# Workshop 10
# Classification: Logistic Regression, Specificity, Sensitivity and ROC

### Exercise 1 Logit function

$\text{logit}(p)$ is defined to be $\log\left(\frac{p}{1-p}\right)$ for $p \in [0, 1]$. Show that if $\text{logit}(p) = a$ then

$$p = \frac{e^a}{1 + e^a}.$$

### Exercise 2 Logistic regression coefficients

A logistic regression with one predictor variable $x$ gives the *linear predictor* $\text{logit}(p) = 4 - 2x$

(a) What value does $p$ take when $x = 0$.

(b) For which value of $x$ is $p = f(x) = 0.5$.

(c) Is the function $f(x)$ monotonic increasing or decreasing with $x$?

(d) For the linear predictor $\text{logit}(p) = \beta_0 + \beta_1 x$, summarise in words the effect of the coefficients $\beta_0$ and $\beta_1$ on the probability curve?

### Exercise 3 Classification matrix

A company has designed a pattern recognition program to identify iPhones from a set of photos of either iPhones or of Android phones. In an experiment, the program correctly identifies 98 from 124 of the iPhones and correctly identifies 90 of the 117 Android phones.

For the purposes of this exercise consider the iPhone as the "positive case".

(a) Construct the classification matrix for this experiment.

(b) Calculate the sensitivity and specificity of the program.

**Exercise 4 Bayes' Theorem**

Show, using the notation on slides 24-25, that

$$P(B) = P(B|A)P(A) + P(B|\overline{A})P(\overline{A}).$$

This formula is called the "law of total probability". It is sufficient to show that the two forms of Bayes' Theorem are equivalent.

**Exercise 5 Prosecutor's fallacy**

The police in a large city of (1 million inhabitants) arrest a man for theft. A DNA test shows a positive match with a DNA sample taken at a murder crime scene which so far remains unsolved. There is no other evidence linking the thief with the murder case. This type of DNA matching claims that: if he was at the crime scene the probability of a positive test result is is 1 (certain); if he was not at the crime scene the probability of a positive test result is is $10^{-5}$ (one in a hundred-thousand).

The thief is appears in court being tried for the murder case. The prosecutor claims "the DNA test shows that probability this man is innocent is one in a hundred-thousand"

(a) What is wrong with the prosecutor's probability statement?

(b) Use Bayes' theorem to calculate the probability that this thief is the murderer.

**Exercise 6 Classificatzion in R (logistic regression)**

The `Diabetes` data set in Moodle is part of a larger data set collected by the *US National Center for Health Statistics*. We will fit a logistic regression model to the variable `YN`, which takes the values "Yes" for patients with Diabetes and "No" otherwise. This dataset contains two explanatory variables `BMI` (Body mass index) and `Age`.

First of all, to get to know the data, read in the `Diabetes` data set and obtain the following summary statistics.

(a) How many observations are there?

(b) obtain the frequency table for diabetes status.

(c) What is the mean and standard deviation of `BMI` and `Age`?

(d) Plot a histogram of `BMI` and `Age`, and a scatter plot of the two.

(e) Create a box plot of `BMI` against `YN` and `BMI` against `YN`.

The command in R to fit a logistic regression for diabetes status dependent on BMI is

```
glm.obj<-glm(YN~BMI,data=train,family="binomial")
```

The term `family="binomial"` is required because there are different types of generalised linear models (GLMs) all fitted using the `glm()` function and the logistic regression model uses theory based on the binomial distribution. Just as an example, another possibility is modelling count data ($Y \in \{0, 1, 2, 3, \ldots\}$), in which case `family="poisson"`.

The R script file `DiabetesLogReg.R` is a template to analyse the logistic regression for `YN~BMI`.
An outline of the code is given below.
► Once you have worked through this model, copy and edit the code to fit a model `YN` depending on `Age`.
► Once you have done this, fit another model with `YN` depending on `BMI` and `Age`.
► Once you have calculated the sensitivity, specificity, ROC Curve and AUC for all 3 models, find the equivalent statistics using the test data and compare them with the results for the training data.

To plot the ROC curve you will need to install the package `ROCR` pronounced "Rocker"! [1] The syntax is quite complicated but the code given in `DiabetesLogReg.R` works for many different classification examples with just the obvious editing of variable names.

Outline of the template code:

- Split the data into a training and test data set, with 2000 ($\approx 20\%$) observations in the test data set.
- Fit the logistic regression model and look at the model summary.
- Define "High Risk of Diabetes" using a cut off of $\alpha = 0.5$. Obtain the classification matrix
- Compute the sensitivity and specificity.
- Plot the ROC curve.
- Calculate the AUC.

**Exercise 7 Optional: Challenger Catastrophe**
The following is a nice true-life example and the small data set makes it easier to understand the ideas behind fitting a logistic regression model. I recommend this exercise for those not taking the Ma Data Science *regression* course or if you feel you need some revision of the basic principles of logistic regression.

On 28th January 1985 the space Shuttle Challenger exploded soon after take off, killing all the astronauts on board. An investigation by NASA was carried out to find the cause of the explosion. Before the launch, the on-site engineers had no reason to consider the weather conditions as particularly unusual or dangerous and so gave clearance for the launch.

[1] Website: `https://rocr.bioinf.mpi-sb.mpg.de/` Max Plank inst. Informatik.

The following investigation found that the problem lay with the rubber O-Rings, fitted between sections of the rocket casing. In particular the physical properties of O-rings were sensitive to the temperature of the launch site.

In the experiment, the temperature was controlled for 23 O-Rings, and their flexibility measured. The O-Rings that were too brittle are recorded as "failure". These data are now used as a standard example for analyses using logistic regression.

The raw data are:

```
Temp <- c(66, 70, 69, 68, 67, 72, 73, 70, 57, 63, 70, 78, 67,
          53, 67, 75, 70, 81, 76, 79, 75, 76, 58)
Failure <- c(0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0,
             0, 0, 0, 0, 1, 0, 1)
```

Use R to work through the exercise.

(a) Plot the points `Failure` ($y$-axis) against `Temp` in a scatter plot.

(b) Calculate the overall failure rate of the O-rings. Hint `table(?)` and `prop.table(?)`

(c) Calculate the failure rate in the following temperature intervals $[52, 57), [57, 62), \ldots$ (`cut(?,right=F)`)

(d) As $\text{logit}(0) = -\infty$ and $\text{logit}(1) = -\infty$, adjust any zero or one frequencies to 0.01 and 0.99 respectively. Calculate the logits for the the frequencies in each temperature group:

$$\text{logit}(p) = \log(p/(1-p))$$

(e) Plot the estimated frequencies against the temperature mid-points.

(f) Plot the logit frequencies against the temperature mid-points.

(g) Fit a regression line with the „logits" dependent on Temperature.

(h) Add the regression line to the plot in part (f).

(i) Repeat the plot in part (e) and add the predicted curve using `curve(?,add=T)`. You will need to calculate your predicted values by defining an R function for the inverse logit function. This converts the linear predictor values line onto a probability scale.

$$\pi = \frac{e^{\gamma}}{1 + e^{\gamma}}$$

(j) Use this function to predict the probability of a failure when the temperature is 61 degrees Fahrenheit (16 Celsius).

(k) The temperature when the Challenger launched was 36 degrees Fahrenheit. Estimate the probability that an O-Ring fails at this temperature. Comment on the fitted probability of failure at this temperature.

(l) A logistic regression does not assign the temperatures into intervals but considers temperature as a continuous variable. Fit a logistic regression in R using:

```
summary(fit) <- glm(Failure~Temp, family=binomial))
```

(m) Plot a graph of predictor function $f(x)$ against Temperature. Hint see `predict` in the Diabetes sample code.