37262 Mathematical Statistics

Lecture 9

Modelling Different Response Types

- A simple linear regression model of the form $y_i = \alpha + \beta x_i + \varepsilon_i$ can (depending on what values x can take) return values for y which can be arbitrarily large, positively or negatively.

- This response type simply doesn't suit many datasets. For example, if we are seeking to predict how many times a given outcome will be observed, the predicted response must be a count (i.e. a non-negative integer – $\{0,1,2,3,...\}$

- Alternatively, we might simply be trying to predict whether or not a given outcome will or will not occur given some predictor variables. In this case, we have a binary response (i.e. one of two possible responses.)

- Such models can be obtained through **generalised linear models** or **GLMs**.

Generalised Linear Models

- A generalised linear model applies a **link function** to a standard linear model to ensure that the response type is appropriate for the dataset.

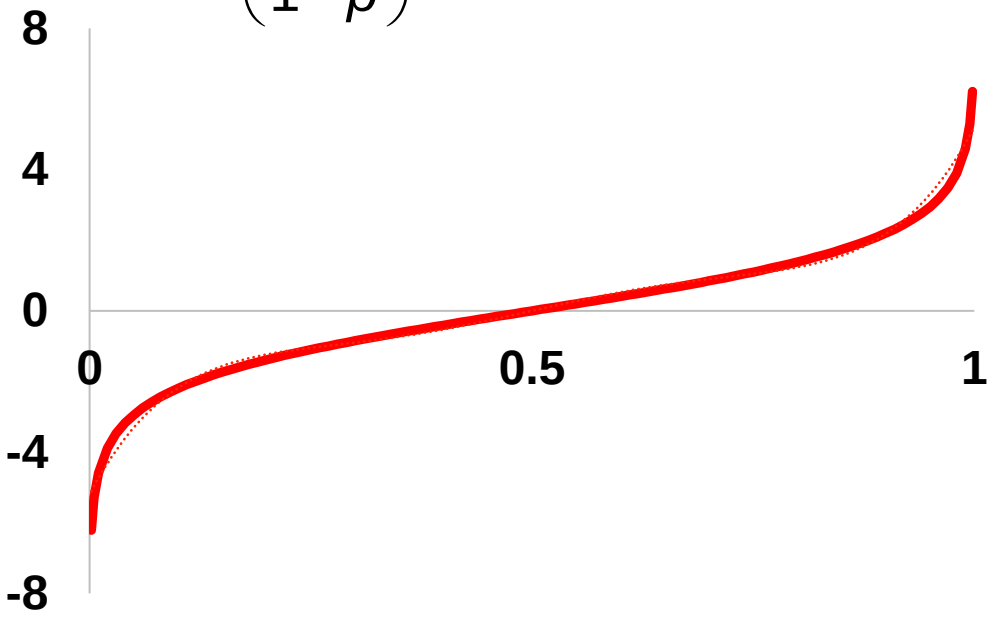
- For example, if we are seeking a binary regression model, we seek to predict if a given outcome will (1) or will not (0) occur given predictor information.

- We do this by using a function which transforms the possible outcomes of a linear regression (which could, at least in theory, take any real values) and converting this to values between 0 and 1.

- The two most common functions for binary regression are the **logit** and **probit** functions.

Binary Logistic Regression

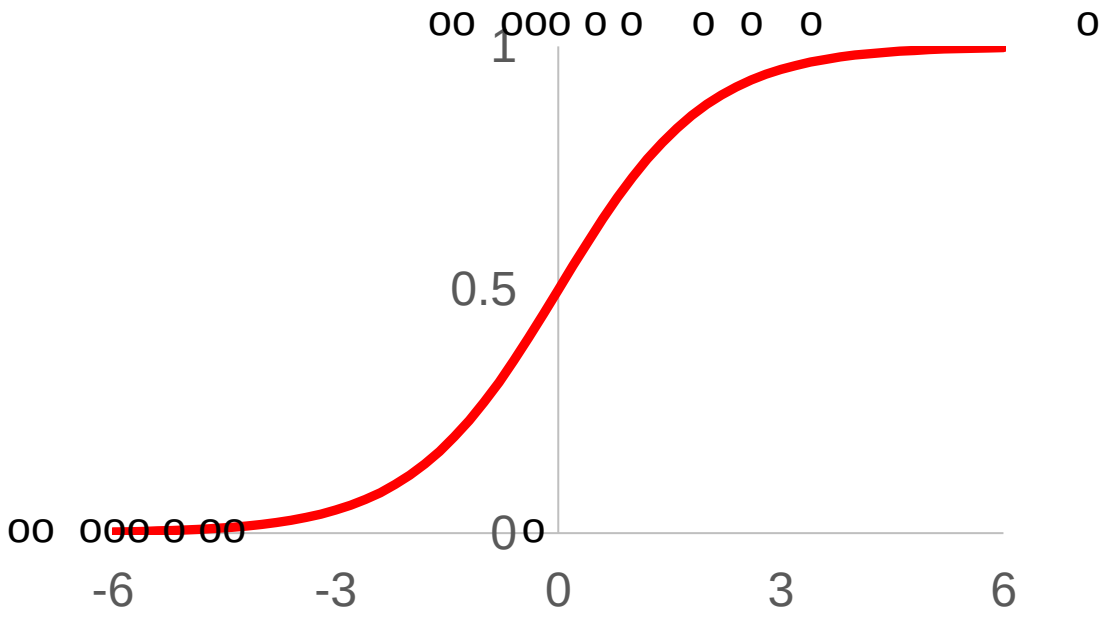
- The logit of a probability p $(0 < p < 1)$ is defined as $\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$.



- As p approaches zero, the logit function tends off towards minus infinity.
- As p approaches one, the logit function tends off towards positive infinity.
- We therefore run a regression model to predict the inverse of the logit function and convert this to a probability between 0 and 1.

Binary Logistic Regression

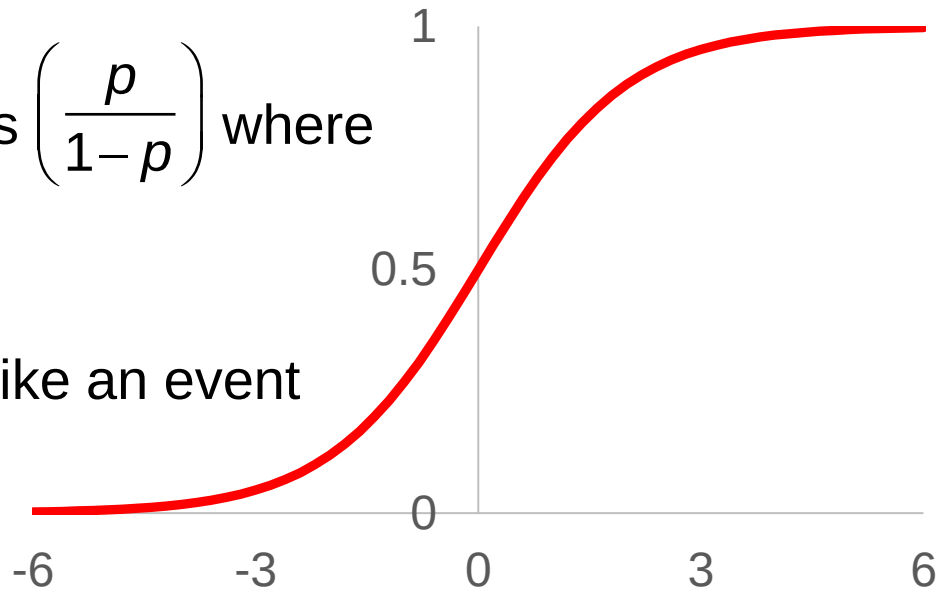
- Fitting to an observed dataset, we already know the 0s and 1s (i.e. whether or not the outcome occurred) and want to work backwards to see which predictor values are more associated with 1s than 0s.



- We therefore need the inverse of the logit function.
- If $y = \ln\left(\frac{p}{1-p}\right)$, then $e^{-y} = \frac{1-p}{p} = \frac{1}{p} - 1$.
- $p = \frac{1}{1+e^{-y}} = \frac{e^{y}}{1+e^{y}}$
- We then undertake a linear regression to find the value of *y* which best fits this curve.

Binary Logistic Regression: Odds Ratio

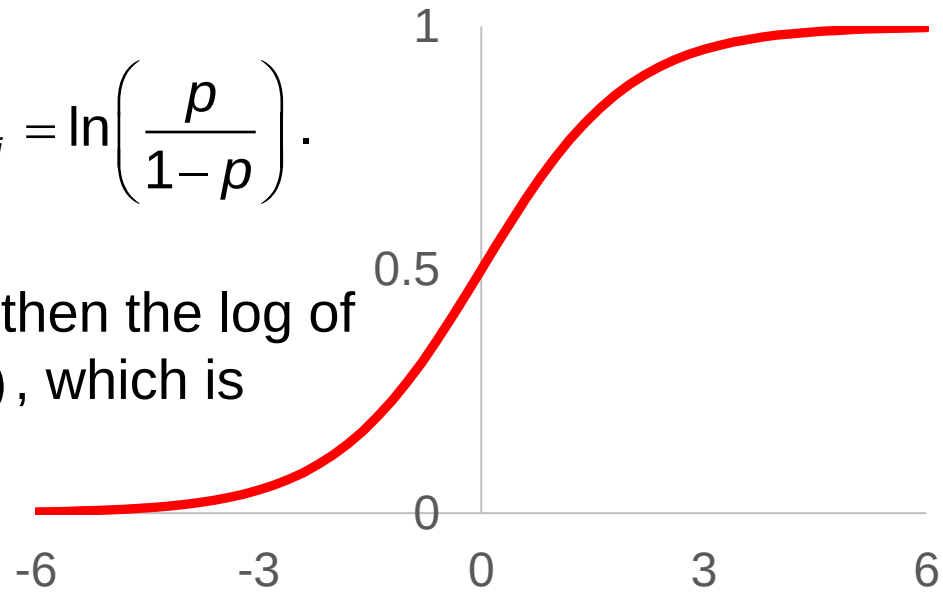
- One of the strengths of the logit function for a generalised linear model is that it gives easily interpreted parameters.
- The **odds ratio** of an outcome occurring are defined as $\left(\frac{p}{1-p}\right)$ where p is the probability that it occurs.
- You can think of odds as being how many times more like an event is to happen than not. For example, odds of 3 mean that an event has 75% chance of occurring and 25% chance of not occurring (since 75/25 = 3.)



- An event has odds of 1 if is equally likely to happen as not (i.e. 50% chance.)

Binary Logistic Regression: Odds Ratio and Interpretation

- When we fit use the logit function as the link function for a generalised linear model, we automatically get the log of the odds ratio.
- We effectively fit (for a simple linear model) $y_i = \alpha + \beta x_i = \ln\left(\frac{p}{1-p}\right)$.
- We note that if the predictor variable *x* increases by 1, then the log of the odds ratio increases from $(\alpha + \beta x_i)$ to $(\alpha + \beta(x_i + 1))$, which is an increase of β.
- This means that the odds of the event changes by a factor of e^{β}.



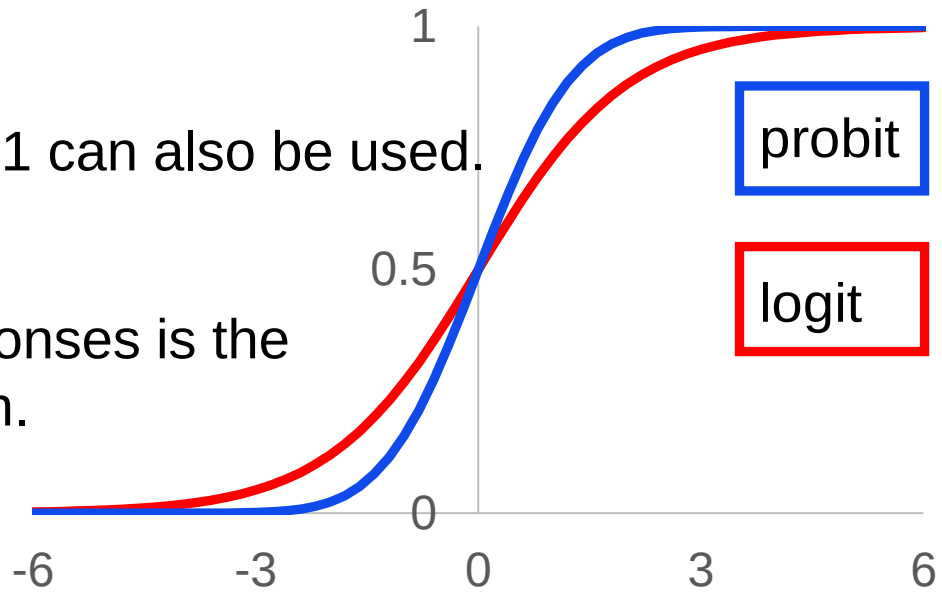
Binary Regression

- The logit function $\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$ is not the only possible function which can be used for binary responses (i.e. 0s and 1s.)

- Other functions which take only values between 0 and 1 can also be used.

- Another (less commonly-used) function for binary responses is the probit function, based on a standard normal distribution.

- $\text{probit}(p) = \left\{ L \text{ such that } p = \int_{-\infty}^{L} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \right\}$



Exponential Family

- We have a general framework for fitting models to many of the most common distributions.

- A distribution which depends on a single parameter θ is said to belong to the **exponential family** if its probability density or probability mass function can be written in the form $f(x) = a(x)b(\theta)\exp\big(c(\theta)d(x)\big)$ and where the range of x does not depend on θ.

- A similar statement also applies for multiparameter distributions. In this case, the distribution belongs to the exponential family if, for parameter vector $\boldsymbol{\Theta} = (\theta_1,...,\theta_n)$, it can be written in the form $f(x) = a(x)b(\boldsymbol{\Theta})\exp\big(\boldsymbol{c}(\boldsymbol{\Theta}).\boldsymbol{d}(x)\big)$ where . denotes the dot (or scalar) product ahd, again, the range of the variable does not depend on the parameter values.

Exponential Family

- All of the following distributions belong to the exponential family.

- Some of these are easily shown to be expressible in the form of the exponential family, other are more less obvious.

Normal	Geometric
Exponential	Poisson
Gamma	Beta
Bernoulli	Chi-squared
Binomial (if the number of trials is known)	
Negative Binomial (if the number of required successes is known)	

- Note that distributions such as the uniform variable $U[a,b]$ with range $[a,b]$ and the Pareto variable $Pareto(m,\alpha)$ with range $[m,\infty)$ do not belong to the exponential family as their ranges depend on the parameters.

Exponential Family: Example

- Consider $X \sim Bin(n,p)$ where n is known.

- This has probability mass function $f(x) = P(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$.

- $f(x) = \frac{n!}{x!(n-x)!}\left(\frac{p}{1-p}\right)^x (1-p)^n = \frac{n!}{x!(n-x)!}(1-p)^n \exp\left(x \ln\left(\frac{p}{1-p}\right)\right)$

- $f(x) = \boxed{\frac{n!}{x!(n-x)!}} \boxed{(1-p)^n} \exp\left(\boxed{x} \boxed{\ln\left(\frac{p}{1-p}\right)}\right)$

$= \boxed{a(x)} \times \boxed{b(p)} \times \exp\left(\boxed{c(p)} \times \boxed{d(x)}\right)$

Exponential Family: Canonical Form

- Every member of the exponential family can be expressed in **canonical form** or **standard form**.

- If, in the form $f(x) = a(x)b(\theta)\exp\big(c(\theta)d(x)\big)$, then $c(\theta)$ is said to be the **natural parameter**.

- The canonical form is then $f(x) = a(x)b(c)\exp\big(c \times d(x)\big)$ in terms of this natural parameter.

- For example, it is possible to define a Bernoulli variable with known number of trials, not by the success probability p, but rather by the log-odds ratio $c(p) = \ln\left(\frac{p}{1-p}\right)$

Exponential Family: Example

- Consider $Q \sim Poi(\lambda)$.
- This has probability mass function $f(q) = P(Q = q) = \frac{e^{-\lambda}\lambda^q}{q!}$.
- $f(q) = \frac{e^{-\lambda}\lambda^q}{q!} = \left(\frac{1}{q!}\right)e^{-\lambda}\exp\left(q\ln(\lambda)\right)$
- This is in canonical form with natural parameter $c(\lambda) = \ln(\lambda)$.

Exponential Family: Mean

- Consider the general form of the exponential family $f(x) = a(x)b(\theta)\exp\left(c(\theta)d(x)\right)$
- Let us first redefine this in terms of the natural parameter c. $f(x) = a(x)b(c)\exp\left(cd(x)\right)$
- We know that $\int f(x)dx = b(c)\int a(x)\exp\left(cd(x)\right)dx = 1$
- $\frac{\partial}{\partial c} b(c)\int a(x)\exp\left(cd(x)\right)dx = \frac{\partial}{\partial c}(1) = 0$
- $b'(c)\int a(x)\exp\left(cd(x)\right)dx + b(c)\int a(x)\frac{\partial}{\partial c}\left(\exp\left(cd(x)\right)\right)dx = 0$

Exponential Family: Mean

- $b'(c)\int a(x)\exp\left(cd(x)\right)dx + b(c)\int a(x)\frac{\partial}{\partial c}\left(\exp\left(cd(x)\right)\right)dx = 0$
- $\frac{b'(c)}{b(c)}\int a(x)b(c)\exp\left(cd(x)\right)dx + b(c)\int a(x)d(x)\exp\left(cd(x)\right)dx = 0$
- $\frac{b'(c)}{b(c)}\int f(x)dx + \int d(x)f(x)dx = 0$
- $\frac{\partial}{\partial c}\ln(b(c)) + E(d(x)) = 0$ hence $E(d(x)) = -\frac{\partial}{\partial c}\ln(b(c))$.
- If $d(x) = x$ then hence $E(X) = -\frac{\partial}{\partial c}\ln(b(c))$.

Exponential Family: Mean

- For the Poisson distribution, our distribution was $f(q) = \frac{e^{=\lambda}\lambda^q}{q!} = \left(\frac{1}{q!}\right)e^{=\lambda}\exp\left(q\ln(\lambda)\right)$
- This gave the natural parameter of $c = \ln(\lambda)$.
- We also obtained $b(\lambda) = e^{-\lambda}$ hence $b(c) = e^{-e^c}$.
- By the method on the previous slide, we obtain $E(Q) = -\frac{\partial}{\partial c}\ln(b(c)) = -\frac{\partial}{\partial c}\ln(e^{-e^c}) = \frac{\partial}{\partial c}e^c$
- This gives $E(Q) = \frac{\partial}{\partial c}e^c = e^c = \lambda$.

Exponential Family: Variance

- Although a little messier to justify, we also have the property that, for natural parameter c, the variance of a distribution defined by the exponential family $f(x) = a(x)b(c)\exp\left(cd(x)\right)$ is easily calculated by $Var(d(X)) = -\frac{\partial^2}{\partial c^2}\ln(b(c))$.

- For the Poisson, we therefore have $Var(Q) = -\frac{\partial^2}{\partial c^2}\ln(b(c)) = -\frac{\partial^2}{\partial c^2}\ln(e^{-e^c}) = \frac{\partial^2}{\partial c^2}e^c = e^c = \lambda$.

Exponential Family: Variance

- Recall now $X \sim Bin(n,p)$

- This can be written as $f(x) = \frac{n!}{x!(n-x)!}(1-p)^n \exp\left(x \ln\left(\frac{p}{1-p} \right) \right) = \frac{n!}{x!(n-x)!}(1-e^{\theta})^n \exp(x\theta)$

where $\theta = \ln\left(\frac{p}{1-p} \right)$.

Exponential Family: Variance

- The mean is therefore $E(X) = -\frac{\partial}{\partial\theta}\ln((1-e^{\theta})^{n}) = -n\frac{-e^{\theta}}{1-e^{\theta}} = n\left(\frac{p}{1-p}\right)\Big/\left(\frac{1}{1-p}\right) = np.$
- Similarly, the variance is $Var(X) = -\frac{\partial^2}{\partial\theta^2}\ln((1-e^{\theta})^{n}) = n\frac{\partial}{\partial\theta}\left(\frac{e^{\theta}}{1-e^{\theta}}\right) = n\left(\frac{e^{\theta}}{(1-e^{\theta})^2}\right)$
- $Var(X) = n\left(\frac{e^{\theta}}{(1-e^{\theta})^2}\right) = n\left(\frac{p}{1-p}\right)\Big/\left(\frac{1}{1-p}\right)^2 = np(1-p).$

Generalised Linear Models

- We return now to generalised linear models.

- A generalised linear model consists of three elements.
- Given an observation $\boldsymbol{X} = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_n \end{pmatrix}$ of n predictor variables, we have a linear predictor $\eta = \boldsymbol{X}^t \begin{pmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}$
- We also have an assumed distribution for modelling $\boldsymbol{Y}$ – usually a member of the exponential family.
- Finally, we have a link function g such that $g(\mu) = \eta$ where $\mu = E(\boldsymbol{Y} \mid \boldsymbol{X})$

Generalised Linear Models

- Trivially, if we set the link function $g(\mu) = \mu$ then we obtain $E(y_i \mid x_i) = \mu_i = \eta_i = \alpha + \beta_1 x_{1i} + ... + \beta_n x_{ni}$ i.e. the systematic (not random) component of a standard linear model.

- In general, for other types of regression model, we can select the natural parameter from the canonical form of the exponential family as the link function. For example, Poisson regression can be done using $\eta = \boldsymbol{X}^t \boldsymbol{\beta} = \ln(\mu)$ since the natural parameter for the Poisson distribution was the log of the mean.

- Similarly, we have that the logit function is a link function for binomial regression.

Generalised Linear Models

- Recall that we could use ordinary least squares to fit models where the residuals had some covariance structure (other than the usually-assumed structure of independent and identically distributed structure) via generalised least squares.

- It is beyond the scope of 37262, but we can similarly fit GLMs by **an iterated weighted least squares** approach.

- This minimises a weighted sum of squared residuals for estimated parameters, then recalculates the weights and then is repeated until the parameters converge. Most modern statistical packages can do this very efficiently.

Generalised Linear Models: Score Function

- Recall from Lecture 8 that, for a loglikelihood function $\ell(\boldsymbol{X} \mid \theta) = \ln(L(\boldsymbol{X} \mid \theta))$, we had the score function $V = Score(\boldsymbol{X} \mid \theta) = \frac{\partial}{\partial \theta} \ell(\boldsymbol{X} \mid \theta)$.
- We had that $E(V) = E\left(\frac{\partial}{\partial \theta} \ln L(\boldsymbol{X} \mid \theta) \right) = \int \frac{\frac{\partial}{\partial \theta} L(\boldsymbol{X} \mid \theta)}{L(\boldsymbol{X} \mid \theta)} L(\boldsymbol{X} \mid \theta) d\boldsymbol{X} = \frac{\partial}{\partial \theta} \int L(\boldsymbol{X} \mid \theta) d\boldsymbol{X}$.
- As $\int L(\boldsymbol{X} \mid \theta) d\boldsymbol{X} = 1$, $E(V) = \frac{\partial}{\partial \theta} 1 = 0$.
- We further defined $I(\theta) = Var(V) = -E\left(\frac{\partial^2}{\partial \theta^2} \ln L(\boldsymbol{X} \mid \theta) \right)$ to be the Fisher information, a measure of information about the parameter θ contained in the sample $\boldsymbol{X}$.

Generalised Linear Models: Score Function

- In the case of multiple parameters, $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}$ the score statistic is similarly defined

$$V = Score(\boldsymbol{X} \mid \boldsymbol{\beta}) = \frac{\partial}{\partial \beta_1} \ell(\boldsymbol{X} \mid \boldsymbol{\beta}) + \ldots + \frac{\partial}{\partial \beta_n} \ell(\boldsymbol{X} \mid \boldsymbol{\beta}) \quad = \nabla \ell(\boldsymbol{X} \mid \boldsymbol{\beta}).$$

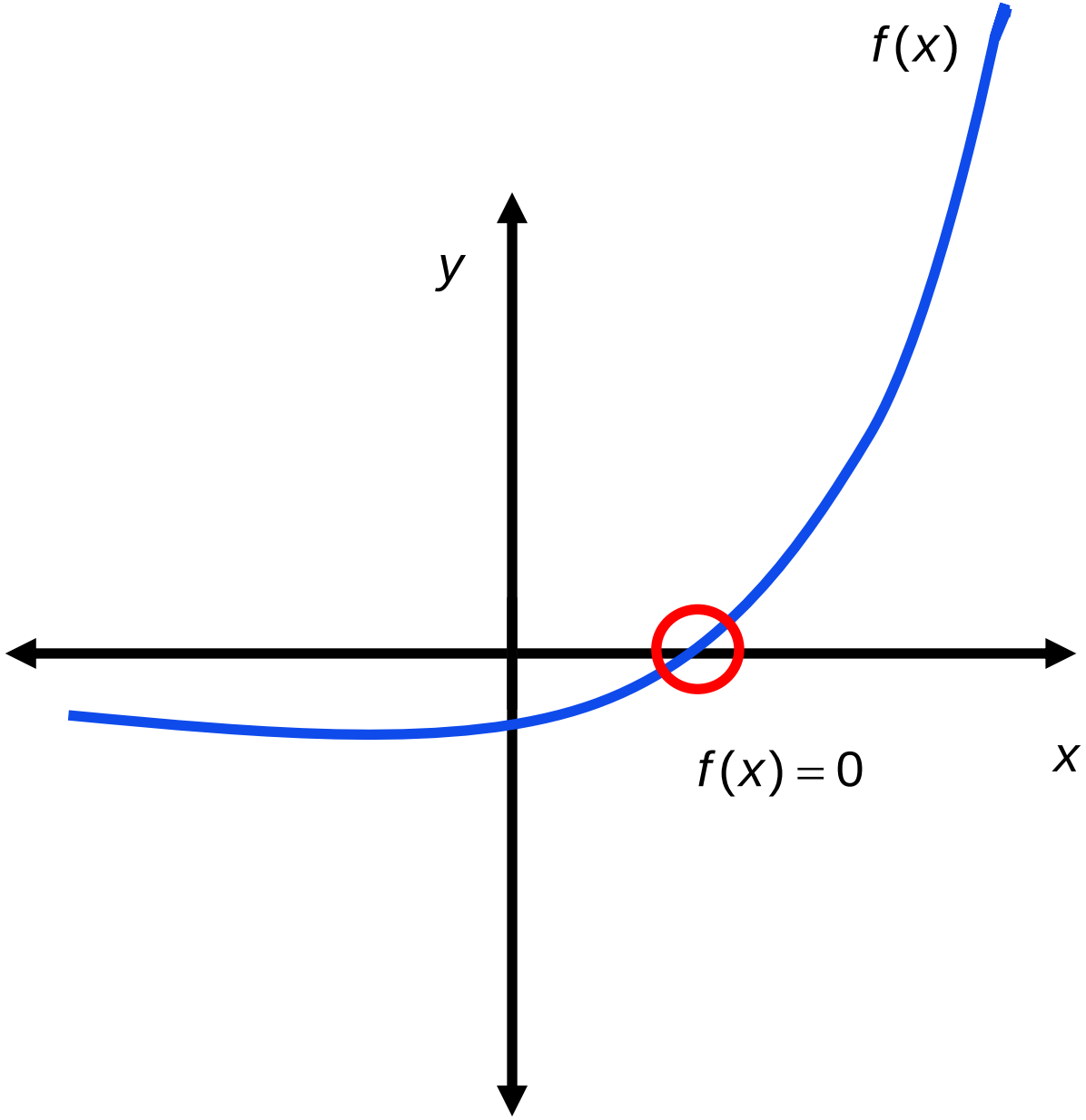
- With link function $\boldsymbol{\eta} = \boldsymbol{X}^t \boldsymbol{\beta}$, we obtain $V_\beta = \boldsymbol{X} V_\eta$

- Via the matrix chain rule, we obtain the Fisher information $I(\boldsymbol{\beta}) = \boldsymbol{X} I(\boldsymbol{\eta}) \boldsymbol{X}^t$

Newton-Raphson Method

- Recall the **Newton-Raphson Method** (or Newton's Method), which involves repeatedly solving simpler functions which approximate to the original function.

- For example, consider the problem of finding a root of the equation $f(x) = 0$.

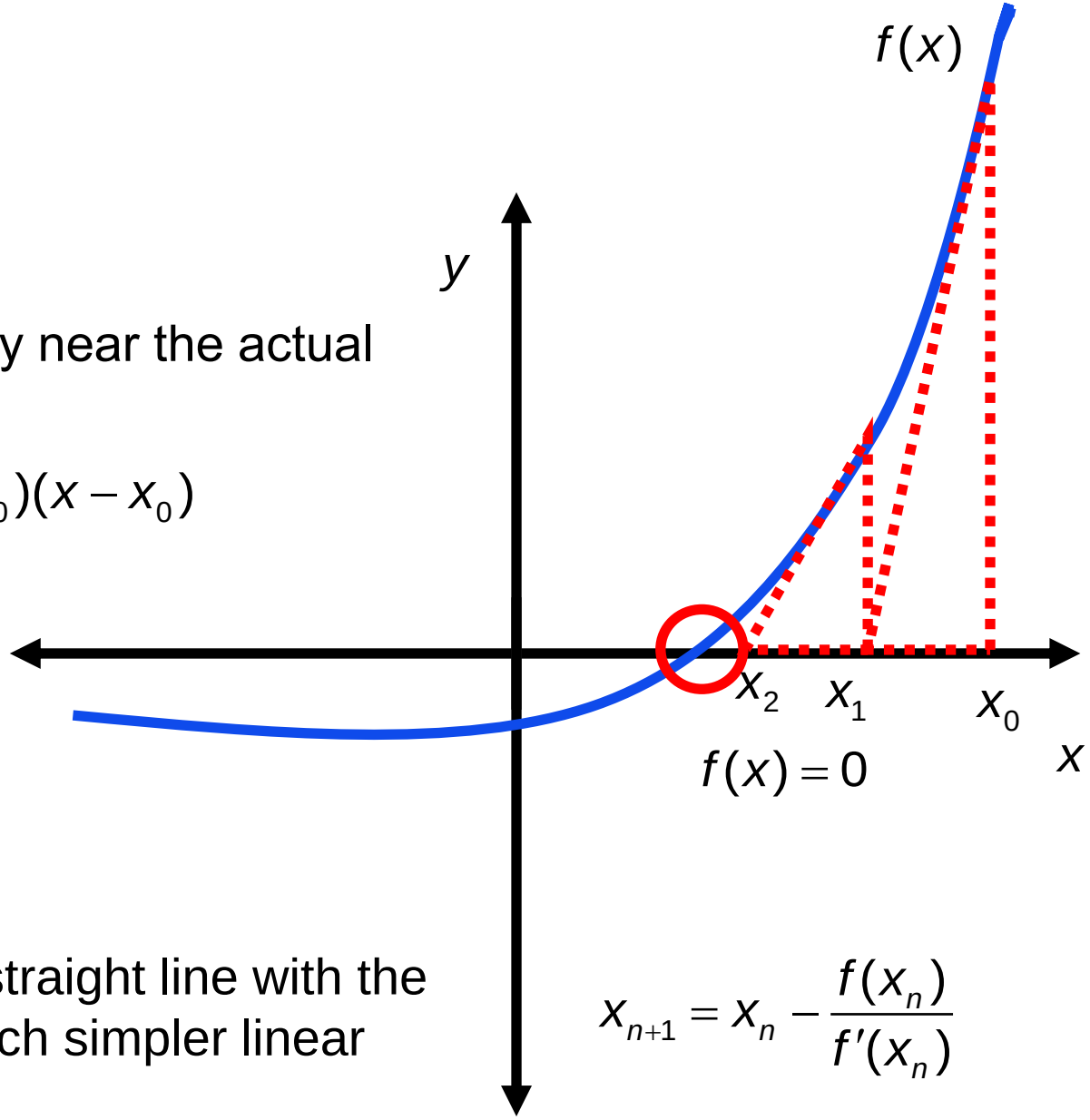


Newton-Raphson Method

- We start with a “guess” of the solution, hopefully near the actual solution. Call this point x_0 .
- Expanding $\Delta f \approx \Delta x \frac{df}{dx}$ gives $f(x) - f(x_0) \approx f'(x_0)(x - x_0)$

- We solve this to find the solution which will be our next “guess” x_1 .
- $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$

- This works by approximating our function to a straight line with the same gradient at each step and solving the much simpler linear function instead.



$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

Fisher Scoring

- Similarly, the weighted least squares problem $I(\boldsymbol{\beta}) = \boldsymbol{X}I(\boldsymbol{\eta})\boldsymbol{X}^t$ is solved with an iterative approach similar to the Newton-Raphson method.

- We know that the likelihood is maximised when the expected value of the score function is zero, so it is simply an exercise in finding a root of the equation setting the expectation of the score function to zero.

- This approach is the **Fisher Scoring** algorithm which solves by weighted least squares, at each step of the algorithm, re-estimating the weights until the system converges to a solution.

Fisher Scoring

- For a single variable with initial estimate η_0 , this algorithm is simply

$$\eta_{m+1} = \eta_m + \frac{V(\boldsymbol{X} \mid \eta_m)}{I(\boldsymbol{X} \mid \eta_m)}$$

- Where $V(\boldsymbol{X} \mid \eta_m)$ and $I(\boldsymbol{X} \mid \eta_m)$ are, respectively, the score function and Fisher information associated with the sample $\boldsymbol{X}$.

Generalised Linear Models

- It should be noted, however, that this approach can produce some confusing outputs which require careful attention.

- For example, although the iterated least squares procedure relies on the Fisher information (based on the second derivates of loglikelihood), it is essentially aiming to maximise likelihoods.

Generalised Linear Models

- If we seek to fit a Poisson $Q \sim Poi(\lambda)$ model to a dataset of "observations" $\{q_1, q_2, q_3\} = \{1, 2, 1.5\}$ it will produce an estimate based on the mean of these being 1.5.

- In reality, of course, the likelihood of these observations is zero, since we cannot observe a Poisson count of 1.5, which is not an integer.

- The likelihood of these observations given any possible parameter is also zero and hence the proposed solution is no less likely than that any other parameters.

- Caution should therefore be used when relying on modern efficient computational packages.