# Lab 5 - Logistic Regression

## Lab Goals

In this lab we will explore logistic regression models for binomial data in R. In particular, we will examine:

1. the `glm()` function
2. interpreting coefficients
3. diagnostics

## Cocaine Treatment - Binary Responses

An experiment was conducted to evaluate the effectiveness of two different drugs to prevent relapses in cocaine addiction. A random sample of 72 former cocaine addicts were randomly assigned to one of three drug treatments (`Drug`). One treatment group received the drug Lithium. Another treatment group received the drug Desipramine. The last treatment group received a placebo. After a specified treatment period, the binary response `Relapse` was recorded where `Relapse = yes` indicates the individual returned to cocaine use and `Relapse = no` indicates the individual did not return to cocaine use. The data can be found in the *CocaineTreatment.csv* file in the folder for Lab 5.

Load this data set both into the console below and in this R Markdown document using the code chunk provided for you.

```
CocaineTreatment = read.csv('CocaineTreatment.csv')
```

### Data Organized as Bernoulli Trials

Here the data are considered as 72 bernoulli trials. We will run a logistic model with response `Relapse` and predictor `Drug`. By default, the model estimates $p_i = P(Relapse_i = yes)$ (Why? It chooses the last level alphabetically as the "success" level). However, you can change the ordering of the levels to estimate $p_i = P(Relapse_i = no)$. Since we are interested in how well the drugs prevented relapse, we will change the ordering here. Furthermore, we would like to have the placebo as the reference level for `Drug`. So, we will also re-order `Drug` to make `Placebo` the first level.

```
#Change no to the last level of Relapse
CocaineTreatment$Relapse = factor(CocaineTreatment$Relapse, c("yes","no"))

#Change Placebo to the first level of Drug
CocaineTreatment$Drug = factor(CocaineTreatment$Drug, c("Placebo","Desipramine","Lithium"))
```

The `glm()` function will fit generalized linear models in R. The logistic regression models are just one type of generalized linear models. Here is the generic format of the `glm()` function:

–

glm(*formula*, family = *family*, data=*data*)

–

The arguments in *italics* need to be replaced by the specific information for your model. Where

1) *formula* is of the form `Response ~ predictor1 + predictor2 + predictor3 + ...`
2) for logistic regression *family* is `binomial`
3) *data* is the name of your data set

Note: These are not the only arguments that are valid for this function. We will explore this function more later.

Here we will fit the model `coc1.glm` and run a summary of the output.

```
#Fit the model
coc1.glm=glm(Relapse~Drug,family=binomial,data=CocaineTreatment)

#Summary
summary(coc1.glm)
```

```
##
## Call:
## glm(formula = Relapse ~ Drug, family = binomial, data = CocaineTreatment)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3232  -0.7585  -0.6039   1.0383   1.8930
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.6094     0.5477  -2.938   0.0033 **
## DrugDesipramine   1.9459     0.6866   2.834   0.0046 **
## DrugLithium       0.5108     0.7226   0.707   0.4796
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 91.658  on 71  degrees of freedom
## Residual deviance: 81.220  on 69  degrees of freedom
## AIC: 87.22
##
## Number of Fisher Scoring iterations: 4
```

1. Suppose this is an appropriate model for these data. Does there appear to be any outliers?
2. What is the fitted model for the treatment group that received Lithium?
3. What is the fitted model for the treatment group that received Desipramine?
4. For which drug was the estimated rate of relapse less than 50%?
5. How can we interpret the partial slope associated with `DrugDesipramine`?
6. Using the summary information, perform a test to determine whether `Drug` is a significant predictor for this model.

i) State the null and alternative hypotheses.

ii) The test statistic.

iii) The reference distribution

iv) The following code will calculate the p-value for this test. What do we conclude?

7. The `anova()` function will also perform this test. First we will need to run the intercept only model. The reduced (intercept only) model should be the first argument of `anova()`. The second argument should be the full model. You also need to include the option `test = "LRT"`. Verify the p-value for this test is the same as the one calculated above.

```
#Intercept only
cocint.glm=glm(Relapse~1,family=binomial,data=CocaineTreatment)

#LRT
anova(cocint.glm,coc1.glm,test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Relapse ~ 1
## Model 2: Relapse ~ Drug
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        71     91.658
## 2        69     81.220  2   10.438 0.005413 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## Data Organized as Binomial Observations

Since we have 24 independent observations for each drug treatment, we can also run the logistic model in R with Binomial observations. The data set *CocaineTreatment2.csv* contains a Binomial observation for each level of `Treatment`. The variable, `Yes`, is the number who relapsed on the drug treatment. `No` is the number who did not relapse on the drug treatment. The total number of observations (trials) for each drug treatment is `Yes` + `No`.

```
CocaineTreatment2 = read.csv('CocaineTreatment2.csv')
```

We can still use the `glm()` function to run a logistic regression model when the data follow a Binomial distribution. In comparison to the run for Bernoulli data, the only change we need to make is to the *formula*.

The format of *formula* should now be: `cbind(Successes,Failures) ~ predictor1 + predictor2 + ...` where

i) `Successes` is the number of successes for each Binomial observation

ii) `Failures` is the number of failures for each Binomial observation

Here we will refit the logistic regression model for predicting the probability of not having a relapse. Again, the first step is to reorder the levels of `Treatment`. Verify that the parameter estimates and Wald tests for this model match those from our first model.

```
CocaineTreatment2$Treatment=factor(CocaineTreatment2$Treatment,c("Placebo","Desipramine","Lithium"))

coc2.glm=glm(cbind(No,Yes)~Treatment,family=binomial,data=CocaineTreatment2)

summary(coc2.glm)
```

```
##
## Call:
## glm(formula = cbind(No, Yes) ~ Treatment, family = binomial,
##     data = CocaineTreatment2)
##
## Deviance Residuals:
```

```
## [1]  0  0  0
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -1.6094     0.5477  -2.938   0.0033 **
## TreatmentDesipramine   1.9459     0.6866   2.834   0.0046 **
## TreatmentLithium       0.5108     0.7226   0.707   0.4796
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.0438e+01  on 2  degrees of freedom
## Residual deviance: 1.3323e-15  on 0  degrees of freedom
## AIC: 16.08
##
## Number of Fisher Scoring iterations: 3
```

Another way…

```
coc3.glm=glm(No/(No+Yes)~Treatment,family=binomial,weights=(No+Yes),data=CocaineTreatment2)

summary(coc3.glm)
```

```
##
## Call:
## glm(formula = No/(No + Yes) ~ Treatment, family = binomial, data = CocaineTreatment2,
##     weights = (No + Yes))
##
## Deviance Residuals:
## [1]  0  0  0
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -1.6094     0.5477  -2.938   0.0033 **
## TreatmentDesipramine   1.9459     0.6866   2.834   0.0046 **
## TreatmentLithium       0.5108     0.7226   0.707   0.4796
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.0438e+01  on 2  degrees of freedom
## Residual deviance: 1.3323e-15  on 0  degrees of freedom
## AIC: 16.08
##
## Number of Fisher Scoring iterations: 3
```

1. Why are the deviance residuals all 0?
2. What do you notice about the `Null deviance`? Has the LRT changed?