Generalized Linear Models

The statistical analyses that we have considered so far are suitable only for continuous response variables that have a linear relationship with the explanatory variable(s). For example, a simple linear regression model, is of the form

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

where β_0 represents the intercept, β_1 represents the slope of the line and e_i represents the error term. Another way of writing this is to define a linear relationship between the expected value of the response variable y_i, (which is μ_i) and the explanatory variable x, along with the variance

$$E[y_i] = \mu_i = \beta_0 + \beta_1 x_i$$

$$Var[y_i] = \sigma^2$$

It is assumed that:

- the error terms follow a normal distribution $e_i \sim N(0, \sigma^2)$
- the variance of the error terms is constant
- the relationship between x_i and y_i is linear

The linear model can be extended to include more than one explanatory variable, in which case it is a multiple regression model:

$$E[y_i] = \mu = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}$$

where β_0 represents the intercept, β_j represents the strength of the linear relationship between the explanatory variable x_{ij} and the response variable y_i. The set of models referred to as **general linear models** take the form:

$$E[y] = \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

where the explanatory variables x_{ij} can be continuous or categorical. It is possible to have continuous and categorical explanatory variables in the same model (ANCOVA).

All general linear models assume that:

- **the errors have a normal distribution,** $e_i \sim N(0, \sigma^2)$
- **the variance of the error terms is constant**
- y_i **is a linear combination of the** x_{ij}

In practice, many response variables violate these assumptions, for example:

- binary response variables where the response is one of two possible categories (yes/no, present/absent, dead/alive, etc.)
- nominal response variables where there are more than two possible categories (yes/no/don't know/not applicable)
- count data (always positive)
- proportions (bounded between 0 and 1)
- time to an event

The class of models known as **generalized linear models** allows for response variables that:

- have error distributions other than the normal distribution
- non constant variance
- a nonlinear relationship between the explanatory variable(s) and the response variable .

A generalized linear model relates the expected value of the response variable y_i (which is μ_i) to a value predicted by a linear combination of the explanatory variables $x_{ij,}$:

$$g(E[y_i]) = g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}$$

The linear combination of explanatory variables is known as the **linear predictor** and denoted by η i.e.

$$\eta = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}$$

The non-linear function that defines the relationship between μ_i and η is known as the **link function,** denoted by g. Using this notation, we may write the form of a generalized linear model as:

$$g(\mu_i) = \eta$$

or equivalently

$$\mu_i = g^{-1}(\eta)$$

Summary of general linear models (vs) generalized linear models:

	General LM	Generalized LM
Distribution of y_i	$y_i \sim N(\mu_i, \sigma^2)$	$y_i \sim$exponential family
Linear predictor	$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$	$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$
Link function	$\mu_i = \eta$	$g(\mu_i) = \eta$

Many of the properties of the Normal distribution are shared by a wider class of distributions called the **exponential family of distributions**. Three common distributions belonging to the exponential family are the Normal distribution, the Poisson distribution, and the Binomial distribution.

Error Distribution	Data type	Link function
Normal	Continuous, Symmetric	Identity
Poisson	Count	Log
Binomial	Binary/Proportion	Logit

Logistic Regression Preliminaries

The Binomial Distribution Recall, a Bernoulli trial is a procedure which has two possible results – success or failure. The probability of success is called p and the probability of failure is then $(1-p)$. The binomial distribution is used to find probabilities associated with a fixed number of repeated Bernoulli trials. The binomial distribution is used when:

1. each trial is repeated a fixed number of times
2. each trial can result in either success or failure
3. there is a fixed probability of success which does not vary from one trial to the other

The third condition above means that the outcome of one trial does not have any effect on any subsequent trial. So for example if a fair coin is tossed five times, the outcome of the first four tosses has no influence on the fifth toss. Even if the first four tosses have come up heads there is still a 50-50 chance of getting a head on the fifth toss.

Definition 1 *The formula for calculating the probability of r successes out of n trials is*

$$P(r,n) = C_r^n p^r (1-p)^{n-r}$$

where p is the probability of success for each trial and $(1-p)$ is the probability of failure for each trial.

The probability mass function for a variable with a binomial distribution where $n = 20$ and $p = 0.5$ is shown below:

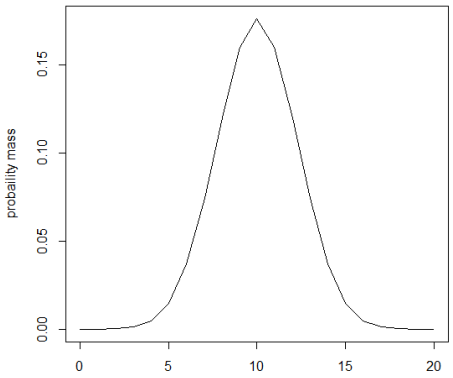

Figure 1

If y is a binomially distributed random variable i.e. $y \sim Bin(n, p)$ then

$$\begin{aligned} E[y] &= np \\ Var[y] &= np(1-p) \end{aligned}$$

Note that the variance of y is **not** constant with respect to the mean.
If the mean changes so does the variance, as shown in Figure 2.

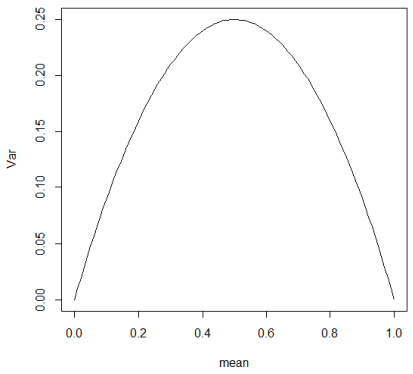

Figure 2

The Exponential Function

The exponential function is the function e^x (often written as $\exp(x)$) where e is the number (2.718281828....). The domain of e^x is all real numbers $(-\infty, \infty)$ and the range is all the real numbers greater than 0 $(0, \infty)$ as shown in Figure 3(a) below. The function e^x is its own derivative.
The natural logarithm, $\ln(x)$ is the inverse function of e^x so that $\ln(e^x) = x$. The domain of $\ln(x)$ is all real numbers greater than 0, $(0, \infty)$ and the range is all real numbers $(-\infty, \infty)$ as shown in Figure 3(b) below.

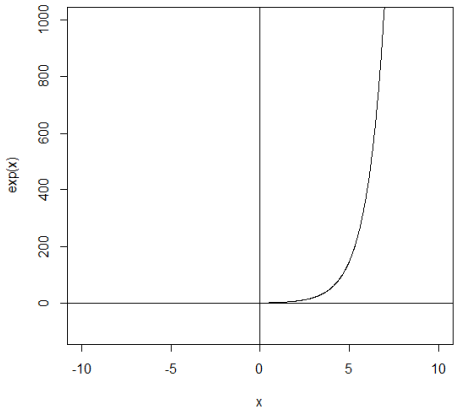

Figure 3(a)

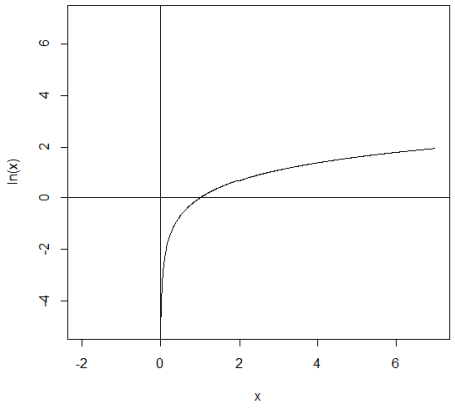

Figure 3(b)

Logistic Regression

Linear regression is used to model outcomes for variables with a normal distribution, for example, it can be used to predict a person's height based on their weight. Logistic regression is used to model outcomes for binary variables, for example:

- will a patient survive an illness based on current health and treatment?
- will a customer pay back a loan based on their income and credit rating?
- will it rain tomorrow based on current weather conditions?

In logistic regression, the response variable y_i $(i = 1...N)$ is defined to be the outcome of n_i Bernoulli trials, coded as "0" or "1". For example the outcome that there is rain tomorrow is coded as "1" and the outcome that there is no rain tomorrow is coded as "0". If $n_i = 1$, then the response variable y_i has a binary outcome

$$y_i = \begin{cases} 0 \text{ with probability } 1 - p_i \\ 1 \text{ with probability } \quad p_i \end{cases}$$

If there are n_i Bernoulli trials then the response variable follows a binomial distribution, i.e. $y_i \sim Bin\left(n_i, p_i\right)$ where

$$\begin{aligned} E\left[y_i\right] &= n_i p_i \\ Var\left[y_i\right] &= n_i p_i \left(1 - p_i\right) \end{aligned}$$

Suppose we have a binary response variable y_i and k explanatory variables $x_{i1}, x_{i2}, ..., x_{ik}$, we are interested in modeling the probability that the outcome of y_i is "1":

$$E[y_i] = p_i = P(y_i = 1)$$

The logistic regression model is given by:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

sometimes it is written as:

$$\log it(p_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

We do not model p_i directly as a linear combination of the explanatory variables, instead we model the function $\log\left(\frac{p_i}{1-p_i}\right)$as a linear combination of explanatory variables. In terms of generalized, linear models, the link function g is of the form $g(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$.

Why don't we model p_i directly?

The linear predictor $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$ can take on any value in the interval $(-\infty, \infty)$. If we model p_i directly using a linear predictor $p_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$, it is possible that p_i can take on values outside the range $[0, 1]$ which is undesirable since p_i is a probability (see Fig 4(a))

We wish to map any value of the linear predictor in the range $(-\infty, \infty)$ to a value in the range $[0, 1]$.

If we model $\log\left(\frac{p_i}{1-p_i}\right)$ as a linear combination of the explanatory variables then, as we show below, we restrict the range of values that p_i can take on to the interval $[0, 1]$.
The function $\exp(x)$ will take any number between $-\infty$ and ∞ and map it to a number between 0 and ∞. Therefore if we write:

$$p_i = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k),$$

we restrict the range of values that p_i can take on to the interval $[0, \infty)$ (see Fig.4(b)).
Finally, if we write

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)}{(1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k))}$$

then we restrict the range of values that p_i can take on to the interval $[0, 1]$ (see Fig.4(c)).

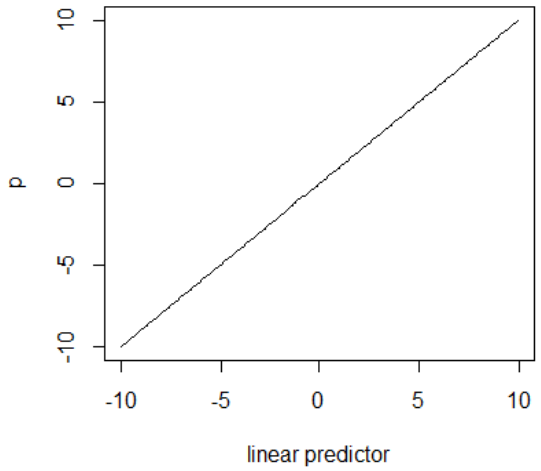

Figure 4(a)

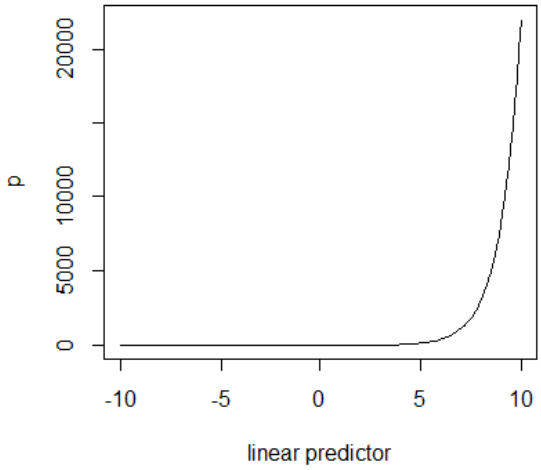

Figure 4(b)

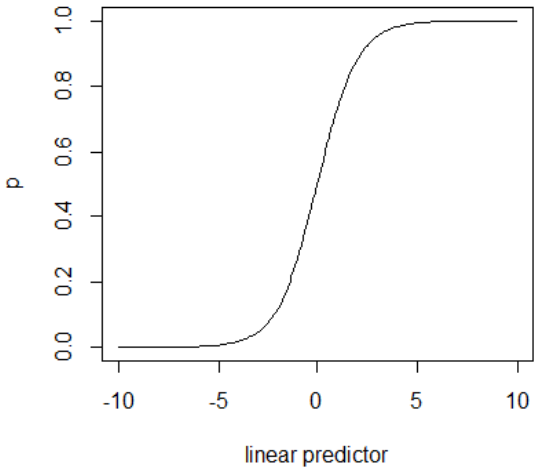

Figure 4(c)

The equation

$$p_i = \frac{\exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k\right)}{\left(1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)\right)}$$

describes the logistic curve (see Fig 4(c)) and is often rearranged to give:

$$\log\left(\frac{p_i}{(1 - p_i)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

Note that for p_i values around 0.5, the rate of change in probabilities is larger than for p_i values close to 0.9 and 0.1.

Odds

The ratio $\frac{p_i}{1-p_i}$ represents the ratio of the probability that the event $y_i = 1$ will occur divided by the probability that the event $y_i = 1$ will not not occur and is known as the **odds** in favour of the event.
For example if the probability that a certain team will win a match is given as $\frac{2}{3}$ then the odds in favour of that team winning are:

$$\frac{\frac{2}{3}}{1-\frac{2}{3}} = \frac{\frac{2}{3}}{\frac{1}{3}} = 2$$

Odds are usually written as $2:1$. Logistic regression models the *log of the odds* as a linear combination of the explanatory variables.

Logistic regression with a binary response variable and a single continuous explanatory variable

If we have just one continuous explanatory variable x_1 then the logistic regression model will be of the form:

$$\log\left(\frac{p_i}{(1-p_i)}\right) = \beta_0 + \beta_1 x_1$$

Example 1 In January 1986, the space shuttle Challenger exploded shortly after launch. An investigation was launched into the cause of the crash and attention focused on the rubber O-ring seals in the rocket boosters. At lower temperatures, rubber becomes more brittle and is less effective as a sealant. At the time of the launch, the temperature was 31^oF. Could the failure of the O-rings have been predicted? In the 23 previous shuttle missions for which data exists, some evidence of damage was recorded on some O-rings. Each shuttle has two boosters, each with three O-rings. For each mission, we know the launch temperature and the corresponding number of O-rings out of six showing damage.
Our response variable y_i has two outcomes; the event $y_i = 1$ represents the outcome that there is damage to at least one of the six O-rings and the event $y_i = 0$ represents the event that there is no damage on any of the six O-rings.

$$y_i = \begin{cases} 0 \text{ with probability } 1-p_i \\ 1 \text{ with probability } \quad p_i \end{cases}$$

The explanatory variable, x_i, is the temperature in ^oF at the time of launch, table 1 below shows the data from the 23 shuttle missions prior to the Challenger launch.

Table 1

Temperature ^oF (x_i)	Damage present (y_i)
53	1
57	1
58	1
63	1
66	0
67	0
67	0
67	0
68	0
69	0
70	1
70	0
70	1
70	0
72	0
73	0
75	0
75	1
76	0
76	0
79	0
79	0
81	0

The data from Table 1 is shown in Fig. 5 below.

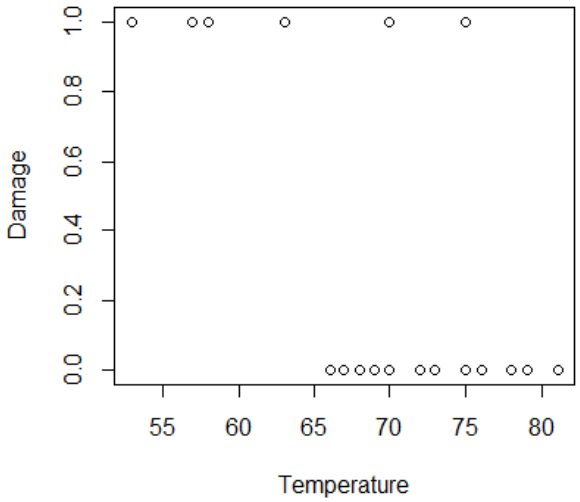

Figure 5

It seems that O-ring damage is more likely to occur at low temperatures. If we try to fit a linear model to this data (i.e. a model of the form $y_i = \beta_0 + \beta_1 x_i + e_i$), we run into problems as shown in Fig. 6.

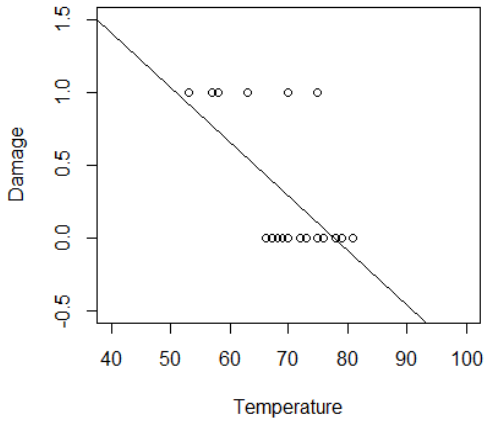

Figure 6

It is clear that the linear model predicts outcomes outside of the interval $[0, 1]$.

Instead, we will use the logistic model $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x$ which fits the data well and does not predict outcomes outside the interval $[0, 1]$ (see Figure 7).

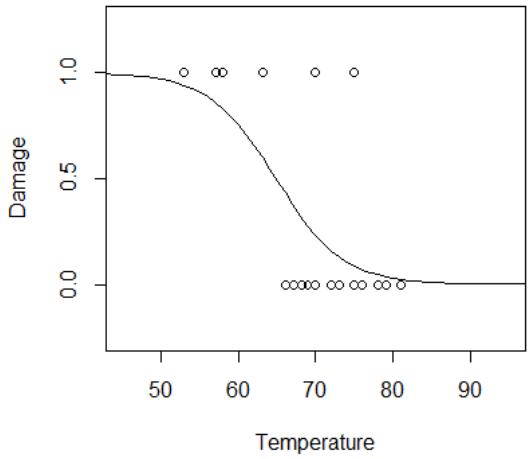

Figure 7

Figure 7 shows that as the temperature drops below $50^o F$ the probability of damage occurring to at least one of the O-rings approaches 1. At $65^0 F$ the probability of damage occurring to at least one of the O-rings is approximately 0.5 and as the temperature increases above $80^o F$ the probability of damage occurring to at least one of the O-rings approaches 0.

Note that for temperatures between $60^0 F$ and $80^0 F$, the rate of change in the probability of damage occurring to at least one of the O-rings is large compared with samples with smaller and larger temperatures i.e. at extreme temperature values, a change in temperature has less effect on the probability compared with average temperatures.

The model was fitted in R by estimating the values of β_0 and β_1 using a method that minimises the error between the observations and the fitted values. Part of the model summary is shown in Table 2 below.

Table 2

	Estimate	Std Error	z value	$Pr\,(>z)$
Intercept (β_0)	15.04	7.37	2.04	0.04
Temperature (β_1)	-0.23	0.11	-2.15	0.03

Using the estimates for β_0 and $\beta_{1,}$ the model is:

$$\begin{aligned} \log\left(\frac{p_i}{(1-p_i)}\right) &= 15.04 - 0.23x \\ \frac{p_i}{(1-p_i)} &= e^{15.04-0.23x} \\ \frac{p_i}{(1-p_i)} &= e^{15.04}e^{-0.23x} \end{aligned}$$

So, the relationship between the odds of O-ring failure and temperature is modelled in a non-linear way. If the regression parameter β_1 is estimated as 0, then the exponentiated value ($e^{\beta} = e^0$) is 1, and has no effect on the odds. Therefore, a logistic regression parameter β that is zero has no effect on the odds. A positive regression parameter corresponds to an increase in the odds and a negative value to a decrease.

How to interpret the model The negative value of the coefficient β_1 indicates that the probability of O-ring damage was greater at low temperatures. A unit increase in temperature will **decrease** the **log odds** of O-ring damage by 0.23. Alternatively we can say that a unit increase in temperature will **decrease** the **odds** of O-ring damage by $e^{(0.23)} = 1.2586$.

More generally:

- a unit increase in the predictor variable, x, will change the **log odds** by β_1
- a unit increase in the predictor variable, x, will change the **odds** by a factor of e^{β_1}

In general, we cannot comment on the effect of unit increase in temperature on the **probability** of O-ring failure since the rate of change in the probability of O-ring failure depends on the value of the temperature. We can only comment on the effect of unit increase in temperature on the **probability** of O-ring failure at fixed temperatures.
For example, we can calculate the change in the probability of O-ring failure when the temperature changes from 60^0F to 61^0F as follows:

When the temperature is 60^0F, the probability of O-ring failure is:

$$\begin{aligned}
\log\left(\frac{p_i}{(1-p_i)}\right) &= 15.04 - 0.23x \\
\log\left(\frac{p_i}{(1-p_i)}\right) &= 15.04 - 0.23 \times 60 \\
\log\left(\frac{p_i}{(1-p_i)}\right) &= 1.24 \\
\frac{p_i}{(1-p_i)} &= e^{1.24} \\
\frac{p_i}{(1-p_i)} &= 3.4556 \\
p_i &= 3.4556\,(1-p_i) \\
p_i &= 3.4556 - 3.4556p_i \\
p_i + 3.4556p_i &= 3.4556 \\
p_i(1+3.4556) &= 3.4556 \\
p_i &= \frac{3.4556}{(1+3.4556)} = 0.77556
\end{aligned}$$

Similarly, when the temperature is 60^0F, the probability of O-ring failure is:

$$\begin{aligned}
\log\left(\frac{p_i}{(1-p_i)}\right) &= 15.04 - 0.23x \\
\log\left(\frac{p_i}{(1-p_i)}\right) &= 15.04 - 0.23 \times 61 \\
\log\left(\frac{p_i}{(1-p_i)}\right) &= 1.01 \\
\frac{p_i}{(1-p_i)} &= e^{1.01} \\
\frac{p_i}{(1-p_i)} &= 2.7456 \\
p_i &= \frac{2.7456}{(1+2.7456)} = 0.73302
\end{aligned}$$

So if the temperature changes from 60^0F to 61^0F then the probability of O-ring failure changes from 0.77556 to 0.73302 i.e. it decreases by 0.0425.

Exercise

If the temperature changes from 70^0F to 71^0F,what is the change in the probability of O-ring failure?