MATH3823 - Solutions to Chapter 5 Exercises

Exercise 5.1

a. Logistic. A volunteer has, or does not have, a negative reaction and the number of volunteers having negative reactions for each dose can be counted. It is likely that volunteers react independently. This is a classical situation for a logistic model, with a binomial response, and is referred to as a dose-response study.

b. Geometric. The response variable is the number of transaction before the first fraudulent transaction is identified. Each transaction is fraudulent or not, and transactions should be independent. This is a standard setting for the geometric. There is no explanatory variables here and so no regression model is appropriate.

c. Poisson. The response is the number of phone calls per day and there is no explanatory variable. It is not clear if the phone calls regarding an internet failure will be independent as a single failure might create many calls. This may mean that a Poisson model does not fit well, but it is the first to try.

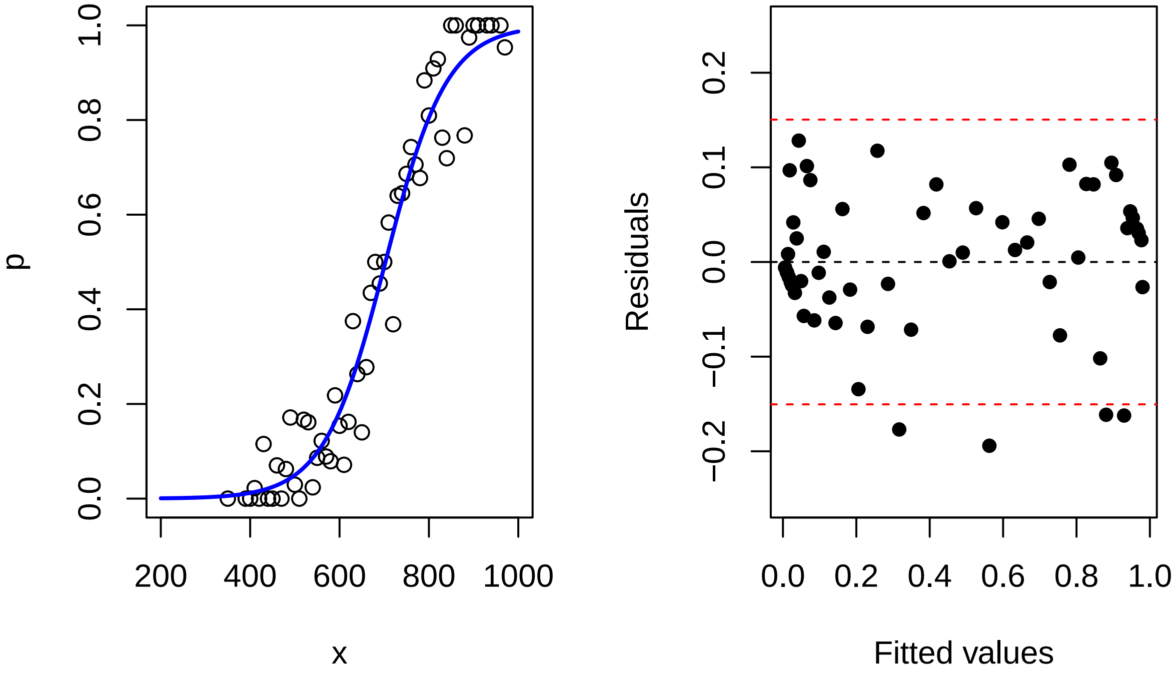
d. Logistic. Students passing their exams or not is a Bernoulli trial. It is reasonable to think that students pass, or not, independently and the response is the number of students passing. The explanatory variable is the application score which might help predict the success of students.

Exercise 5.2 For each of the following situations follow the appropriate steps previously discussed for linear or logistic model fitting.

a. The response variable is the number of successful loan applications where the number of loan applications is known (the data contains the number successful and the number declined). This is a standard binomial modelling situation where the probability of an application being successful depends on the credit score. The number of successful loan applications is the response, Y, and the credit score is the explanatory variable, x. The model is then $Y \sim \text{Bin}(m, p)$ where $p = \exp(\alpha + \beta x)/(1 + \exp(\alpha + \beta x))$ where parameters α and β are to be determined from the data.

```
#################################################
# First load the data
loans = read.csv(file="https://rgaykroyd.github.io/MATH3823/Datasets/loans.csv")

# Calculate the number of Bernoulli trials
m = loans$success+loans$declined

# Get plot area ready as a 1 x 2 grid
par(mfrow=c(1,2))

# Plot data (perhaps re-label axes and change scales)
plot(loans$score, loans$success/m,
```

```r
     xlim=c(200,1000), ylim=c(0,1),
     xlab="x", ylab="p")

# Define local variable names -- needed for x in prediction step, but easier throughout. Needed for resp
x = loans$score
y = cbind(loans$success, loans$declined)

# Fit logistic by using family=binomial
model = glm(y ~ x, family = "binomial")

# Evaluate value on fitted model for many points
xnew = seq(200,1000, length.out=100)
fitted = predict(model, newdata = data.frame(x=xnew), type="response")

# Add fitted curve to scatter plot
lines(xnew,fitted, lwd=2, col="blue")

# Scatter plot of residual & fitted values
residuals = resid(model, type="response")
plot(model$fitted.values, residuals, pch=16,
     xlab="Fitted values", ylab="Residuals", ylim=c(-0.25,0.25))
abline(h=0, lty=2)

r.sd = sd(residuals)
abline(h= -2*r.sd, lty=2, col="red")
abline(h=  2*r.sd, lty=2, col="red")
```

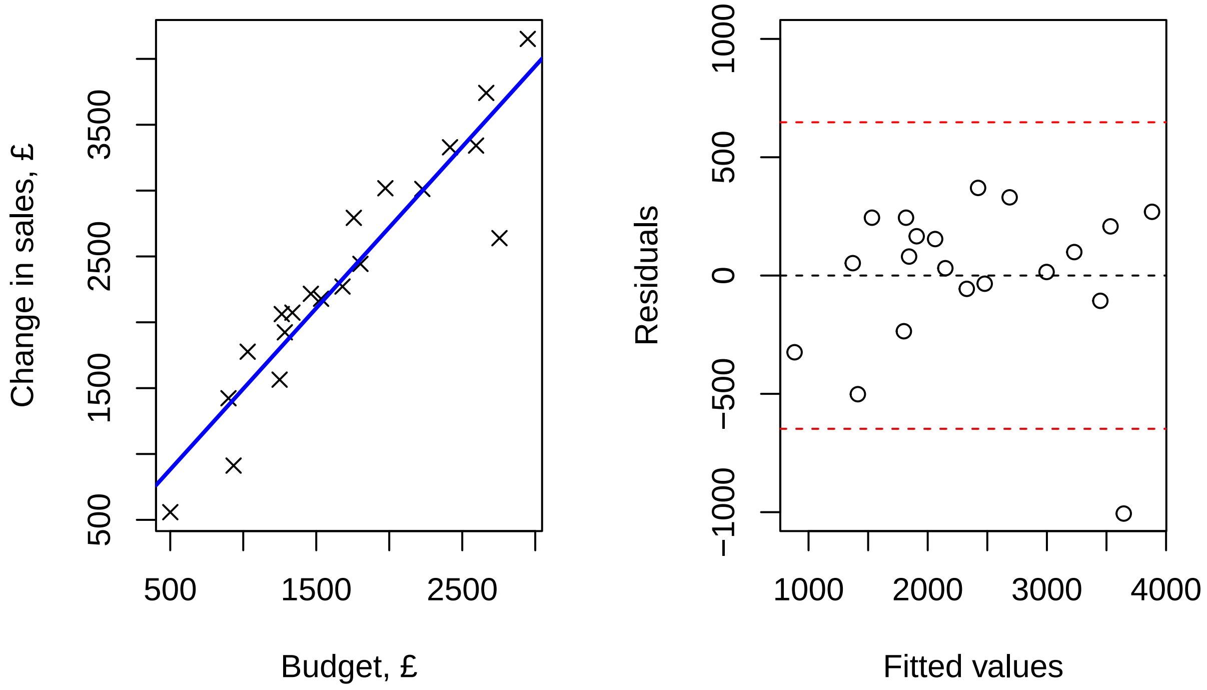


The logistic model is a good fit with the fitted curve following the data points well and there is no substantial pattern in the residuals. Do not worry about a few points outside the 2 standard deviation interval, as it is only there for *guidance*. The fitted line, however, looks a very fit and hence a logistic model should be adequate to describe the relatinship.

b. The description of the problem does not indicate that a binomial response is appropriate and hence does not indicate a logistic regression. Instead, it seems that both the response, change in sales, and the explanatory, advertising budget, are continuous measurements. It would be reasonable to consider a linear model.

```
##################################################
# First load the data
advertising = read.csv(file="https://rgaykroyd.github.io/MATH3823/Datasets/advertising.csv")

# Get plot area ready as a 1 x 2 grid
par(mfrow=c(1,2))

# Plot data (perhaps relabel axes and change scales)
plot(advertising$budget, advertising$sales, pch=4,
     xlab="Budget, £", ylab="Change in sales, £")

# Define local variable names -- needed for x in prediction step, but easier throughout. Needed for res
x = advertising$budget
y = advertising$sales

# Fit linear model
```

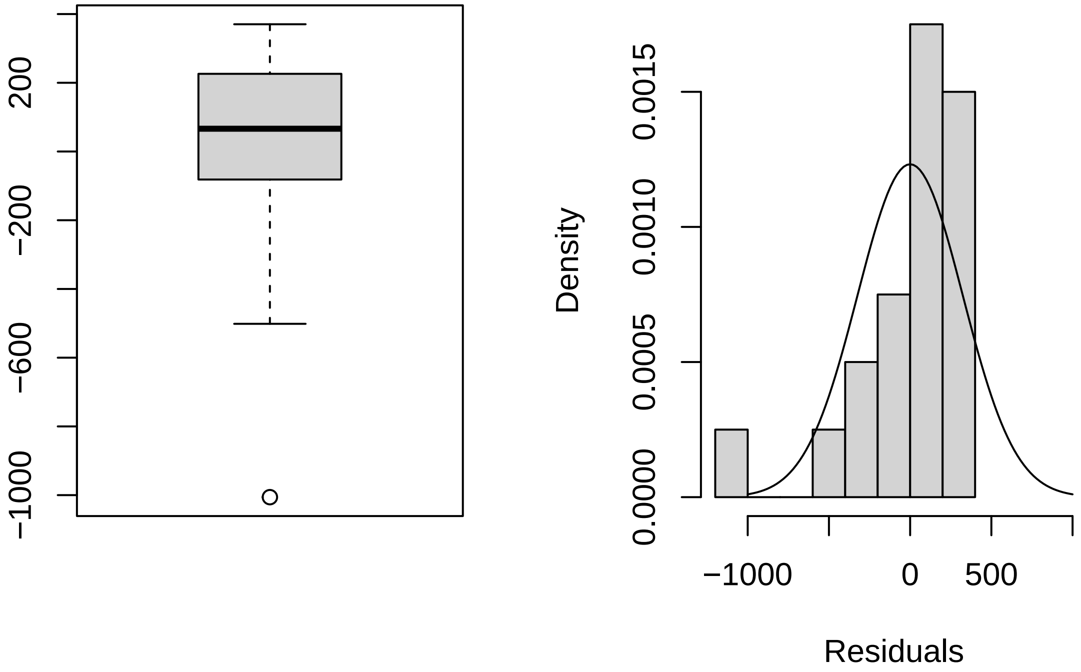
```r
model = lm(y ~ x)

# Add fitted curve to scatter plot
abline(model, lwd=2, col="blue")

# Scatter plot of residual & fitted values
plot(model$fitted.values, model$residuals,
     xlab="Fitted values", ylab="Residuals",
     ylim=c(-1000,1000))
abline(h=0, lty=2)

r.sd = sd(model$residuals)
abline(h= -2*r.sd, lty=2, col="red")
abline(h=  2*r.sd, lty=2, col="red")
```



The fitted line describes the data well, although there is a point at the far right which is substantially below the line. This can also clearly be seen in the residual plot.

```r
# Get plot area ready as a 1 x 2 grid
par(mfrow=c(1,2))

boxplot(model$residuals)
hist(model$residuals, probability = T, main="",
     xlab="Residuals", xlim=c(-1200,1000))

x.mean = mean(model$residuals)
```

```
x.sd = sd(model$residuals)
curve(dnorm(x,x.mean,x.sd), -1000, 1000, add=T)
```



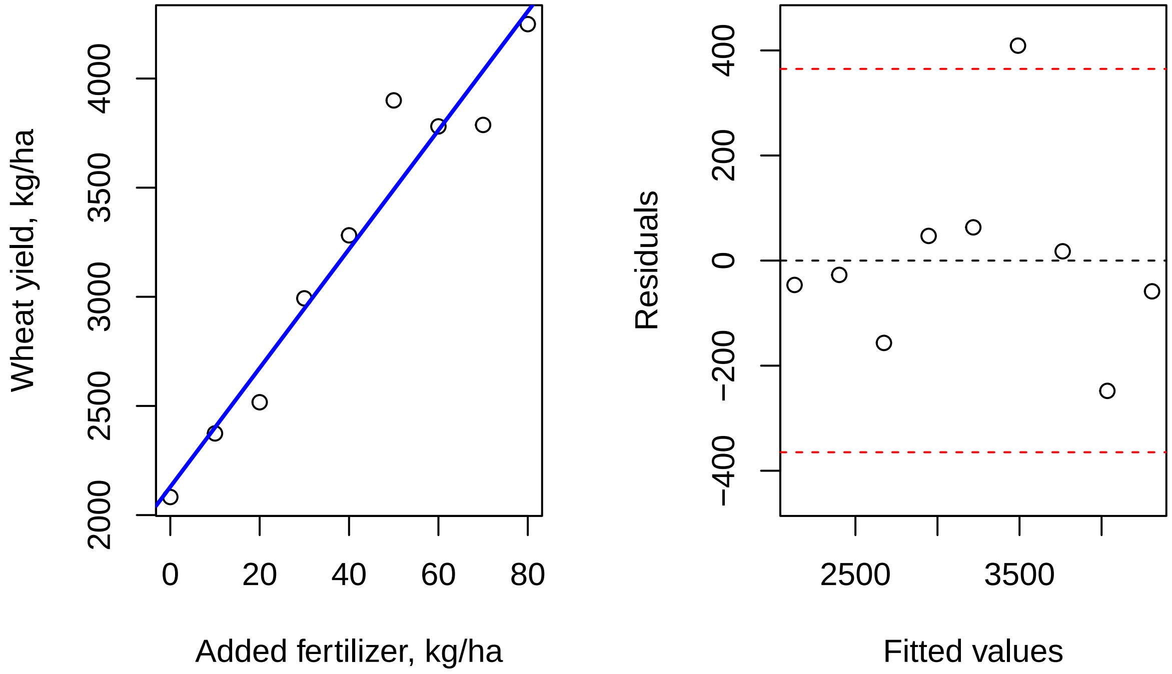
The boxplot also shows this extreme negative residual. A (density-scaled) histogram shows the negative skew nature of the residuals, which isn't consistent with a normal distribution. A normal density is added to the histogram and does not fit the distribution of residuals at all well. It is reasonable to say that the model assumptions do not hole and that an alternative model should be considered.

Remark: From the modelling situation and the general shape of the data, then it could be that using a logistic equation would be worth trying – you will see elsewhere that the generalised linear model framework can be used in such situations.

c. Again, the content does not suggest that a logistic regression with a binomial response variable is appropriate. Hence consider a linear model.

```
##################################################
# First load the data
fertilizer = read.csv(file="https://rgaykroyd.github.io/MATH3823/Datasets/wheat.csv")

# Get plot area ready as a 1 x 2 grid
par(mfrow=c(1,2))

# Plot data (perhaps relabel axes and change scales)
plot(fertilizer$P205, fertilizer$yield,
     xlab="Added fertilizer, kg/ha", ylab="Wheat yield, kg/ha")
```

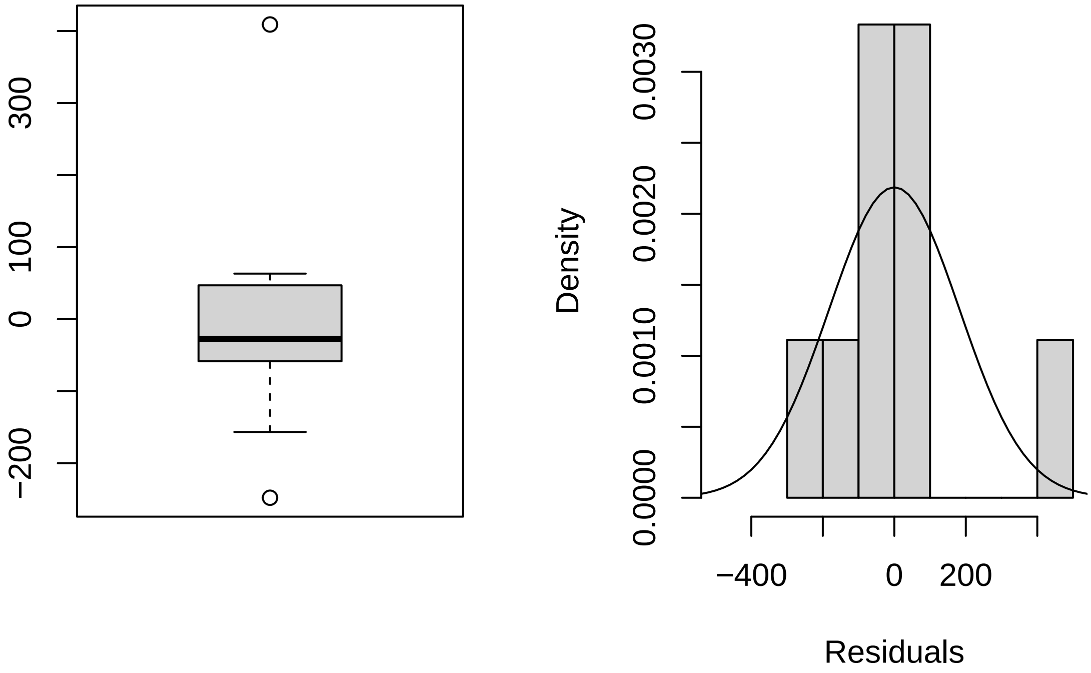
```
# Define local variable names -- not required  but easier throughout
x = fertilizer$P205
y = fertilizer$yield

# Fit linear model
model = lm(y ~ x)

# Add fitted curve to scatter plot
abline(model, lwd=2, col="blue")

# Scatter plot of residual & fitted values
plot(model$fitted.values, model$residuals,
     xlab="Fitted values", ylab="Residuals",
     ylim=c(-450,450))
abline(h=0, lty=2)

r.sd = sd(model$residuals)
abline(h= -2*r.sd, lty=2, col="red")
abline(h=  2*r.sd, lty=2, col="red")
```



The linear model fits the data well with only one slightly extreme residual. There is a slight curved pattern in the residuals, but the sample size is very small.

```
# Get plot area ready as a 1 x 2 grid
par(mfrow=c(1,2))
```

```r
boxplot(model$residuals)
hist(model$residuals, probability = T, main="",
     xlab="Residuals", xlim=c(-500,500))

x.mean = mean(model$residuals)
x.sd = sd(model$residuals)
curve(dnorm(x,x.mean,x.sd), -1000, 1000, add=T)
```



The boxplot and histogram suggest that the normal distribution assumption about the residuals may not be valid. Again another type of model could be tried – see the remark after he previous question.

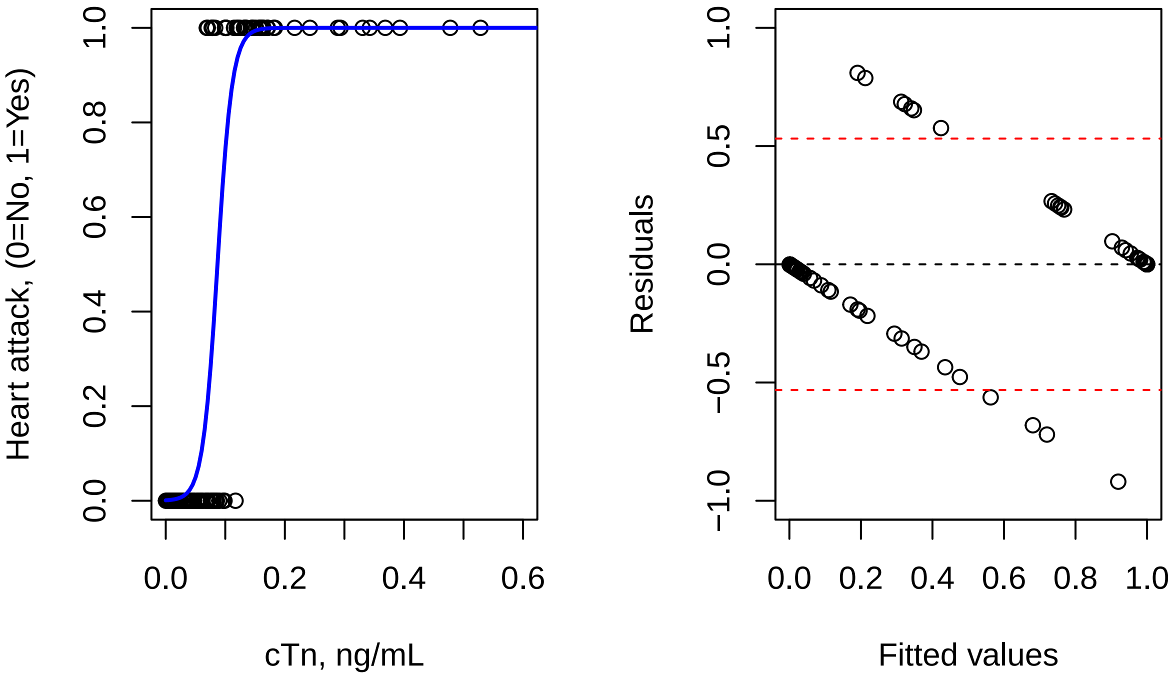
d. This is a less obvious example of logistic regression but clearly each patient is either diagnosed as having had a heart attack or not – a single Bernoulli trial. Recall that the Bernoulli is a special case of the binomial distribution with $m = 1$ and hence we can use the logistic model with the probability parameter depending on the cTn value.

```r
##################################################
# First load the data
health = read.csv(file="https://rgaykroyd.github.io/MATH3823/Datasets/heart.csv")

# Set the number of trials
m=1

# Get plot area ready as a 1 x 2 grid
par(mfrow=c(1,2))
```

```r
# Plot data (perhaps relabel axes and change scales)
plot(health$cTn, health$diagnosis,
     xlim=c(0,0.6), ylim=c(0,1),
     xlab="cTn, ng/mL", ylab="Heart attack, (0=No, 1=Yes)")

# Define local variable names as easier throughout
x = health$cTn
y = cbind(health$diagnosis, m-health$diagnosis)

# Fit logistic by using family=binomial
model = glm(y ~ x, family = "binomial")
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
# Evaluate value on fitted model for many points
xnew = seq(0,1, length.out=200)
fitted = predict(model, newdata = data.frame(x=xnew), type="response")

# Add fitted curve to scatter plot
lines(xnew,fitted, lwd=2, col="blue")

# Scatter plot of residual vs fitted values
residuals = resid(model, type="response")
plot(model$fitted.values, residuals,
     xlab="Fitted values", ylab="Residuals", ylim=c(-1,1))
abline(h=0, lty=2)

r.sd = sd(residuals)
abline(h= -2*r.sd, lty=2, col="red")
abline(h=  2*r.sd, lty=2, col="red")
```



The scatter plot with the fitted curve should suggest that the residual plot may not be as expected and you will see a *warning* message part-way through the when you perform the calculations. With all data values being 0 or 1 gives a very odd appearance to the residual plot. In such situation gives greater weight to the fitted curve than the residual plot – though you will see how to create a better residual plot for these situations elsewhere. The fitted curve describes the data well and so a logistic model is appropriate for this data set.

Exercise 5.3 Use the fitted logistic regression model to predict what dose of gaseous carbon disulphide would kill 90% of beetles.}

The model connects probability of death p to dose x via

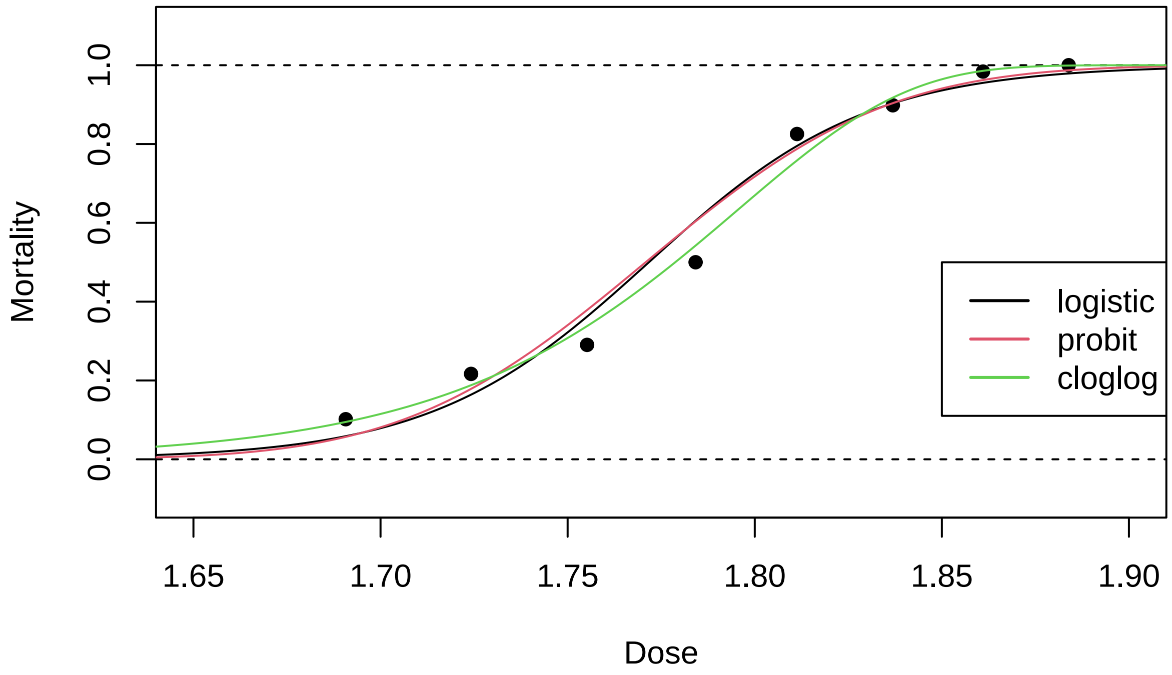
$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \eta(x) = \beta_0 + \beta_1 x.$$

Hence the required dose for any p_0 is $x(p) = (\text{logit}(p) - \beta_0)/\beta_1$. Substituting the given $p = 0.9$ and fitted $\hat{\beta}_0 = -60.717$ and $\hat{\beta}_1 = 34.270$, we obtain

$$x(0.9) = (\log(0.9/(1-0.9)) + 60.717)/34.270 = 1.83584.$$

```
beetle = read.table("https://rgaykroyd.github.io/MATH3823/Datasets/beetle.txt", header=T)

dose = beetle$dose
mortality = beetle$died/beetle$total
```

```r
plot(dose, mortality, pch=16,
     xlim=c(1.65, 1.90), xlab ="Dose",
     ylim=c(-0.1, 1.1),  ylab="Mortality")
abline(h=c(0,1), lty=2)

y = cbind(beetle$died, beetle$total-beetle$died)
output.dose = seq(1.6,1.95,0.001)

# logistic link
glm.fit = glm(y ~ dose, family=binomial(link='logit'))
fitted = predict(glm.fit, data.frame(dose=output.dose), type="response")
lines(output.dose, fitted)

# probit link
glm.fit = glm(y ~ dose, family=binomial(link='probit'))
fitted = predict(glm.fit, data.frame(dose=output.dose), type="response")
lines(output.dose, fitted, col=2)

# Complementary log-log link
glm.fit = glm(y ~ dose, family=binomial(link='cloglog'))
fitted = predict(glm.fit, data.frame(dose=output.dose), type="response")
lines(output.dose, fitted, col=3)

legend(1.85, 0.5, legend=c("logistic","probit","cloglog"), col=c(1,2,3), lty=c(1,1,1), lwd=1.5)
```



The choice of link function appears to make very little difference in this example as all three fitted curves are very close together.

Exercise 5.4

- We will have $\phi = 1$ if and only if $\{\pi_1/(1-\pi_1)\}/\{\pi_2/(1-\pi_2)\} = 1$ iff $\pi_1/(1-\pi_1) = \pi_2/(1-\pi_2)$. Since the "odds' ' map $\pi \to \pi/(1-\pi)$ is monotone from $(0,1)$ to $(0,\infty)$ (check its derivative is everywhere positive), this last inequality holds iff $\pi_1 > \pi_2$.

- The log odds ratio log ϕ_j can be written as

  $$\begin{aligned} \log\ \phi_j &= \operatorname{logit}\pi_{1j} - \operatorname{logit}\pi_{2j} \\ &= \alpha_1 - \alpha_2 + (\beta_1 - \beta_2)x_j, \quad j = 1, ..., m, \end{aligned}$$

  which is constant in j iff $\beta_1 - \beta_2 = 0$.

  The parameters α_i and β_i describe the intercept and slope of a logistic regression of π_{ij} on x_j. Thus the log odds ratio also varies linearly with x_j, with slope $\alpha_1 - \alpha_2$ and intercept $\beta_1 - \beta_2$.

- Here is an example where this model might be plausible. Suppose a group of workers has been exposed to a dangerous pollutant, and this group has been matched to an unexposed control group. Further, suppose the smoking habits of each worker are recorded and given a score $x = 0, 1, 2$ (for none, medium, high). The workers are followed up for 40 years and the numbers dying of lung cancer are recorded. Let π_{ij} denote the probability of dying from lung cancer for workers in group i with smoking level j, $i = 1, 2$, $j = 0, 1, 2$. Then, with risk measured in terms of logit π_{ij},

  - α_1 and α_2 represent the risk to nonsmokers in the exposed and unexposed groups. We might expect $\alpha_1 > \alpha_2$.
  - β_1 and β_2 represent the increased risk as x increases by 1 for each group. We expect $\beta_1, \beta_2 > 0$.

  If the risks of smoking act "independently' ' from those of exposure, we expect $\beta_1 = \beta_2$. But if there is a biological interaction between pollution and smoking, we might expect $\beta_1 > \beta_2$.

- Let `count` be a $2m \times 2$ matrix containing the numbers of diseased and non-diseased people for each combination of i and j. Let `status` be a 2-level factor giving the group information i and let `x` be a quantitative variable storing the x_j information. This model can be fitted by the command `glm(count ~ status*x, binomial)`. \end{enumerate}

Exercise 5.5 For the *Low dose* group the probability of death is $65/237 = 0.2742616$ and $226/244 = 0.9262295$ for the *High dose* group – a very big difference. Then, the risk of death due to High dose exposure relative to Low dose is $0.9262295/0.2742616 = 3.377175$. This value means that the risk of death is more than 3 times more likely in the *High dose* group than in the *Low dose* group. Clearly, if there is no difference, then relative risk would be around 1 – random variation would prevent a value of exactly 1 – and hence a value substantially different to 1 indicates an association and an extreme value would indicate a strong association.

The odds of death in the *Low dose* group is $65/172 = 0.377907$ – less likely to die than not die – and $226/18 = 12.55556$ for the *High dose* group. The second odds is much greater than the first, that is the odds are much greater in the *High dose* group than the *Low dose* group.

These various *scales*, probability, relative risk and odds, all measure the same basic properties and each is used in some situations and hence it is important that we can move between them as the application changes.

End of Solutions to Chapter 5 Exercises