Stat 6021 R Tutorial: Logistic Regression

Today, we will learn how to fit a (binary) logistic regression model in R. Recall that in a logistic regression model, we have

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

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. 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

   (a) Fitting logistic regression model. We will start by modeling the log odds of survival against fare paid. To do this, type

   ```
   result<-glm(Survived ~ Fare, family = "binomial")
   ```

   Notice the argument `family` has to be specified as `"binomial"` for a logistic regression. You can actually fit a linear regression model by specifying `"gaussian"` instead. The function `glm()` uses maximum likelihood estimation whereas `lm()` uses ordinary least squares.

   (b) Interpreting estimated coefficients for logistic regression. Use `summary(result)` to obtain the estimated coefficients and interpret them in context.

   (c) Inference. We use the Wald test to test if the coefficients are significant. How is the `z value` calculated? Is the fare paid a significant predictor for survival on the Titanic? How do we construct confidence intervals for β_1?

(d) Null and residual deviance. Suppose instead we fit a “full” model with gender and age of the passenger as additional predictors. Use the `summary` function to obtain the relevant output. What do null deviance and residual deviance tell us? How are their degrees of freedom calculated?

(e) Testing all coefficients. We can test whether all coefficients in a logistic regression model are zero, i.e.

$$\beta_1 = \beta_2 = \beta_3 = 0.$$

by computing ΔG^2 = null deviance − residual deviance, and compare this test statistic against a χ^2_{p-1} distribution. What is the ΔG^2 test statistic for this model telling us?

(f) Testing some coefficients. Suppose we want to test the following

$$\beta_2 = \beta_3 = 0.$$

We do so by comparing the residual deviances of the models in questions 1a and 1d. Calculate the test statistic and state a relevant conclusion.

2. The second example will be based on grouped data. 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.

(a) Plot of sample log odds against the predictor. With grouped data, you can create a plot of sample log odds against the predictor, to have an idea if a linear relationship exists between the two. To do so, type

```
prop<-died/size
plot(logdose, log(prop/(1-prop)))
```

(b) Model fitting. To fit the model with grouped data, type

```
result<-glm(prop~logdose, family="binomial", weights=size)
summary(result)
```

(c) With grouped data, we can carry out goodness of fit testing with Pearson’s χ^2 and deviance goodness of fit tests.

i. For Pearson’s χ^2, type

```
pearson<-residuals(result,type="pearson")
X2<-sum(pearson^2)
X2
```

To find the p-value, type `1-pchisq(X2,9-2)`.

ii. Deviance Goodness of Fit. What is the test statistic for this test? What is the relationship between Pearson’s χ^2 and deviance goodness of fit tests for grouped data?