5229 I_1$$

where $I_1 = 1$ for males and 0 for females. The estimated slope of -2.5229 can be interpreted as:

- the log odds of surviving is 2.5229 less for males than for females, OR
- the odds of surviving for males is $\exp(-2.5229) = 0.0802$ times the odds of surviving for females.

The Z test associated with β_1 is highly significant. The test statistic is $Z = \frac{-2.5229}{0.1868} = -13.506$. The p-value can be found using `pnorm(-13.506)*2` which is close to 0.

The reported null deviance of 951.80 is analogous to the SS_{res} of an intercept-only model. The reported residual deviance of 733.52 is analogous to the SS_{res} of our 1-predictor model.

c) More hypothesis tests in logistic regression

Suppose we consider adding both `Age` and `Fare` as predictors to the model

```r
##full model
full<-glm(Survived ~ ., family = "binomial", data=Data)
summary(full)
```

```
##
## Call:
## glm(formula = Survived ~ ., family = "binomial", data = Data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3366  -0.6233  -0.5843   0.7988   1.9835
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.792778   0.240997   3.290    0.001 **
## Sexmale     -2.401160   0.192239 -12.491  < 2e-16 ***
## Age         -0.005457   0.006655  -0.820    0.412
## Fare         0.012409   0.002704   4.589 4.45e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 951.80  on 706  degrees of freedom
## Residual deviance: 701.61  on 703  degrees of freedom
## AIC: 709.61
##
## Number of Fisher Scoring iterations: 5
```

To test $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ versus H_a : at least one of the coefficients in H_0 is not zero, we calculate a ΔG^2 test statistic, and compare it with a χ^2 distribution with 3 degrees of freedom (3 because we are testing 3 coefficients). The ΔG^2 test statistic in this scenario is just the difference in the null deviance and residual deviance of the 3-predictor model

```r
##test if coefficients for all 3 predictors are 0
##test stat
TS<-full$null.deviance-full$deviance
TS
```

```
## [1] 250.1889
```

```
##pvalue
1-pchisq(TS,3)
```

```
## [1] 0
```

So we reject the null hypothesis. The 3-predictor model is chosen over the intercept-only model; the 3-predictor model is useful.

Another hypothesis test that we can do is to compare between the 1-predictor model and the 3-predictor model. In this scenario, we are testing $H_0 : \beta_2 = \beta_3 = 0$, H_a : at least one of the coefficients in H_0 is not zero.

We calculate a ΔG^2 test statistic, and compare it with a χ^2 distribution with 2 degrees of freedom (2 because we are testing 2 coefficients). The ΔG^2 test statistic in this scenario is the difference in the residual deviances of both models

```
##test if additional predictors have coefficients equal to 0
##test stat
TS2<-result$deviance-full$deviance
##pvalue
1-pchisq(TS2,2)
```

```
## [1] 1.178861e-07
```

So we reject the null hypothesis, and prefer the 3-predictor model over the 1-predictor model.

Notice that the coefficient for `Age` in the 3-predictor model is insignificant, so we can actually drop `Age` from the model and fit a 2-predictor model.

2) Grouped data

In this second example, we are dealing with grouped data. Observations with the same value on the predictor(s) are grouped together.

We will use `dose.txt`. The first column denotes the dose level of a chemical given to a group of insects on a $\log_{10}$ scale, the second column denotes the number of insects that died in that group, and the third column denotes the number of insects in that group.

```
library(tidyverse)
Data2<-read.table("dose.txt", header=T)
Data2
```

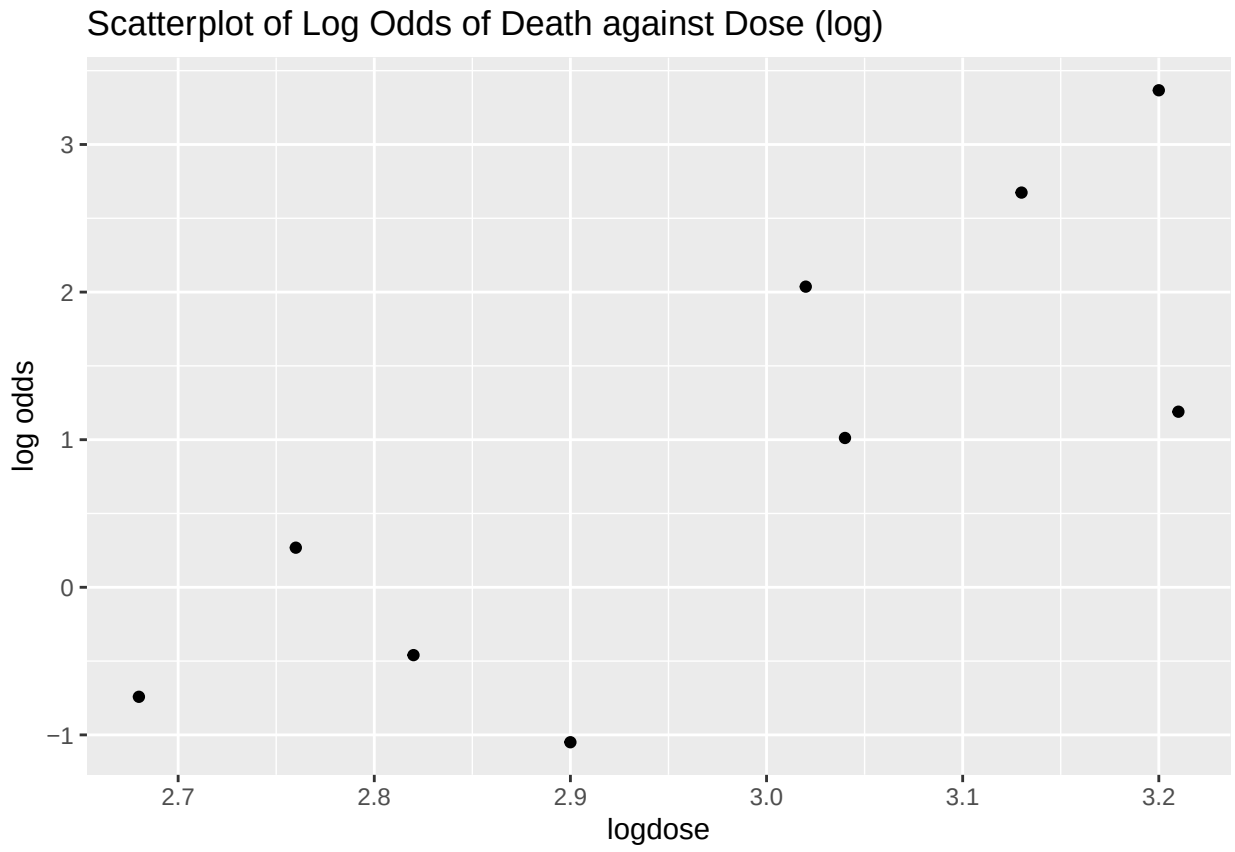
a) Exploratory data analysis

Recall that in logistic regression, we model the log odds of "success" against the linear combination of coefficients and predictors. With grouped data, we can create a scatterplot to visualize this relationship, with sample log odds against the predictor. So we need to calculate the sample log odds of dying for each group

```
##calculate proportion that died
Data2$prop<-Data2$died/Data2$size
##calculate log odds that died
Data2$log.odds<-log(Data2$prop/(1-Data2$prop))
```

And then create our scatterplot

```
##plot log odds against dose
ggplot(Data2, aes(x=logdose,y=log.odds))+
  geom_point()+
  labs(x="logdose", y="log odds",
       title="Scatterplot of Log Odds of Death against Dose (log)")
```



If a logistic regression is appropriate, we expect a linear relationship between log odds of death and the predictor in this scatterplot. We do not have a linear relationship, so we should not proceed with the rest of the analysis.

b) Fitting logistic regression with grouped data

We only proceed to show how to fit a logistic regression when we have grouped data

```
##fit logistic regression with grouped data
result2<-glm(prop~logdose, family="binomial",
             weights=size, data=Data2)
```

Notice that we use the vector of proportions as the response variable, and we add an argument `weights=size`, which is a vector of the number of observations in each group.

c) Goodness of fit tests

Goodness of fit (GOF) tests can be performed with grouped data. The hypothesis are

$$H_0 : \log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_1$$

and

$$H_0 : \log(\frac{\pi}{1-\pi}) \neq \beta_0 + \beta_1 x_1$$

Essentially, the null hypothesis is stating the log odds of dying is a linear combination of coefficients and the predictor, or in other words, the logistic regression fits the data.

There are two main GOF tests: Pearson's χ^2 test and deviance GOF test. Both test statistics are compared with a χ^2_{N-p} distribution, where N denotes the number of groups, and p denotes the number of parameters in the model.

For Pearson's χ^2

```
N<-dim(Data2)[1]
p<-2
##pearson chi square goodness of fit
pearson<-residuals(result2,type="pearson")
##calculate pearson's chi sq
X2<-sum(pearson^2)
X2
```

```
## [1] 28.5638
```

```
##p-value
1-pchisq(X2,N-p)
```

```
## [1] 0.0001737254
```

For the deviance GOF

```
##calculate the test stat
result2$deviance
```

```
## [1] 29.34616
```

```
##p-value
1-pchisq(result2$deviance,N-p)
```

```
## [1] 0.0001250955
```

Notice that the test statistics and p-values from both tests are fairly similar. The tests are asymptotically equivalent.

We end up rejecting the null hypothesis, and so the data supports the alternative hypothesis, that the logistic regression does not fit the data well, which is not surprising given the scatterplot that we produced earlier.