University of Glasgow

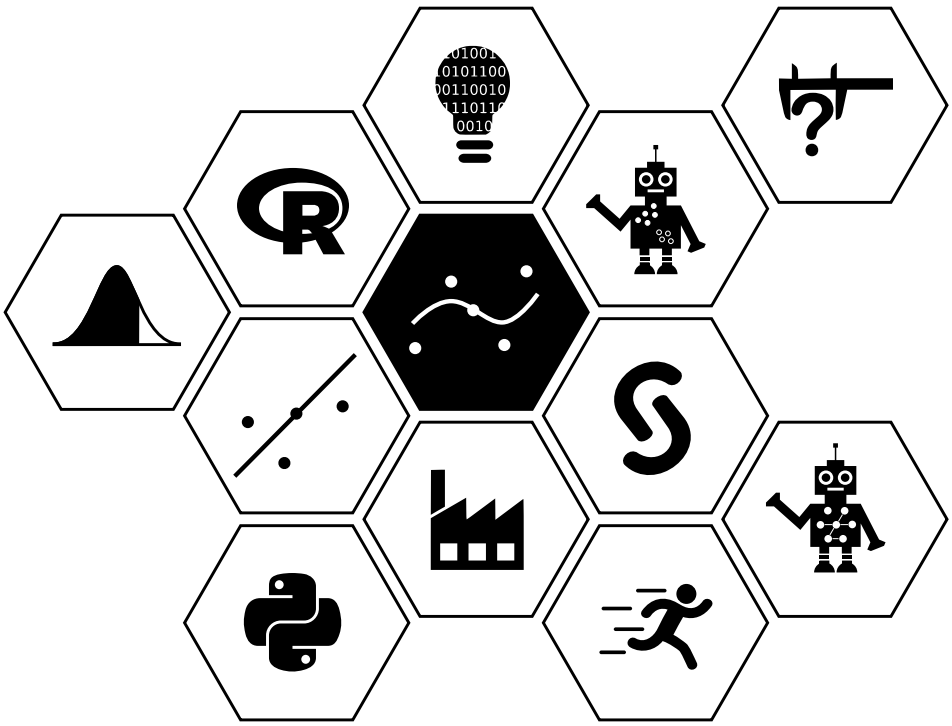
Advanced Predictive Models

Tereza Neocleous

Academic Year 2021-22

Week 2:

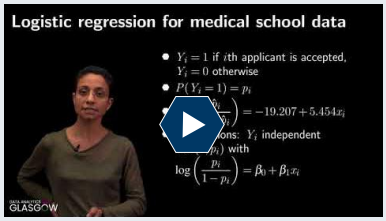
Introduction to generalised linear models



DATA ANALYTICS GLASGOW

Introduction to Generalised Linear Models

Introduction and motivating examples



Introduction to GLMs

https://youtu.be/5u1w6eROypI

Duration: 9m57s

In this course we will extend the theory of linear regression models to non-normal responses. Before we begin with introducing the class of models known as Generalised Linear Models (GLMs), we will briefly illustrate why linear models are not sufficient for all types of data. Throughout the course, we will see how we can deal with a variety of situations where the linear model may not be adequate.

The main objective of this week's learning material is to introduce Generalised Linear Models (GLMs), which extend the linear model framework to outcome variables that don't follow the normal distribution. GLMs can be used to model non-normal continuous outcome variables, but they are most frequently used to model **binary**, **categorical** or **count data**. We will focus on these latter types of outcome variables. To see why extensions to the normal linear model are needed, let's look at a couple of examples, one where the normal linear model is appropriate and one where it's not.

Example 1 (Bollywood box office revenue).

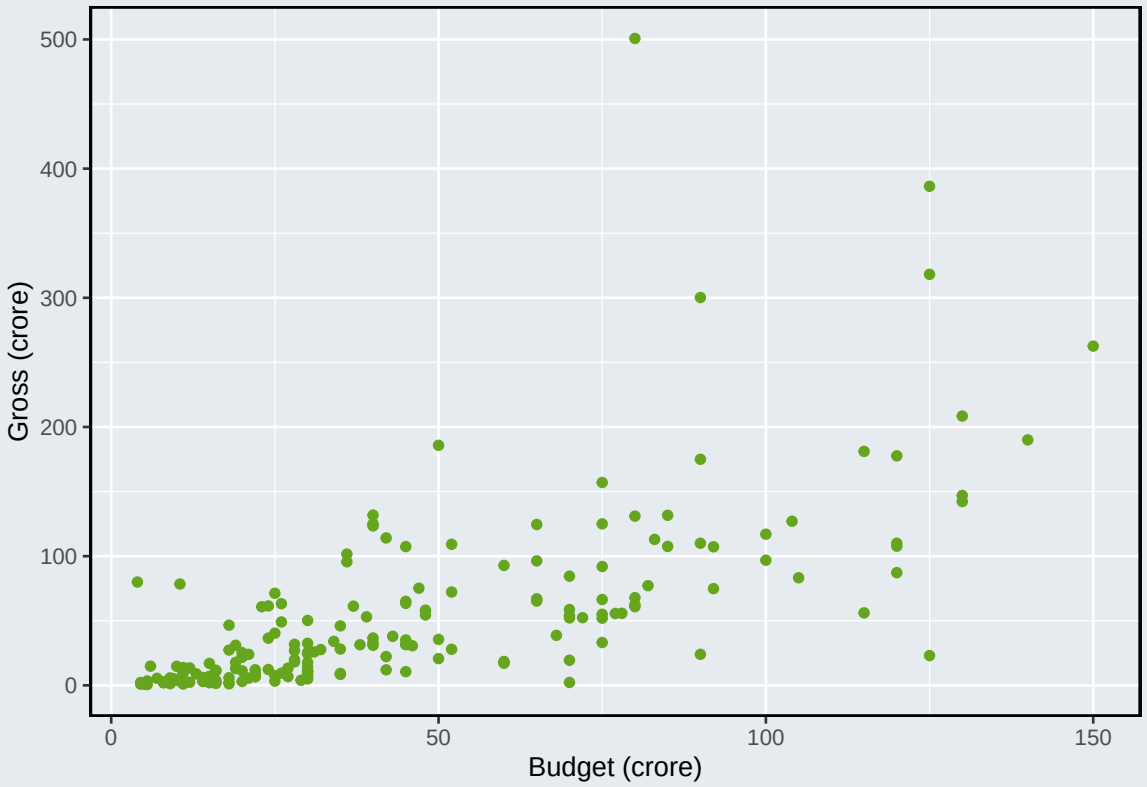
Possibly the simplest scenario of a predictive model is when we want to predict an outcome variable based on a predictor which displays a linear relationship to the variable of interest. Consider the following dataset on Bollywood film revenues (sourced from http://www.bollymoviereviewz.com](http://www.bollymoviereviewz.com)) which contains data on 190 films made during the period 2013-2017. We would like to predict the gross revenue of a film from the film's budget. Both the gross revenue and the budget are measured in crore. Here are the first few rows of the data:

```
bollywood <-
  read.csv(url("http://www.stats.gla.ac.uk/~tereza/rp/bollywood_boxoffice.csv"))
head(bollywood)
```

Movie	Gross	Budget
Ek Villain	95.64	36.0
Humshakals	55.65	77.0
Holiday	110.01	90.0
Fugly	11.16	16.0
City Lights	5.19	9.5
Kuku Mathur Ki Jhand Ho Gayi	2.23	4.5

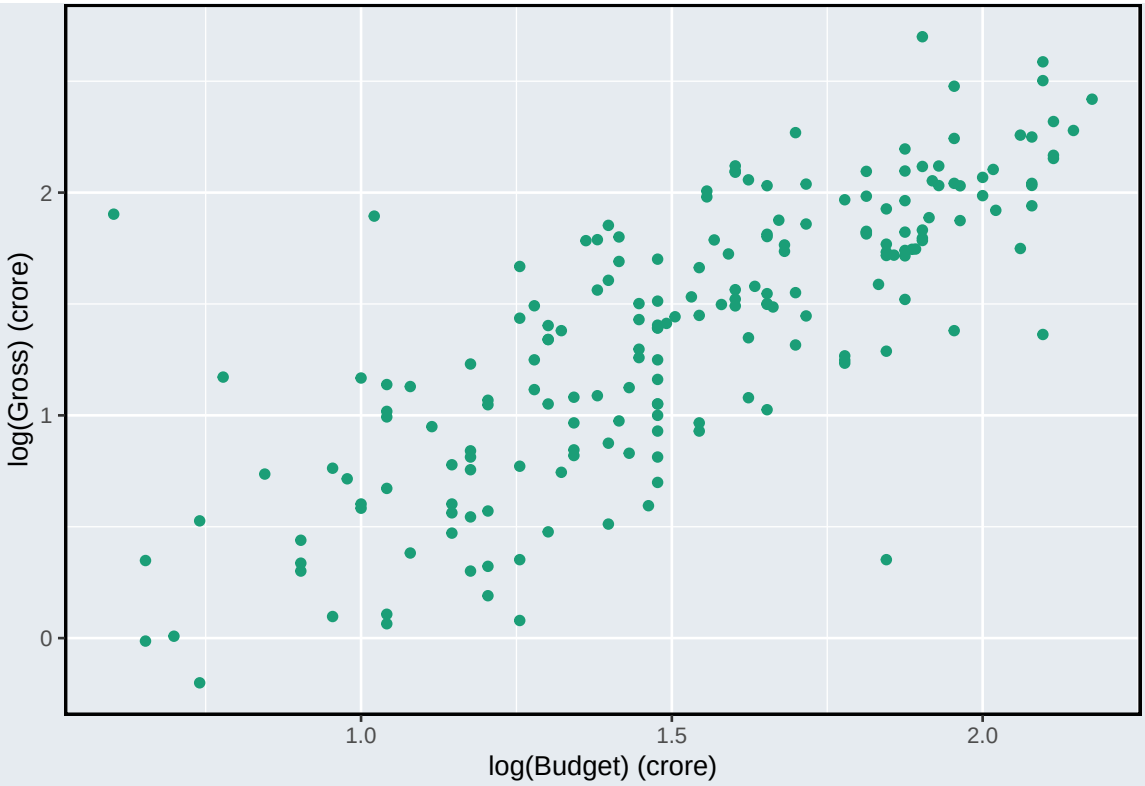
We can plot the gross revenue against the budget to explore the relationship between the two variables.

```
b.plot <- ggplot(data = bollywood, aes(y = Gross, x = Budget)) +
          geom_point(col = "#66a61e") +
          scale_x_continuous("Budget (crore)") + scale_y_continuous("Gross (crore)")
```



Looking at the scale of the values on both the horizontal and vertical axes, we might want to transform the data by taking logs.

```
b.plot.l <- ggplot(data = bollywood, aes(y = log10(Gross), x = log10(Budget))) +
            geom_point(col = "#1b9e77") +
            scale_x_continuous("log(Budget) (crore)") +
            scale_y_continuous("log(Gross) (crore)")
```



Now let's fit a model with the $\log_{10}$ transformed gross revenue as the response (Y_i) and the $\log_{10}$ transformed budget (x_i) as the explanatory/predictor variable. We can use the `lm()` function to fit this linear model in R.

```
bol.lm <- lm(log10(Gross) ~ log10(Budget), data = bollywood)
```

The model equation in mathematical notation is

$$E(Y_i) = \mu_i = \beta_0 + \beta_1 x_i; \quad \text{where the } Y_i \text{ are independent } N(\mu_i, \sigma^2), \quad i = 1, ..., 190$$

The model fit is shown below:

```
summary(bol.lm)
```

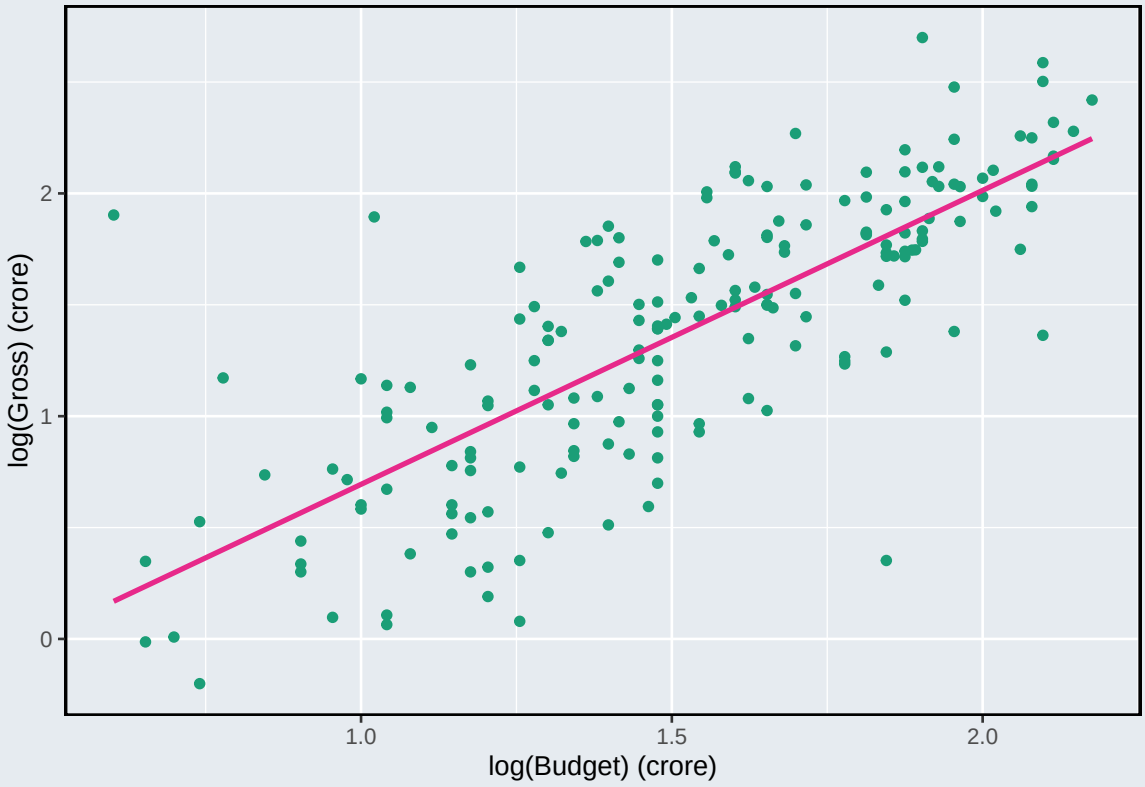
```
Call:
lm(formula = log10(Gross) ~ log10(Budget), data = bollywood)

Residuals:
     Min       1Q   Median       3Q      Max
-1.45702 -0.24470  0.00807  0.24600  1.73413

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.62549    0.12338  -5.069 9.51e-07 ***
log10(Budget)  1.31955    0.07887  16.730  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3921 on 188 degrees of freedom
Multiple R-squared:  0.5982,    Adjusted R-squared:  0.5961
F-statistic: 279.9 on 1 and 188 DF,  p-value: < 2.2e-16
```

We can visualise this regression model by plotting the data and fitted regression line:

```
b.plot.lm <- ggplot(data = bollywood, aes(y = log10(Gross), x = log10(Budget))) +
          geom_point(col ="#1b9e77") +
          scale_x_continuous("log(Budget) (crore)") +
```

```
        scale_y_continuous("log(Gross) (crore)") +
        geom_smooth(method = lm, colour="#e7298a", se=FALSE)
```

```
`geom_smooth()` using formula 'y ~ x'
```



Task 1.

Using the fitted model equation, predict the gross revenue for a film with a budget of (i) 10, (ii) 50, and (iii) 100 crore.

Hint: Remember that the variables have been log-transformed.

Example 2 (GPA and admission to medical school).

Now let's look at a different kind of dataset, where the outcome we want to predict is not continuous-valued but binary. This is a dataset on admissions to US medical schools, which gives the admission status, GPA and standardised test scores for 55 medical school applicants from a liberal arts college in the US Midwest and it can be loaded from the `Stat2Data` package in R.

```
library(Stat2Data)
data(MedGPA)
```

The first few rows of the data are given below.

Accept	Acceptance	Sex	BCPM	GPA	VR	PS	WS	BS	MCAT	Apps
D	0	F	3.59	3.62	11	9	9	9	38	5
A	1	M	3.75	3.84	12	13	8	12	45	3
A	1	F	3.24	3.23	9	10	5	9	33	19
A	1	F	3.74	3.69	12	11	7	10	40	5
A	1	F	3.53	3.38	9	11	4	11	35	11
A	1	M	3.59	3.72	10	9	7	10	36	5

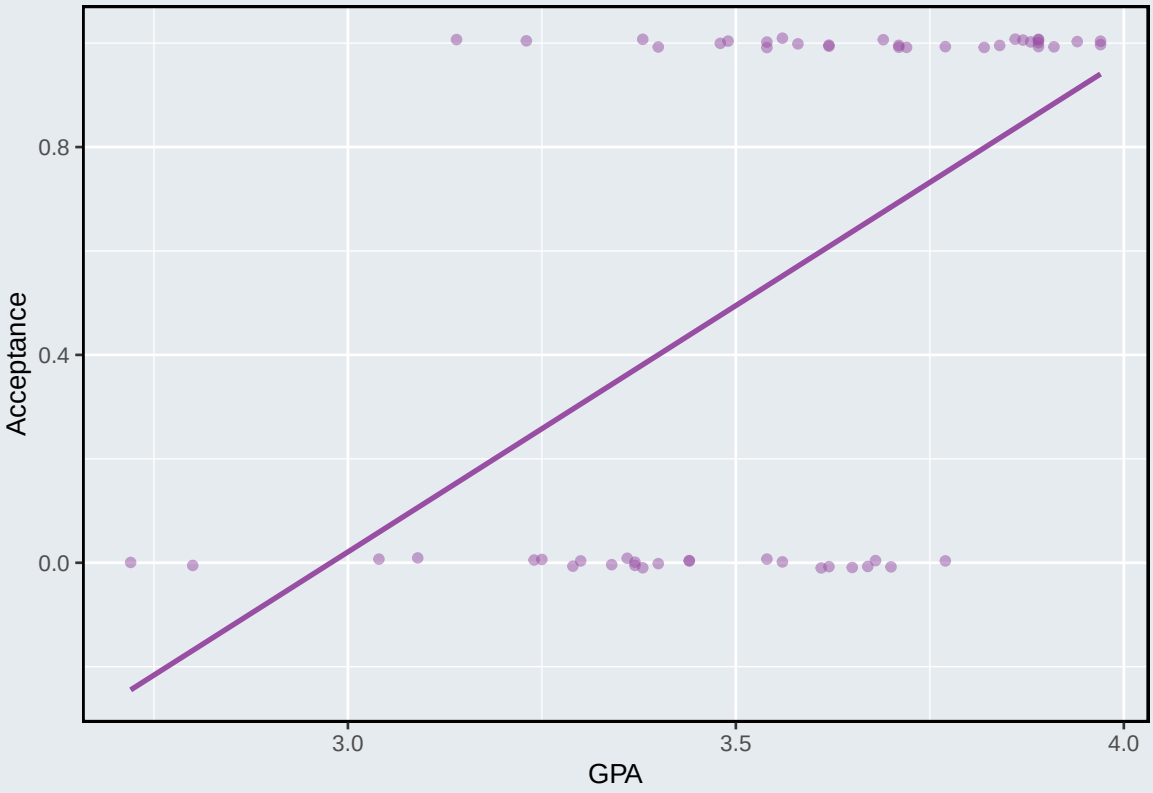
Let us look at a plot of acceptance against GPA, adding a bit of jitter to make overlapping points more visible.

```
medgpa.plot <- ggplot(data = MedGPA, aes(y = Acceptance, x = GPA)) +
               geom_jitter(width =0, height =0.01, alpha =0.5, colour ="#984ea3")
```

We can add the linear regression line for `Acceptance` as a function of `GPA` to the plot.

```
medgpa.plot + geom_smooth(method = "lm", se = FALSE,
                          fullrange = TRUE, colour = "#984ea3")
```

```
`geom_smooth()` using formula 'y ~ x'
```



The R code for fitting the model and the model output is shown below.

```
med.lm <- lm(Acceptance ~ GPA, data=MedGPA)
summary(med.lm)
```

```
Call:
lm(formula = Acceptance ~ GPA, data = MedGPA)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7510 -0.3717  0.1352  0.3059  0.8464

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.8240     0.7226  -3.908 0.000266 ***
GPA           0.9483     0.2027   4.678 2.04e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4267 on 53 degrees of freedom
Multiple R-squared:  0.2922,    Adjusted R-squared:  0.2788
F-statistic: 21.88 on 1 and 53 DF,  p-value: 2.043e-05
```

In mathematical notation, the value of the independent response Y_i is equal to 1 if the ith applicant is accepted and $Y_i = 0$ otherwise, where x_i refers to to the ith applicant's college GPA for $i = 1, \ldots, 55$. The

normal linear model assumes that the Y_i are independent $N(\mu_i, \sigma^2)$ with $\mu_i = \beta_0 + \beta_1 x_i$, with the fitted model equation given by $\hat{\mu}_i = -2.8240 + 0.9483 x_i$.

One issue with this fit is that the predicted values of the response can take any real values, while acceptance can only take the value 0 or 1. And it is hard to argue that a variable taking values of 0 or 1 is normally distributed. Instead, we can use a logistic regression model for the *probability* of acceptance. Let's first write it down in mathematical notation by letting $p_i = P(Y_i = 1)$ denote the probability of acceptance for the ith applicant. We assume that the Y_i are independent random variables which follow the Bin$(1, p_i)$ (or Bernoulli(p_i)) distribution with

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

This is equivalent to:

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}.$$

(Solve the first equation for p_i to verify this!)

We can fit this model in R using the `glm()` function:

```
med.glm <- glm(Acceptance ~ GPA, data = MedGPA, family = binomial)
```

The argument `family=binomial` specifies that `Acceptance` follows a binomial distribution, with probability of success p_i (i.e. the probability of the ith applicant being accepted is p_i). In addition, the probability p_i is a function of x_i (i.e. the probability of the ith applicant being accepted is a function of that applicant's GPA). The default link function, corresponding to the logit link $\log\left(\frac{p_i}{1-p_i}\right)$, is used here. That is, `family = binomial` implies `family = binomial(link="logit")`.

The model fit is shown below.

```
summary(med.glm)
```

```
Call:
glm(formula = Acceptance ~ GPA, family = binomial, data = MedGPA)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7805  -0.8522   0.4407   0.7819   2.0967

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -19.207      5.629  -3.412 0.000644 ***
GPA            5.454      1.579   3.454 0.000553 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 75.791  on 54  degrees of freedom
Residual deviance: 56.839  on 53  degrees of freedom
AIC: 60.839

Number of Fisher Scoring iterations: 4
```

The regression equation for the fitted model is

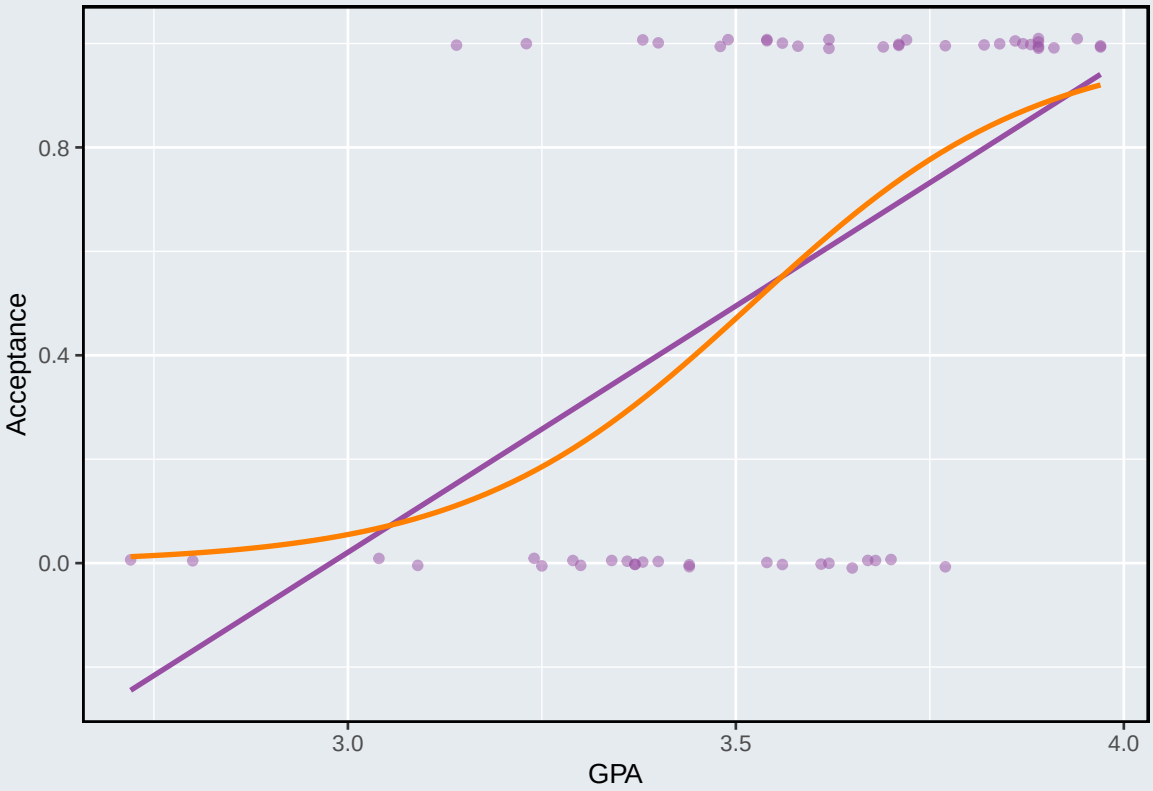
$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = -19.207 + 5.454 x_i,$$

or equivalently

$$\hat{p}_i = \frac{\exp(-19.207 + 5.454x_i)}{1 + \exp(-19.207 + 5.454x_i)}.$$

The fitted curve for the probability of acceptance is shown in orange below.

```
`geom_smooth()` using formula 'y ~ x'
`geom_smooth()` using formula 'y ~ x'
```



We can see that that this curve fits the data better than the linear regression line, and that it gives predicted probabilities between 0 and 1, as desired. We could add predictors to the model to improve predictive performance – we'll see more about that later.

The regression equation we have obtained allows us to predict the acceptance probability for a given GPA.

Task 2.

Predict the acceptance probability for an applicant with a GPA of (i) 2.5, (ii) 3 (iii) 4. First do this "by hand" using the regression equation, then in R using the `predict()` function.

Hint: The `predict()` *function will return values on the linear predictor scale unless you specify* `type='response'` *which returns probabilities instead.*

https://goo.gl/mZHxyN

- Section 2.3 from *Mixed effects models and extensions in ecology with R -Zuur et al.* discusses the appropriateness of the assumptions of the linear model.

What do the Bollywood box office and medical school admission examples have in common?

In both cases we have independent observations and we want to predict an outcome of interest (gross revenue/acceptance) based on an explanatory variable (budget/GPA). In both cases we have a regression equation allowing us to predict the response from a given value of the predictor. However, in one case the response is assumed

to follow the normal distribution, in the other the binomial distribution. In both cases we fit a model to the *mean* of the response: in the normal linear model the mean $E(Y_i) = \mu_i$ is assumed to be a linear function of x_i: $\mu_i = \beta_0 + \beta_1 x_i$, and in the logistic regression model the mean $\mu_i = E(Y_i) = p_i$ is modelled through the *logit link function*. That is, in logistic regression $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i$. In slightly more general notation we have $g(\mu_i) = \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}$ where $\mu_i = E(Y_i)$ and $g(\mu_i)$ for each distribution is given in the following table.

Model	Random component	Systematic component	Link function
Normal model	$y_i \overset{\text{indep}}{\sim} N(\mu_i, \sigma^2)$, $E(Y_i) = \mu_i$	$\boldsymbol{x}_i^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_i$	Identity link $g(\mu_i) = \mu_i$
Logistic regression model	$y_i \overset{\text{indep}}{\sim} Bin(1, p_i)$, $E(Y_i) = p_i$	$\boldsymbol{x}_i^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_i$	Logit link: $g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \log\left(\frac{p_i}{1-p_i}\right)$

Exponential family of distributions

It turns out that the normal and binomial distributions also have something else in common: they are both members of the *exponential family of distributions*. (And so is the Poisson, the negative binomial, gamma distribution and many others.)

Definition 1 (Exponential family of distributions).

Consider a random variable Y whose probability density function (p.d.f.) or probability mass function (p.m.f.) depends on parameter θ. The distribution belongs to the exponential family if it can be written as

$$f(y;\theta) = \exp\left[a(y)b(\theta) + c(\theta) + d(y)\right].$$

*The term $b(\theta)$ is called the **natural parameter**. If $a(y) = y$ the distribution is said to be in **canonical form**.*

Example 3 (Normal distribution is a member of exponential family).

Consider $Y \sim N(\theta, \sigma^2)$ with p.d.f.

$$f(y;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}(y-\theta)^2\right], \qquad -\infty < y < \infty. \tag{1}$$

If we are interested in estimating θ, the variance, σ^2, can be regarded as a nuisance parameter. By rewriting the p.d.f. as

$$f(y;\theta) = \exp\left[-\frac{y^2}{2\sigma^2} + \frac{y\theta}{\sigma^2} - \frac{\theta^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right] \tag{2}$$

we can see that this is of exponential family form with $a(y) = y$ (hence in canonical form) and natural parameter $b(\theta) = \theta/\sigma^2$.

Task 3.

Show that the binomial distribution Bin(n, p) is a member of the exponential family.

The exponential family of distributions has several interesting and useful properties. It can be shown that the expectation and variance for members of the exponential family can be expressed as

$$E[a(Y)] = -\frac{c'(\theta)}{b'(\theta)}$$

and

$$\text{Var}[a(Y)] = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}.$$

Some further useful properties of exponential family distributions relate to the **score function**.

Definition 2 (Score statistic).

*$U = \dfrac{dl(\theta; y)}{d\theta}$ is called the **score statistic**, and is equal to the derivative of the log-likelihood $l(\theta; y)$ with respect to the parameter θ.*

For exponential family distributions with log-likelihood $l(\theta; y) = a(y)b(\theta) + c(\theta) + d(y)$, the score is

$$U(\theta; y) = \frac{dl(\theta; y)}{d\theta} = a(y)b'(\theta) + c'(\theta) \tag{3}$$

We use the score (derivative of the log-likelihood) to solve the likelihood equation $U = \dfrac{dl(\theta; y)}{d\theta} = 0$ in order to obtain the maximum likelihood estimate $\hat{\theta}$ for a number of distributions.

We can think of the score statistic, $U = a(Y)b'(\theta) + c'(\theta)$, as a random variable in its own right, which means we can calculate its expectation

$$\mathrm{E}(U) = b'(\theta)E[a(Y)] + c'(\theta) = b'(\theta)\left[-\frac{c'(\theta)}{b'(\theta)}\right] + c'(\theta) = 0,$$

and its variance

$$\mathrm{Var}(U) = [b'(\theta)^2]\mathrm{Var}[a(Y)].$$

This leads us to a very important concept in statistical inference, called **Fisher information**, which we can use to obtain standard errors for maximum likelihood estimates of GLM coefficients.

Definition 3 (Fisher's information).

*The **Fisher Information**, denoted as $\mathcal{I}$, is given by:*

$$\mathcal{I} = \mathrm{Var}(U) = E(U^2) = E\left[\left(\frac{dl(\theta; y)}{d\theta}\right)^2\right] = E\left[\frac{d^2l(\theta; y)}{d\theta^2}\right]. \tag{4}$$

The variance of the maximum likelihood estimates tells us about the amount of *information* that an observed random variable carries about an unknown parameter in the model that is linked to a distribution.

As we will see shortly, the score and information play a key role in parameter estimation and in obtaining standard errors for the coefficient estimates of a GLM.

Having defined the exponential family of distributions, we are now ready to formally define a GLM.

https://goo.gl/mZHxyN

- Chapter 8 from *Mixed effects models and extensions in ecology with R -Zuur et al.* contains a more in-depth discussion of the exponential family.

Generalised Linear Models

Definition 4 (Generalised Linear Models).

Let Y_i be independent responses from an exponential family distribution in canonical form and $\mu_i = E(Y_i)$, $i = 1, \ldots, n$. A generalised linear model *(GLM) is a model of the form $g(\mu_i) = \boldsymbol{x}_i^\intercal \boldsymbol{\beta}$ where $\boldsymbol{\beta}$ is a p-dimensional parameter vector, $\boldsymbol{x}_i^\intercal$ is the ith row of the design matrix $\boldsymbol{X}$, and $g()$ is a monotonic, differentiable function called the* link function.

A GLM generalises the normal linear model by allowing:

1. a response variable with a distribution other than normal, but a member of the exponential family of distributions; and

2. a relationship between the response and the linear component of the form $g(\mu_i) = \boldsymbol{x}_i^\intercal \boldsymbol{\beta}$ where $g()$ is the *link function*.

Components of a generalised linear model

1. The *random component*: Suppose $Y_1, \ldots, Y_n$ are independent random variables which follow an exponential family distribution such that $f(y_i; \theta_i) = \exp[y_i b(\theta_i) + c(\theta_i) + d(y_i)]$ for $i = 1, \ldots n$. The joint p.d.f. (or p.m.f.) of the Y_i is

$$\begin{aligned} f(y_1, \ldots, Y_n; \theta_1, \ldots, \theta_n) &= \prod_{i=1}^{n} \exp[y_i b(\theta_i) + c(\theta_i) + d(y_i)] \\ &= \exp\left[\sum_{i=1}^{n} y_i b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i)\right] \end{aligned} \tag{5}$$

   The distribution of each Y_i is in canonical form and depends on a single parameter θ_i.

2. The *systematic component*: Associated with each y_i is a vector $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\intercal$ of values of p explanatory variables. The response, Y_i, depends on the explanatory variables through a linear component, $\eta_i = \boldsymbol{x}_i^\intercal \boldsymbol{\beta} = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ for $i = 1, \ldots, n$ where $\boldsymbol{x}_i^\intercal$ is the ith row of the design matrix $\boldsymbol{X}$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\intercal$ is the parameter vector. As in linear models, the design matrix is given by

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^\intercal \\ \vdots \\ \boldsymbol{x}_n^\intercal \end{bmatrix} = \begin{bmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \ldots & x_{np} \end{bmatrix}.$$

3. The *link function*: The parameters θ_i in equation (5) are usually not of direct interest. Instead, we are interested in a smaller set of parameters $(\beta_1, \ldots, \beta_p)$, and assume that Y_i depends on these through the linear predictor η_i. The link between the distribution of the Y_i and the linear predictor η_i is provided by the link function $g()$, for which $g(\mu_i) = \eta_i = \boldsymbol{x}_i^\intercal \boldsymbol{\beta}$. Here $\mu_i = E(Y_i)$ and $g()$ is a monotone, differentiable function. Although any one-to-one function could be used in principle, certain choices of link function can offer great simplification. In particular, the link function can be chosen so that the natural parameter, $b(\theta_i)$, is proportional to the linear component $\eta_i = \boldsymbol{x}_i^\intercal \boldsymbol{\beta}$. Such a link function is known as the *canonical link*. The following table shows the canonical link function for some of the most common distributions.

Distribution	Natural parameter	Canonical link
Normal	$\frac{\theta}{\sigma^2}$	$g(\mu) = \mu$
Poisson	$\log \theta$	$g(\mu) = \log(\mu)$
Binomial	$\log\left(\frac{\theta}{1-\theta}\right)$	$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$

http://encore.lib.gla.ac.uk/iii/encore/record/C__Rb2939999?lang=eng

- Chapter 6 from *Extending linear models with R: generalized linear, mixed effects and nonparametric regression models - Faraway* gives an overview of GLMs and their properties.

https://goo.gl/mZHxyN

- Sections 9.1 and 9.2 from *Mixed effects models and extensions in ecology with R -Zuur et al.* cover the general formulation of GLMs.

Let us now look at some examples of generalised linear models and their components.

Example 4 (Normal linear model).

Consider $E(Y_i) = \mu_i = \boldsymbol{x}_i^\intercal \boldsymbol{\beta}$ where the Y_i are independent $N(\mu_i, \sigma^2)$ for $i = 1, \ldots, n$. Here $g(\mu_i) = \mu_i$, the **identity** link. You may be more familiar with this model written as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$ with the ϵ_i independent, identically distributed $N(0, \sigma^2)$ random variables.

Example 5 (A model for historical linguistics).

If two languages are separated by time t, the probability of having cognate words for a particular meaning can be modelled as $\exp(-\theta t)$. For a test list of n meanings, a linguist judges whether the corresponding words in two languages are cognate or not. For the ith meaning, define

$$Y_i = \begin{cases} 1 & \text{if the two languages have cognate words} \\ 0 & \text{if words are not cognate.} \end{cases}$$

Then $P(Y_i = 1) = \exp(-\theta t) = p$ and $P(Y_i = 0) = 1 - p$ i.e. $Y_i \sim$ Bernoulli(p) (or equivalently Bin$(1, p)$) and $E(Y_i) = p$. The link function is $g(p) = \log p = -\theta t$ so that $g(p)$ is linear in the parameter of interest, θ; also $\boldsymbol{x}_i = -t$; and $\boldsymbol{\beta} = \theta$.

Example 6 (A model for mortality rates).

The number of deaths, Y, in a population can be modelled by the Poisson distribution $f(y; \mu) = \frac{\mu^y e^{-\mu}}{y!}$ where $y = 0, 1, 2, \ldots$ The expected number of deaths per year is $E(Y) = \mu$ and it can be modelled by $E(Y) = \mu = n\lambda(\boldsymbol{x}^\intercal \boldsymbol{\beta})$ where n is the population size and $\lambda(\boldsymbol{x}^\intercal \boldsymbol{\beta})$ is the death rate per 100,000 people per year. Let $Y_1, \ldots, Y_n$ be the numbers of deaths occurring in successive age groups. A possible model is $E(Y_i) = \mu_i = n_i \exp(\theta i)$ where $Y_i \sim$ Poisson(μ_i) and

- $i = 1$ for the age group 30-34,
- $i = 2$ for age group 35-39,
- $\vdots$
- $i = 8$ for age group 65-69.

This can be expressed as a generalised linear model as $g(\mu_i) = \log \mu_i = \log n_i + \theta i$ where $\boldsymbol{x}_i^\intercal = (\log n_i, i)$; and $\boldsymbol{\beta} = (1, \theta)^\intercal$. The term $\log n_i$ is called the *offset*, and we will see more about it when we talk about models for counts.

Task 4.

Formulate the model used in the medical school admissions example as a GLM.

Maximum likelihood estimation of GLM coefficients

In a generalised linear model we are interested in the parameters $\beta_1, \ldots, \beta_p$ that describe how the response depends on the explanatory variables. We use the observed $y_1, \ldots, y_n$ to maximise the log-likelihood function

$$l(\boldsymbol{\beta}, \boldsymbol{y}) = \sum_{i=1}^{n} y_i b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i) \tag{6}$$

obtained from equation (5). This depends on $\boldsymbol{\beta}$ through

$$\begin{aligned} &\mu_i = E(Y_i) = -\frac{c'(\theta_i)}{b'(\theta_i)}; \\ &\mathrm{Var}(Y_i) = [b''(\theta_i)c'(\theta_i) - c''(\theta_i)b'(\theta_i)]/[b'(\theta_i)]^3; \\ &g(\mu_i) = \eta_i = \boldsymbol{x}_i^\intercal \boldsymbol{\beta}, \qquad i = 1, \ldots, n. \end{aligned}$$

The maximisation procedure results in p simultaneous equations for $\hat{\boldsymbol{\beta}}$, which are usually solved numerically using the **method of scoring** (also known as **Fisher's scoring method**) and an algorithm called **iteratively reweighted least squares**.

Supplementary material:

Here we present in some detail how the maximum likelihood estimates of the coefficients of a GLM are obtained. Suppose that we are interested in estimating the parameter vector $(\beta_1, \ldots, \beta_p)^\intercal$ in a GLM. To find the maximum likelihood estimates $\hat{\beta}_j$ we need the scores (multivariate version of the score from Definition 2) expressed as functions of the β_j:

$$U_j = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \left[\frac{\partial l_i}{\partial \beta_j} \right] = \sum_{i=1}^{n} \left[\frac{\partial l_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \beta_j} \right] \tag{7}$$

For an exponential family distribution in canonical form, the components of (7) are:

$$\frac{\partial l_i}{\partial \theta_i} = y_i b'(\theta) + c'(\theta) = b'(\theta) \left[y_i - \left(-\frac{c'(\theta)}{b'(\theta)} \right) \right] = b'(\theta)(y_i - \mu_i) \tag{8}$$

$$\frac{\partial \mu_i}{\partial \theta_i} = -\frac{c''(\theta_i)}{b'(\theta_i)} + \frac{c'(\theta_i) b''(\theta_i)}{[b'(\theta_i)]^2} = b'(\theta_i) \mathrm{Var}(Y_i)$$

$$\Rightarrow \frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b'(\theta_i) \mathrm{Var}(Y_i)} \tag{9}$$

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} x_{ij} = \frac{x_{ij}}{g'(\mu_i)} \tag{10}$$

Substituting (8), (9) and (10) into (7) the expression for the scores becomes

$$U_j = \sum_{i=1}^{n} \left[\frac{(y_i - \mu_i)}{\mathrm{Var}(Y_i)} x_{ij} \frac{\partial \mu_i}{\partial \eta_i} \right] = \sum_{i=1}^{n} \left[\frac{(y_i - \mu_i)}{\mathrm{Var}(Y_i)} \frac{x_{ij}}{g'(\mu_i)} \right]. \tag{11}$$

Note that the scores depend on $\boldsymbol{\beta}$ through $\mu_i = E(Y_i)$ and through $\mathrm{Var}(Y_i)$. The variance-covariance matrix of the U_j has terms $\mathcal{I}_{jk} = E(U_j U_k)$ and is known as the **(Fisher) information matrix**. This is the multivariate version of Definition 3. The elements of matrix $\mathcal{I}$ can be obtained from (11):

$$\begin{aligned} \mathcal{I}_{jk} &= E\left\{ \sum_{i=1}^{n} \left[\frac{(y_i - \mu_i)}{\mathrm{Var}(Y_i)} x_{ij} \frac{\partial \mu_i}{\partial \eta_i} \right] \sum_{l=1}^{n} \left[\frac{(y_l - \mu_l)}{\mathrm{Var}(Y_l)} x_{lk} \frac{\partial \mu_l}{\partial \eta_l} \right] \right\} \\ &= \sum_{i=1}^{n} \frac{E[(Y_i - \mu_i)^2] x_{ij} x_{ik}}{[\mathrm{Var}(Y_i)]^2} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \\ &= \sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{\mathrm{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 = \sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{\mathrm{Var}(Y_i) \left(g'(\mu_i) \right)^2} \end{aligned} \tag{12}$$

Here we have used the fact that $E[(Y_i - \mu_i)(Y_l - \mu_l)] = 0$ by the independence of the Y_i. Notice that the information matrix can be written as

$$\mathcal{I} = \mathcal{I}(\boldsymbol{\beta}) = \boldsymbol{X}^\intercal \boldsymbol{W} \boldsymbol{X} \tag{13}$$

where $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^\intercal \\ \ldots \\ \boldsymbol{x}_n^\intercal \end{bmatrix} = \begin{bmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{np} \end{bmatrix}$, $\boldsymbol{W} = \mathrm{diag}(\boldsymbol{w}) = \begin{bmatrix} w_1 & 0 & \ldots & 0 \\ 0 & w_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \ldots & 0 & w_n \end{bmatrix}$, and

$$w_i = \frac{1}{\mathrm{Var}(Y_i) \left(g'(\mu_i) \right)^2}, \qquad i = 1, \ldots, n. \tag{14}$$

The information matrix $\mathcal{I}(\boldsymbol{\beta})$ depends on $\boldsymbol{\beta}$ through $\boldsymbol{\mu}$ and through $\mathrm{Var}(Y_i)$ for $i = 1, \ldots, n$.

Equation (11) can be written as

$$U_j = \sum_{i=1}^{n} (y_i - \mu_i) x_{ij} w_i g'(\mu_i) = \sum_{i=1}^{n} x_{ij} w_i z_i \qquad j = 1, \ldots, p \tag{15}$$

where $z_i = (y_i - \mu_i)g'(\mu_i)$, so the score can be expressed in vector-matrix form as

$$\boldsymbol{U}(\boldsymbol{\theta}) = \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{z}. \tag{16}$$

Fisher's method of scoring is based on the estimating equation

$$\hat{\boldsymbol{\beta}}^{(m)} = \hat{\boldsymbol{\beta}}^{(m-1)} + [\mathcal{I}^{(m-1)}]^{-1} \boldsymbol{U}^{(m-1)} \tag{17}$$

where $\hat{\boldsymbol{\beta}}^{(m)}$ is the vector of estimates of $(\beta_1, \ldots \beta_p)$ at the mth iteration, $[\mathcal{I}^{(m-1)}]^{-1}$ is the inverse of the information matrix with elements $\mathcal{I}_{jk}$ given by (12), and $\boldsymbol{U}^{(m-1)}$ is the vector of elements given by (11), all evaluated at $\hat{\boldsymbol{\beta}}^{(m-1)}$. Substituting (13) and (16) in (17) we obtain

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{(m)} &= \hat{\boldsymbol{\beta}}^{(m-1)} + \left[\boldsymbol{X}^\top \boldsymbol{W}^{(m-1)} \boldsymbol{X}\right]^{-1} \boldsymbol{X}^\top \boldsymbol{W}^{(m-1)} \boldsymbol{z}^{(m-1)} \\
&= \left[\boldsymbol{X}^\top \boldsymbol{W}^{(m-1)} \boldsymbol{X}\right]^{-1} \boldsymbol{X}^\top \boldsymbol{W}^{(m-1)} \boldsymbol{X} \hat{\boldsymbol{\beta}}^{(m-1)} + \left[\boldsymbol{X}^\top \boldsymbol{W}^{(m-1)} \boldsymbol{X}\right]^{-1} \boldsymbol{X}^\top \boldsymbol{W}^{(m-1)} \boldsymbol{z}^{(m-1)} \\
&= \left[\boldsymbol{X}^\top \boldsymbol{W}^{(m-1)} \boldsymbol{X}\right]^{-1} \boldsymbol{X}^\top \boldsymbol{W}^{(m-1)} (\boldsymbol{\eta}^{(m-1)} + \boldsymbol{z}^{(m-1)})
\end{aligned} \tag{18}$$

This is the same form as the normal equations for weighted least squares, except that it has to be solved numerically because $\boldsymbol{\eta}$, z and $\boldsymbol{W}$ depend on $\hat{\boldsymbol{\beta}}$.

This is why the method to obtain maximum likelihood estimators for GLMs is called **iteratively (re)weighted least squares (IRWLS)**. The procedure begins by using an initial approximation $\hat{\boldsymbol{\beta}}^{(0)}$ to obtain estimates of $\boldsymbol{\eta}$, z and $\boldsymbol{W}$. Then $\hat{\boldsymbol{\beta}}^{(1)}$ is obtained from (18) and is used to update $\boldsymbol{\eta}$, z and $\boldsymbol{W}$. The iterative process continues until the difference between successive approximations $\hat{\boldsymbol{\beta}}^{(m-1)}$ and $\hat{\boldsymbol{\beta}}^{(m)}$ is sufficiently small.

Inference for GLMs

We will now turn our attention to inference for GLMs, mainly through hypothesis tests and the construction of confidence intervals for the parameters of interest. For that we need some sampling distribution results.

Sampling distribution of the MLE

The asymptotic (large sample) distribution for $\hat{\boldsymbol{\beta}}$ is

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \mathcal{I}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{\text{approx}}{\sim} \chi^2(p) \tag{19}$$

or equivalently

$$\hat{\boldsymbol{\beta}} \overset{\text{approx}}{\sim} N_p(\boldsymbol{\beta}, \mathcal{I}^{-1}), \tag{20}$$

where $\chi^2(p)$ refers to the chi-squared distribution with p degrees of freedom and $N_p(\boldsymbol{\beta}, \mathcal{I}^{-1})$ refers to a multivariate (p-dimensional) normal distribution with vector mean $\boldsymbol{\beta}$ and $p \times p$-dimensional variance-covariance matrix $\mathcal{I}^{-1}$.

In the one-dimensional case, what this says is that the MLE $\hat{\beta}$ is approximately normally distributed with mean β and variance $\mathcal{I}^{-1}$.

This allows us to perform hypothesis tests and construct confidence intervals for the model parameterss.

Definition 5 (Wald statistic).

The Wald statistic, *also known as the* z-statistic, *for each of the model parameters* $\{\beta_j\}$, $j = 1, \ldots, p$, *is equal to the coefficient estimate,* $\hat{\beta}_j$, *over its standard error,* $se(\hat{\beta}_j)$.

Under the null hypothesis $H_0 : \beta_j = 0$, and using the asymptotic normality result for the MLE, the Wald statistic is approximately distributed as standard normal. In other words, we have that

$$z = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)} \overset{\text{approx}}{\sim} N(0, 1)$$

This allows us to perform the *Wald test*, which compares the z-statistic to the upper percentile of a standard normal distribution.

Also using this asymptotic normality result, we can construct an approximate 95% confidence interval for β_j by taking

$$\hat{\beta}_j \pm 1.96\text{se}(\hat{\beta}_j).$$

Here 1.96 is the 97.5th percentile of the standard normal distribution.

Example 7 (Hypothesis test and confidence interval for the GPA coefficient in the model for medical school admissions).

Recall the logistic regression model for the medical school admissions data. In the output we see the MLEs of β_0 and β_1 in the `Estimate` column. These are obtained by solving the likelihood equations numerically using Fisher's scoring method (notice the `Number of Fisher Scoring iterations` information towards the end.)

```
summary(med.glm)
```

```
Call:
glm(formula = Acceptance ~ GPA, family = binomial, data = MedGPA)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7805  -0.8522   0.4407   0.7819   2.0967

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -19.207      5.629  -3.412 0.000644 ***
GPA            5.454      1.579   3.454 0.000553 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 75.791  on 54  degrees of freedom
Residual deviance: 56.839  on 53  degrees of freedom
AIC: 60.839

Number of Fisher Scoring iterations: 4
```

We can also see the standard errors for $\hat{\beta}_0$ (`Intercept`) and $\hat{\beta}_1$ (GPA coefficient). These are obtained from the observed information matrix, which is computed during the iterative estimation procedure. Remember that the variance of $\boldsymbol{\beta}$ is estimated by $\mathcal{I}^{-1}$, the inverse of the information matrix. The function `vcov()`returns this estimated variance-covariance matrix of the model coefficients:

```
vcov(med.glm)
```

```
            (Intercept)       GPA
(Intercept)   31.682551 -8.873862
GPA           -8.873862  2.493774
```

The diagonal entries of this matrix are the estimated variances for $\hat{\beta}_0$ and $\hat{\beta}_1$ and the off-diagonal entries give the estimated covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$. The standard errors shown in the output (`Std. Error` column) are the square roots of the estimated variances:

```
sqrt(diag(vcov(med.glm)))
```

```
(Intercept)         GPA
   5.628726    1.579169
```

The `z value` column gives the Wald statistics that test the hypotheses $H_0 : \beta_0 = 0$ and $H_0 : \beta_1 = 0$. Usually we are not as interested in testing whether the intercept term is zero or not and we focus instead on the coefficients of the explanatory variables in the model. So for $H_0 : \beta_1 = 0$ the z value equals 3.454

and is obtained by taking the coefficient estimate and dividing by its standard error

$$z = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)} = \frac{5.454}{1.579} = 3.454.$$

Under H_0, z should approximately follow the standard normal distribution, $N(0,1)$. The p-value shown as `Pr(>|z|)` in the output is the probability of obtaining a value as extreme as z or larger in absolute value, assuming the null hypothesis is true. A small p-value indicates that the z value is unlikely to come from a $N(0,1)$ distribution and leads to rejecting H_0. As the p-value for the GPA coefficient is small (0.000553) we can therefore conclude that the GPA coefficient differs significantly from zero, ie. that the GPA term is worth keeping in the model.

To obtain an approximate 95% confidence interval for the GPA coefficient, we take $\hat{\beta}_j \pm 1.96\text{se}(\hat{\beta}_j)$:

```
5.454-1.96*1.579
```

```
[1] 2.35916
```

```
5.454+1.96*1.579
```

```
[1] 8.54884
```

The resulting interval is $(2.36, 8.55)$, and since it does not include zero, we conclude that the GPA coefficient is significant.

Deviance

To assess the adequacy of a model of interest, we compare it with the **saturated** (or **full**) model, which has the maximum number of parameters that can be estimated. For data with n observations, $y_1, \ldots, y_n$, each with a different parameter in $\boldsymbol{X}^\intercal\boldsymbol{\beta}$, the saturated model can be specified with n parameters. If we have replicates, the maximum number of parameters in the saturated model can be less than n. Let m be the maximum number of parameters that can be estimated, and $\boldsymbol{\beta}_{\max}$ and $\hat{\boldsymbol{\beta}}_{\max}$ be the corresponding parameter vector and MLE.

Let $L(\hat{\boldsymbol{\beta}}_{\max}; \boldsymbol{y})$ be the likelihood evaluated at $\hat{\boldsymbol{\beta}}_{\max}$, that is the likelihood for the full model. Let $L(\hat{\boldsymbol{\beta}}; \boldsymbol{y})$ be the maximum value of the likelihood for a model of interest. The **likelihood ratio**

$$\lambda = \frac{L(\hat{\boldsymbol{\beta}}_{\max}; \boldsymbol{y})}{L(\hat{\boldsymbol{\beta}}; \boldsymbol{y})}$$

provides a measure of how well the model of interest fits compared with the full model. In practice, we often use the logarithm of the likelihood ratio: $\log\lambda = l(\hat{\boldsymbol{\beta}}_{\max}; \boldsymbol{y}) - l(\hat{\boldsymbol{\beta}}; \boldsymbol{y})$. Large values of $\log\lambda$ suggest that the model of interest is a poor description of the data relative to the full model. How large a value of $\log\lambda$? To answer this question we need to obtain a critical region using the sampling distribution of $\log\lambda$. In fact, we will work with the quantity $2\log\lambda$, which is called the **deviance**.

Definition 6 (Deviance). *The deviance, D, is defined as $D = 2\log\lambda = 2[l(\hat{\boldsymbol{\beta}}_{\max}; \boldsymbol{y}) - l(\hat{\boldsymbol{\beta}}; \boldsymbol{y})]$ where $l(\hat{\boldsymbol{\beta}}_{\max}; \boldsymbol{y})$ is the maximised log-likelihood for the saturated model and $l(\hat{\boldsymbol{\beta}}; \boldsymbol{y})$ is the maximised log-likelihood for the model of interest.*

For a GLM that fits the data well, the approximate distribution of the deviance, D, is $\chi^2(m-p)$ where m is the number of parameters in the saturated model and p is the number of parameters in the model of interest.

Example 8 (Deviance for a binomial model).

Suppose $Y_1, \ldots, Y_n$ are independent with $Y_i \sim Bin(n_i, p_i)$ for $i = 1, \ldots, n$. The log-likelihood is

$$l(\boldsymbol{\beta}; \boldsymbol{y}) = \sum_{i=1}^{n} \left[y_i \log p_i - y_i \log(1-p_i) + n_i \log(1-p_i) + \log \binom{n_i}{y_i} \right] \tag{21}$$

For the full model the p_i's are all different, so $\boldsymbol{\beta} = (p_1, \ldots, p_n)^\intercal$ and we can show that $\hat{p}_i = y_i/n_i$. This

gives

$$l(\hat{\boldsymbol{\beta}}_{\max}; \boldsymbol{y}) = \sum \left[y_i \log\left(\frac{y_i}{n_i}\right) - y_i \log\left(\frac{n_i - y_i}{n_i}\right) + n_i \log\left(\frac{n_i - y_i}{n_i}\right) + \log\binom{n_i}{y_i} \right]. \tag{22}$$

For any model with $p < n$ parameters, the MLE of p_i is $\hat{p}_i$ and the fitted values are $\hat{y}_i = n_i \hat{p}_i$. This gives

$$l(\hat{\boldsymbol{\beta}}; \boldsymbol{y}) = \sum_{i=1}^{n} \left[y_i \log\left(\frac{\hat{y}_i}{n_i}\right) - y_i \log\left(\frac{n_i - \hat{y}_i}{n_i}\right) + n_i \log\left(\frac{n_i - \hat{y}_i}{n_i}\right) + \log\binom{n_i}{y_i} \right]. \tag{23}$$

Thus, the deviance is

$$D = 2[l(\hat{\boldsymbol{\beta}}_{\max}; \boldsymbol{y}) - l(\hat{\boldsymbol{\beta}}; \boldsymbol{y})] = 2 \sum_{i=1}^{n} \left[y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i) \log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right) \right]. \tag{24}$$

Example 9 (Deviance for a normal linear model).

Suppose $Y_1, \ldots, Y_n$ are independent with $Y_i \sim N(\mu_i, \sigma^2)$ and $E(Y_i) = \mu_i = \boldsymbol{X}_i^\intercal \boldsymbol{\beta}$ for $i = 1, \ldots, n$. The log-likelihood function is

$$l(\boldsymbol{\beta}; \boldsymbol{y}) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu_i)^2 - \frac{1}{2} n \log(2\pi\sigma^2). \tag{25}$$

For the saturated model, we have n parameters $\mu_1, \ldots, \mu_n$. The MLEs are $\hat{\mu}_i = y_i$ and so the maximum value of the log-likelihood becomes

$$l(\hat{\boldsymbol{\beta}}_{\max}; \boldsymbol{y}) = -\frac{1}{2} n \log(2\pi\sigma^2). \tag{26}$$

For any other model with $p < n$ parameters, the MLE of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \boldsymbol{y}$. The corresponding maximised log-likelihood function is

$$l(\hat{\boldsymbol{\beta}}; \boldsymbol{y}) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\intercal \hat{\boldsymbol{\beta}})^2 - \frac{1}{2} n \log(2\pi\sigma^2). \tag{27}$$

The deviance then is

$$D = 2[l(\hat{\boldsymbol{\beta}}_{\max}; \boldsymbol{y}) - l(\hat{\boldsymbol{\beta}}; \boldsymbol{y})] = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^\intercal \hat{\boldsymbol{\beta}} \right)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 . \tag{28}$$

It turns out that the (exact) distribution of D is $\chi^2(n-p)$. If the model fits the data well, then $D \sim \chi^2(n-p)$, and the expected value of D will be $n - p$, since the expectation of a chi-squared random variable is equal to its degrees of freedom.

However, we are not able to use this chi-squared distribution directly, because the expression for the deviance contains the nuisance parameter σ^2. As you may remember from your linear regression courses, we end up using F tests instead.

Example 10 (Deviance for a Poisson model).

Let $Y_1, \ldots, Y_n$ be independent random variables with $Y_i \sim Po(\mu_i)$. Then the log-likelihood function is

$$l(\boldsymbol{\beta}; \boldsymbol{y}) = \sum y_i \log \mu_i - \sum \mu_i - \sum \log(y_i!). \tag{29}$$

For the full model $\boldsymbol{\beta}_{\max} = (\mu_1, \ldots, \mu_n)^\intercal$, $\hat{\mu}_i = y_i$, and the maximum value of the log-likelihood is

$$l(\hat{\boldsymbol{\beta}}_{\max}; \boldsymbol{y}) = \sum y_i \log y_i - \sum y_i - \sum \log(y_i!). \tag{30}$$

Suppose that for the model of interest with $p < n$ parameters the MLE, $\hat{\boldsymbol{\beta}}$, can be used to obtain $\hat{\mu}_i$ and hence fitted values $\hat{y}_i = \hat{\mu}_i$ (because $E(Y_i) = \mu_i$). For the model of interest the maximum value of the log-likelihood is

$$l(\hat{\boldsymbol{\beta}}; \boldsymbol{y}) = \sum y_i \log \hat{y}_i - \sum \hat{y}_i - \sum \log(y_i!). \tag{31}$$

Hence, the deviance is

$$D = 2[l(\hat{\boldsymbol{\beta}}_{\max}; \boldsymbol{y}) - l(\hat{\boldsymbol{\beta}}; \boldsymbol{y})] = 2\left[\sum y_i \log\left(\frac{y_i}{\hat{y}_i}\right) - \sum (y_i - \hat{y}_i)\right] \tag{32}$$

For most models $\sum y_i = \sum \hat{y}_i$, so the deviance can be written as

$$D = 2\sum y_i \log\left(\frac{y_i}{\hat{y}_i}\right) = 2\sum o_i \log\left(\frac{o_i}{e_i}\right), \tag{33}$$

where o_i denotes the observed value and e_i the expected value of y_i. The deviance can be computed from the data and compared with the $\chi^2(n-p)$ distribution.

Consider the data below for which the Y_i are assumed to be independent observations from a Poisson distribution.

y_i	2	3	6	7	8	9	10	12	15
x_i	-1	-1	0	0	0	0	1	1	1

We fit a model of the form $\mu_i = \beta_1 + \beta_2 x_i$. The fitted values are $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$ where $\hat{\beta}_1 = 7.45163$ and $\hat{\beta}_2 = 4.93530$. The deviance is 1.8947, which is small compared with $n - p = 7$, indicating no lack of fit.

Task 5.

Verify the value 1.8947 for the deviance given in the above example.

Hypothesis testing using the deviance

As we've already seen, we can test hypotheses about the p-dimensional parameter vector $\boldsymbol{\beta}$ by using the Wald statistic and the asymptotic distribution of the MLE.

Alternatively we can compare **nested** models M_0 and M_1 using the difference of their deviances.

Consider $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 = (\beta_1, \ldots, \beta_q)^\intercal$ corresponding to M_0 and $H_1 : \boldsymbol{\beta} = \boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_p)^\intercal$ corresponding to M_1 with $q < p < n$. Test H_0 against H_1 by considering

$$\begin{aligned} D_0 - D_1 &= 2[l(\hat{\boldsymbol{\beta}}_{\max}; \boldsymbol{y}) - l(\hat{\boldsymbol{\beta}}_0; \boldsymbol{y})] - 2[l(\hat{\boldsymbol{\beta}}_{\max}; \boldsymbol{y}) - l(\hat{\boldsymbol{\beta}}_1; \boldsymbol{y})] \\ &= 2[l(\hat{\boldsymbol{\beta}}_1; \boldsymbol{y}) - l(\hat{\boldsymbol{\beta}}_0; \boldsymbol{y})] \end{aligned}$$

If both models describe the data well then $D_0 \sim \chi^2(n-q)$, $D_1 \sim \chi^2(n-p)$ and $D_0 - D_1 \sim \chi^2(p-q)$. If M_1 describes the data well but M_0 does not, then $D_0 - D_1$ will be larger than expected for a value from $\chi^2(p-q)$. So, reject H_0 if $D_0 - D_1 > \chi^2(1-\alpha; p-q)$ that is, if the difference in deviances exceeds the upper $100 \times \alpha\%$ point of the $\chi^2(p-q)$ distribution.

Example 11 (Hypothesis test for the GPA coefficient in the model for medical school admissions, this time using deviances).

Suppose that we want to test $H_0 : \beta_1 = 0$ in the medical school admissions example. We can perform this test using the deviances given in the output.

```
Call:
glm(formula = Acceptance ~ GPA, family = binomial, data = MedGPA)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
```

```
-1.7805  -0.8522   0.4407   0.7819   2.0967

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -19.207      5.629  -3.412 0.000644 ***
GPA            5.454      1.579   3.454 0.000553 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 75.791  on 54  degrees of freedom
Residual deviance: 56.839  on 53  degrees of freedom
AIC: 60.839

Number of Fisher Scoring iterations: 4
```

Here D_0 is the null deviance, that is the deviance in the model that includes only the intercept (and no other predictors). D_1 is the residual deviance, that is the deviance of the model of interest (the model with GPA included as a predictor). Under H_0 $D_0 - D_1$ should be approximately distributed as $\chi^2(1)$. The 95th percentile of the $\chi^2(1)$ distribution is $\chi^2(0.95; 1) = 3.84$, and as $D_0 - D_1 = 75.791 - 56.839 = 18.952 > 3.84$ we can reject the null hypothesis. Again we conclude that GPA is a significant term in the model.

Week 2 learning outcomes

- Know the scope of generalised linear models (GLMs): what do they have in common with the normal linear model and in which ways do they generalise it?
- Be familiar with the properties of the exponential family of distributions and how they relate to GLMs; recognise common distributions that are members of this family.
- Have a basic understanding of how coefficient estimates and standard errors are obtained in a GLM; be familiar with the concepts of likelihood, maximum likelihood estimation, testing, confidence intervals, and goodness of fit statistics in the context of GLMs.
- Be familiar with the main ways of doing inference on the parameters of a GLM – these are based on large sample distribution results for the maximum likelihood estimator and the deviance. Be able to test the significance of a term in a GLM by performing a Wald test or by comparing deviances between nested models.

Answers to tasks

Answer to Task 1. For the *Bollywood box office revenue* example, we can write down the fitted model equation from the `summary(bol.lm)`:

$$\log 10(\text{Gross}) = -0.62549 + 1.31955 \times \log 10(\text{Budget})$$

We can use this equation to predict the gross revenue of a film by simply substituting the relevant budget value, and transforming the result from the log10 scale. Thus:

(i) budget = 10:

$$\log 10(\text{Gross}) = -0.62549 + 1.31955 \times \log 10(10) = 0.69406 \Rightarrow \text{Gross} = 10^{0.69406} = 4.94379$$

(ii) budget = 50:

$$\log 10(\text{Gross}) = -0.62549 + 1.31955 \times \log 10(50) = 1.616386 \Rightarrow \text{Gross} = 10^{1.616386} = 41.34148$$

(iii) budget = 100:

$$\log 10(\text{Gross}) = -0.62549 + 1.31955 \times \log 10(100) = 2.01361 \Rightarrow \text{Gross} = 10^{2.01361} = 103.1834$$

Answer to Task 2. In the *GPA and admission to medical school* example, we can write down the fitted model equation from the `summary(med.glm)`:

$$\log\left(\frac{p_i}{1-p_i}\right) = -19.207 + 5.454 \times \text{GPA}$$

From the fitted equation, we can obtain the acceptance probability by solving for p_i:

$$\hat{p}_i = \frac{\exp(-19.207 + 5.454 \times \text{GPA})}{1 + \exp(-19.207 + 5.454 \times \text{GPA})}$$

To predict the acceptance probability for an applicant we just need to substitute the specified GPA in the equation for $\hat{p}_i$:

(i) GPA = 2.5:

$$\hat{p}_i = \frac{\exp(-19.207 + 5.454 \times 2.5)}{1 + \exp(-19.207 + 5.454 \times 2.5)} \Rightarrow \hat{p}_i = 0.00378$$

(ii) GPA = 3:

$$\hat{p}_i = \frac{\exp(-19.207 + 5.454 \times 3)}{1 + \exp(-19.207 + 5.454 \times 3)} \Rightarrow \hat{p}_i = 0.05494$$

(iii) GPA = 4:

$$\hat{p}_i = \frac{\exp(-19.207 + 5.454 \times 4)}{1 + \exp(-19.207 + 5.454 \times 4)} \Rightarrow \hat{p}_i = 0.93143$$

Alternatively, we can use the `predict()` function in `R` as follows:

```
predict(med.glm, data.frame(GPA = c(2.5, 3, 4)), type = 'response')
```

```
          1           2           3
0.003791903 0.054992029 0.931512655
```

Answer to Task 3. Consider $Y \sim \text{Bin}(n, \theta)$ with p.m.f.

$$f(y;\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y} \text{ for } y = 0, 1, \ldots, n. \tag{34}$$

By rewriting the p.m.f. as

$$f(y;\theta) = \exp\left[y\log\theta - y\log(1-\theta) + n\log(1-\theta) + \log\binom{n}{y}\right] \tag{35}$$

$$= \exp\left[y\log\frac{\theta}{1-\theta} + n\log(1-\theta) + \log\binom{n}{y}\right] \tag{36}$$

we see that this is a member of the exponential family in canonical form, and its natural parameter is given by: $b(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$.

Answer to Task 4. Random component: Let $Y_i = 1$ if the ith applicant is accepted to medical school and $Y_i = 0$ if not. We assume that the Y_i are independent responses from Bin$(1, p_i)$, with $E(Y_i) = p_i$ for i=1,…,55.

Systematic component: $\beta_0 + \beta_1 x_i$ where x_i is the ith applicant's GPA and β_0 and β_1 are parameters to be estimated.

Link function: $g(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$ (logit link)

Equation of the GLM:

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

Answer to Task 5. We need to get the fitted values $\hat{y}_i$ for $i = 1, \ldots, 9$ and then substitute into the expression for the deviance.

y_i	2	3	6	7	8	9	10	12	15
$\hat{y}_i$	2.51633	2.51633	7.45163	7.45163	7.45163	7.45163	12.38693	12.38693	12.38693
$y_i \log\left(\frac{y_i}{\hat{y}_i}\right)$	-0.45931	0.52743	-1.30004	-0.43766	0.56807	1.69913	-2.14057	-0.38082	2.87115

So $\sum_{i=1}^{9} y_i \log\left(\frac{y_i}{\hat{y}_i}\right) = 0.94738$ and $D = 2\sum_{i=1}^{9} y_i \log\left(\frac{y_i}{\hat{y}_i}\right) = 1.89476$.