Lesson 2

Generalized Linear Models

Introduction

- In the first lecture we considered data which has normally distributed errors or could be made normal
- Here, we will consider some common alternative mathematical distributions that describe non-normal data
- We will also show how these non-normal distributions can be combined with normal variables using Generalized Linear Models.

Assumptions of linear models

$$y_{ijk} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon_{ijk}$$

Linear models make many assumptions, including:

1. The model makes biological sense/ physical sense
2. Additivity (terms are added together)
3. Linearity
4. Independence of errors (LATER)
5. Homoscedasticity – equal variance of errors
6. ~~Normality of errors.~~

Alternative Distributions

- Some data which is typically not normal:
  - Proportions (e.g. infection rates, survival rates)
    - Variance will be ∩ - shaped function of mean
  - Count data (insects on a leaf, trees in a plot)
    - Often many zeroes – variance will increase with mean
  - Binary response (dead / alive; present / absent)
    - Non-normal

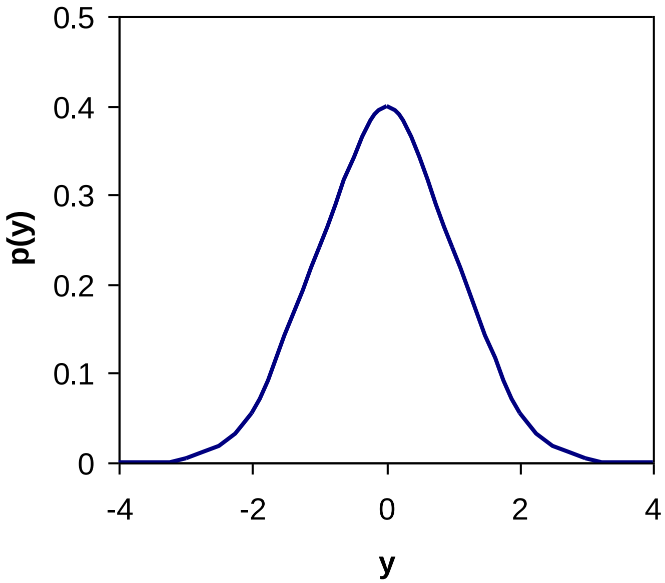
Importantly,

- These data can be described by known distributions
- As such they can be described mathematically
- And we can make inferences about them using those mathematical properties

Normal distribution

- Symmetric, continuous, unimodal, unbounded
- Used for:
  - continuous variates (usually unbounded)
  - Numerous data (growth, flow rates, mass, etc)

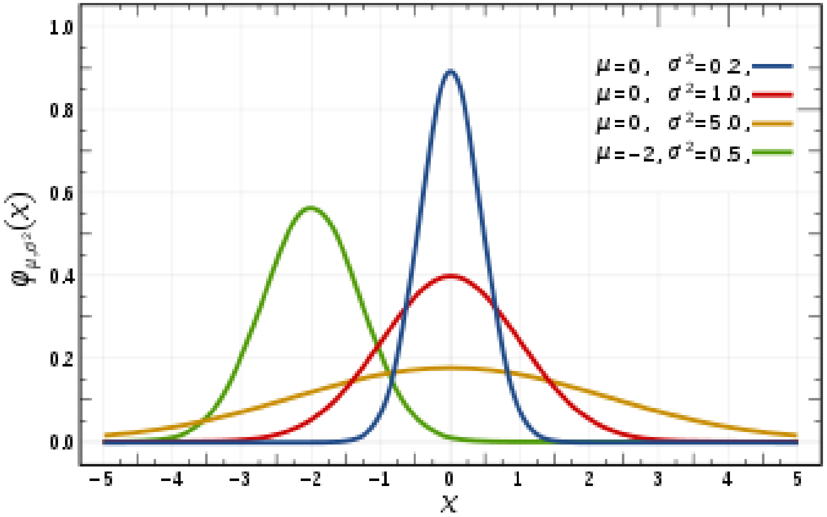
- Properties:
  - mean (μ) and variance (σ^2)
  - σ^2 constant and indept. of μ
  - Unbounded: (-∞ , ∞)



Normal distribution

- Probability density function (pdf)

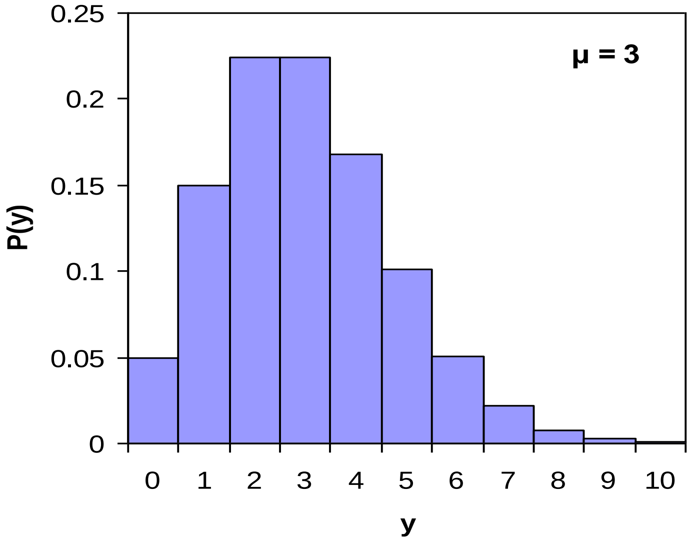
$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)}$$



Poisson distribution

- Asymmetric, discrete, unimodal, bounded below
- Used for:
  - variables describing the number of occurrences of a particular event in an interval of time or space, i.e. count data (e.g. # insects on leaves)

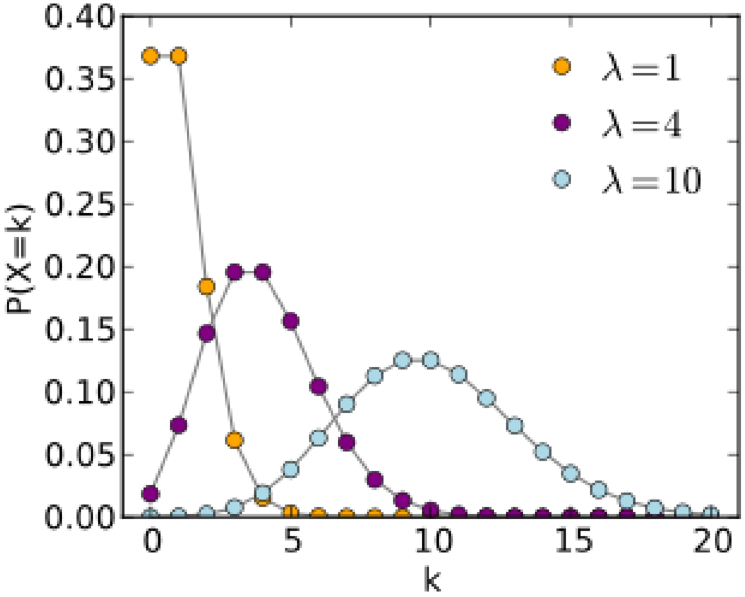
- Properties:
  - $\mu = \sigma^2$ ($=\lambda$)
  - Bounded below: [0,∞)
  - Approx. normal for $\mu >> 0$ but incr. +ve skew as $\mu \rightarrow 0$



Poisson distribution

- Probability density function (pdf)

$$f(y,\lambda) = \frac{\lambda^{y}\, e^{-\lambda}}{y!}$$

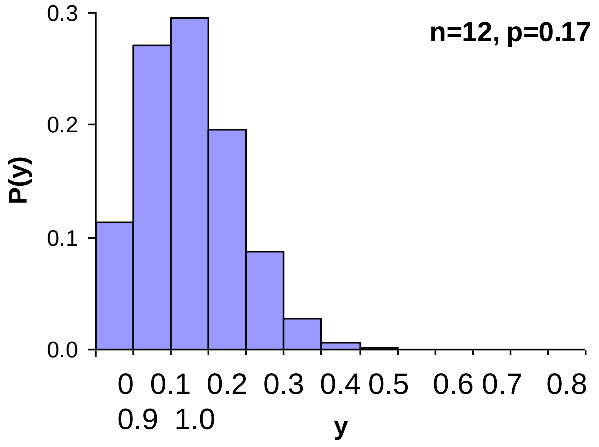


Binomial distribution

- Asymmetric, discrete, unimodal, bounded both sides
- Used for:
  - Binary data (yes/no)
  - summed proportions of binary data in a given number of trials (n trials) (binomial data)

- Properties:
  - $\mu = np$, the prob of success ('yes')

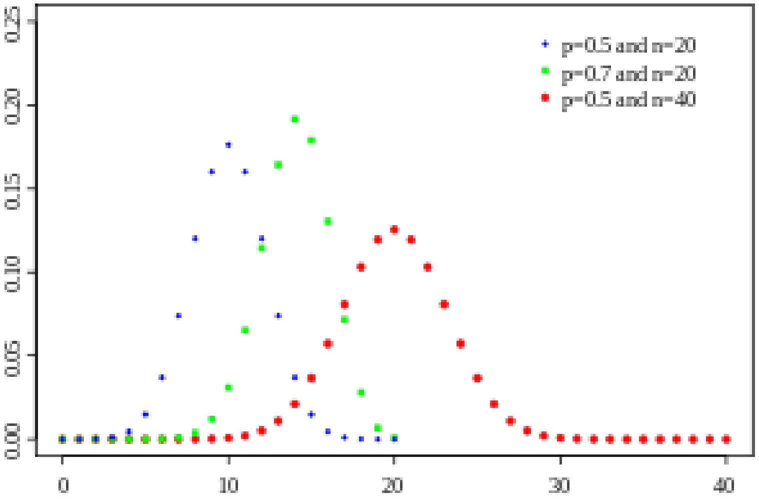
  [(1 - p) = prob. of observing failure]
  - $\sigma^2 = \mu (1 - \mu)$
  - Bounded both sides: [0,1]



Binomial distribution

- Probability density function (pdf)

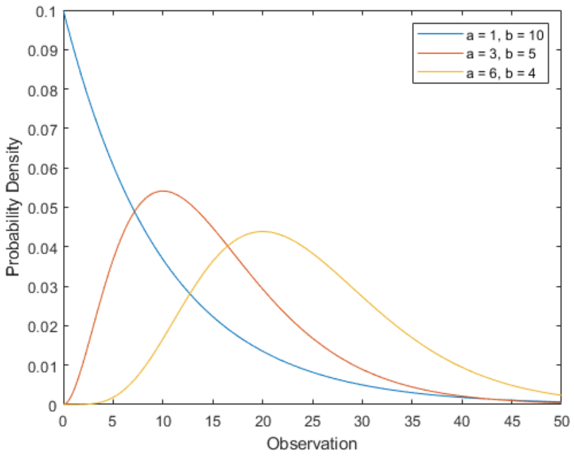
$$\mathrm{f}(n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$



Gamma distribution

- Asymmetric, continuous, unimodal, bounded below
- Used for:
  - continuous variates that are bounded below (e.g. rain)
  - zeros in poisson data (negative binomial model)
  - prior for several distributions (e.g. Normal, Poisson)

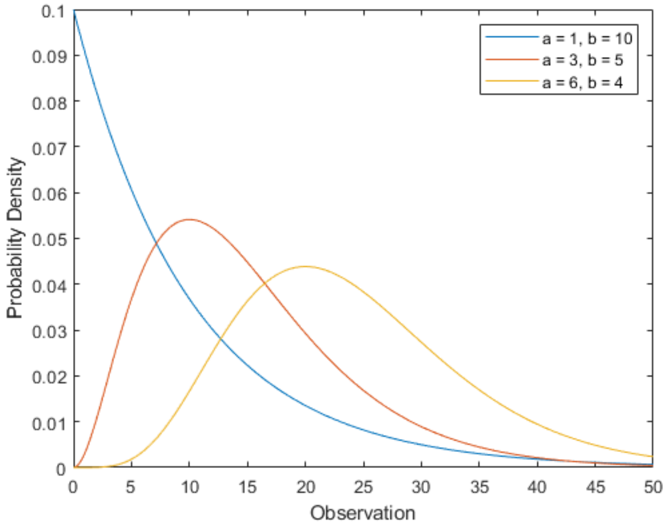
- Properties:
  - 2 parameters: α (shape), β (rate)
  - mean ($\mu = \alpha / \beta$)
  - variance ($\sigma^2 = \alpha / \beta^2$)
  - Bounded below: $(0 , \infty)$



Gamma distribution

- Probability density function (pdf)

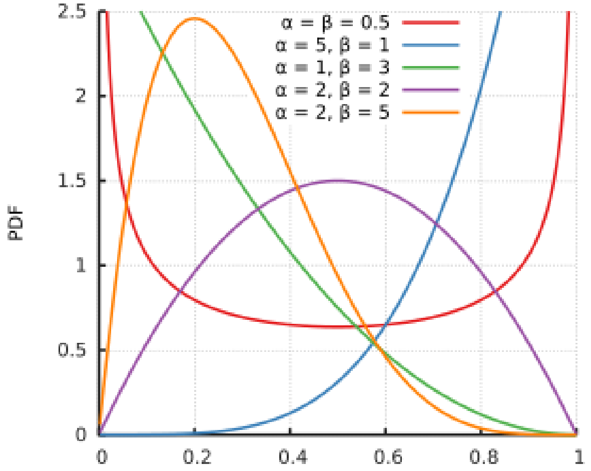
$$\mathrm{f}(\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma\alpha} x^{\alpha-1} e^{-\beta x} \qquad \text{where} \quad \Gamma\alpha = \int_0^{\infty} x^{z-1}\, e^{-z}\, dx$$



Beta distribution

- Asymmetric, continuous, bounded both sides
- Used for:
  - Proportional data
  - Prior for several distributions (Binomial, Neg Bin, Geometric)

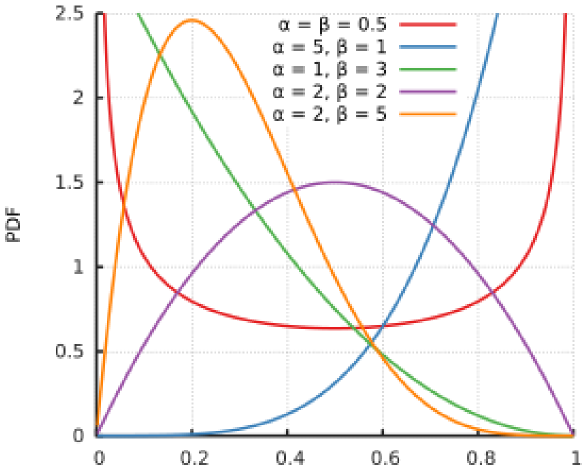
- Properties:
  - 2 shape parameters: α>0, β>0
  - $\mu = \dfrac{\alpha}{\alpha + \beta}$
  - $\hat{\sigma}^2 = \dfrac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$
  - Bounded both sides: [0,1]



Beta distribution

- Probability density function (pdf)

$$\mathrm{f}(\alpha,\beta) = \frac{x^{\alpha-1}e^{-\beta x}}{B(\alpha,\beta)} \quad \text{where} \quad B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$



Generalized linear models

Assumptions of generalized linear models

Linear models make many assumptions, including:

1. The model makes biological sense
2. Additivity (terms are added together)
3. Linearity
4. Independence of errors
5. Homoscedasticity – equal variance of errors
6. ~~Normality of errors.~~

Generalized linear models

- Often we seek to explain non-normal response data using explanatory variables just like in LMs

$$y_{ijk} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon_{ijk}$$

- If the response variable can be described with an alternative parametric distribution, then we can analyse the relationship with regression, using generalized linear models (GLMs)

A problem example:

Problem: At what age do fledgling birds start to fly in a certain population?

Response variable: birds flying or not:

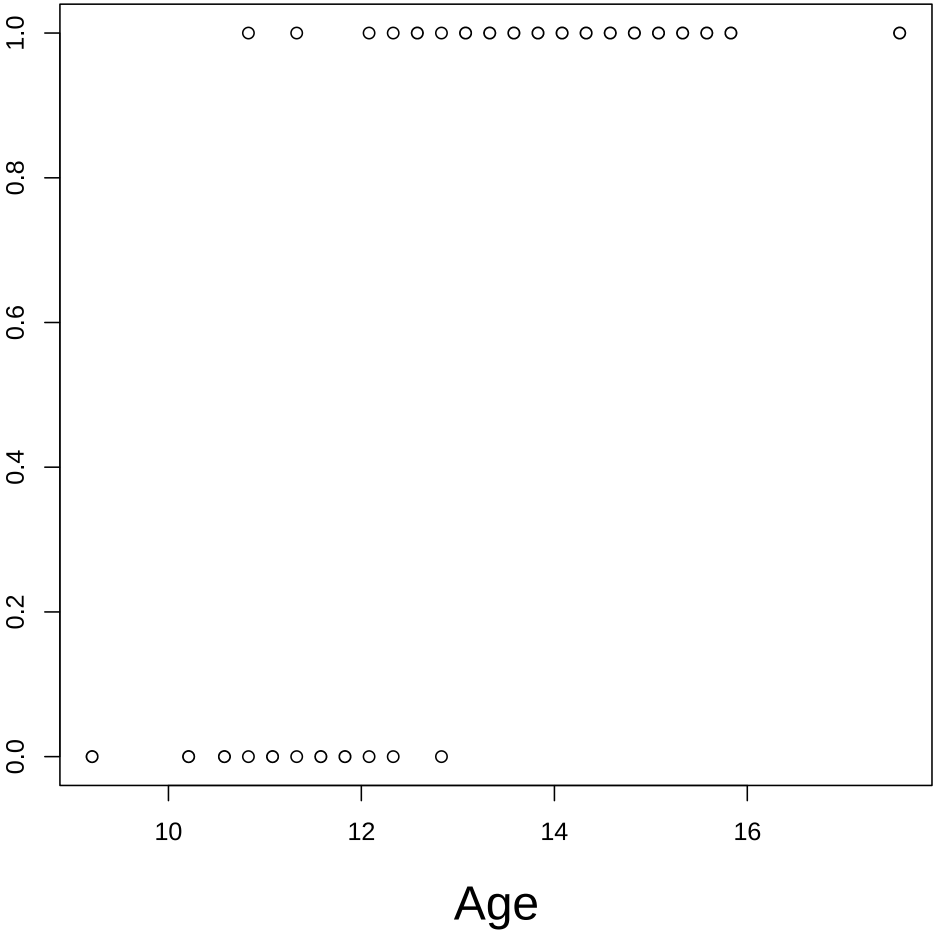
0 – no, 1- yes; -> data is binary

(summed proportion of successes -> binomial data)

Associated explanatory variable: age (days) :

-> data is ordinal

A problem example: binary response



At all low ages, no birds fly, p(x=10) = 0.

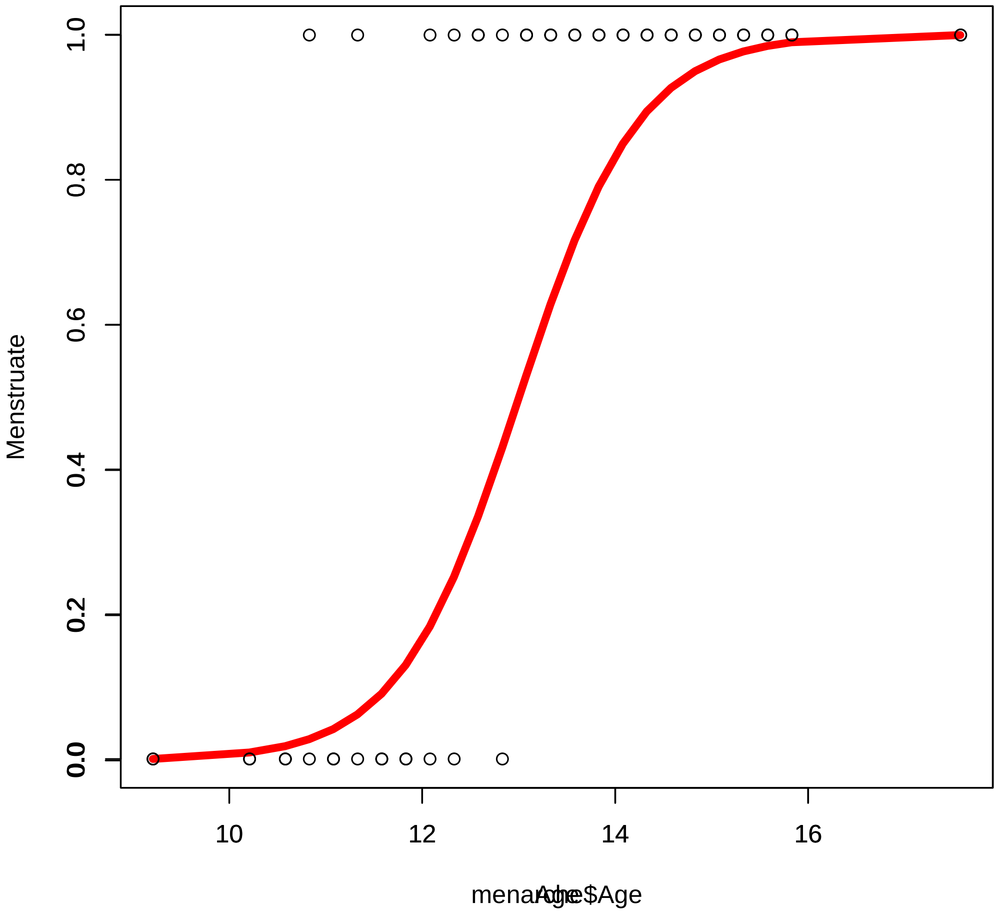
At high ages, all birds fly (=1).

So bounded [0,1]

In-between the proportions change rapidly.

When does the change happen?

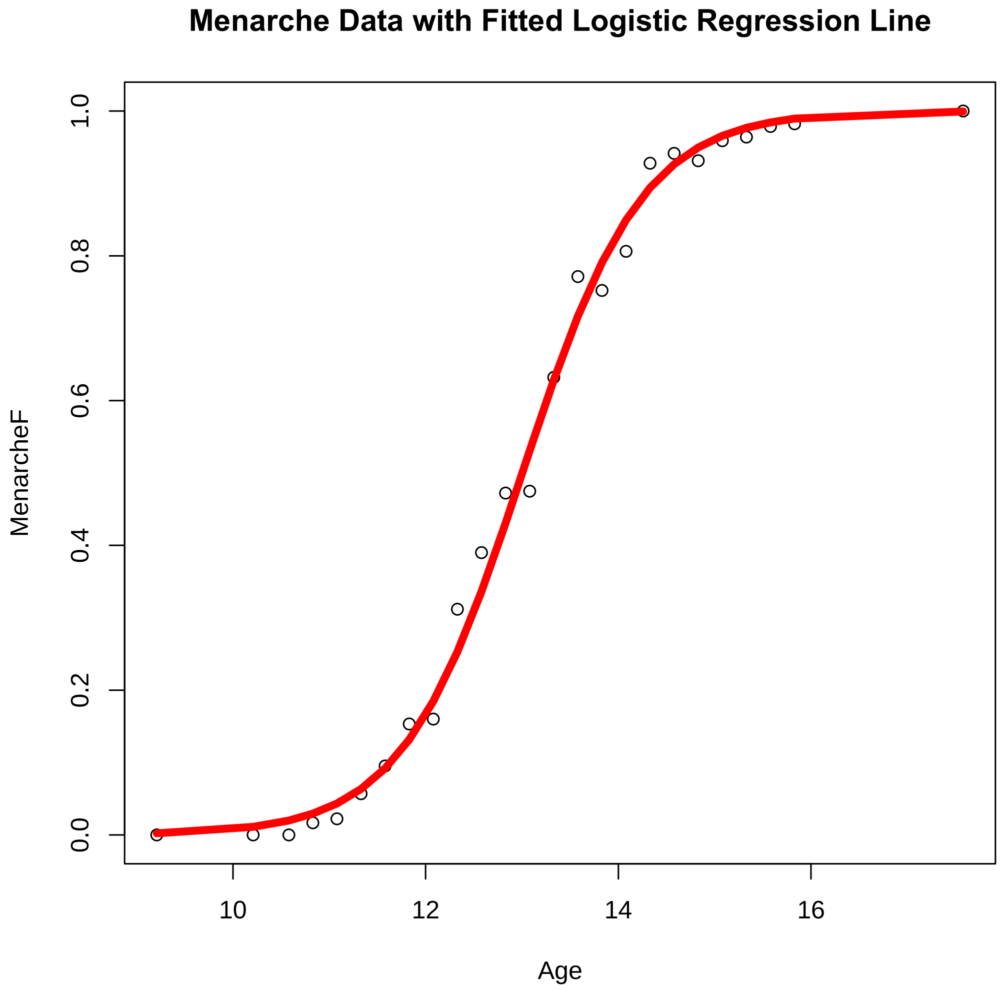
Solution:
logistic regression



Fit a cont. line that switches from 0 to 1 as the x-values change, i.e. as the proportion of “1”s (“yes”) increases.

A logistic function has this property, so we transform our Y-values to this form using the logit function. Then we can model the transformed responses using regression.

logistic regression -the process



The logit function,

$$\eta = \ln(p\ /\ (1\text{-}p))$$

is the relative chance of success (p) over failure (1-p).

The model actually bins the proportions of successes over small intervals, which shows how this is changing with age

Why not transform the response as we did before?

- You may ask yourself: why not just transform the response (Y) to make it normal and apply regular regression?

- Transformation often fails to produce data that are normally distributed with homogeneous variance ✗ (consider binary data!)

- So, we need an analysis method that accommodates non-normal error distributions in the response

Solution: Generalised Linear Models

- A family of statistical methods that unify many classical methods within a single framework.
- They cover all distributions in the exponential family (e.g. normal, Poisson, binomial, gamma, beta)
- Like general linear models, they can handle a variety of explanatory variable types:
  - variates, factors, interactions
- They allow non-normal error structures

Generalised Linear Models

- Three important components:

1. The error structure of the response variable, Y (Normal, Poisson, Binomial, etc)

2. The link function, g(): a transformation of Y that linearizes its values

$$h_i = g(y_i)$$

3. The linear predictor of predictor variables, X_j , to regress against the transformed response:

$$h = a + b_1 x_1 + b_2 x_2 ...$$

Notes: Link function, g()

- It is a transformation function that changes the response to an unbounded, linearized form, η_i , suitable for regression.

- It does not necessarily make the response normal.

- It ensures that estimated values are subject to the bounds of the original distribution of the response.

Notes: Link function, g()

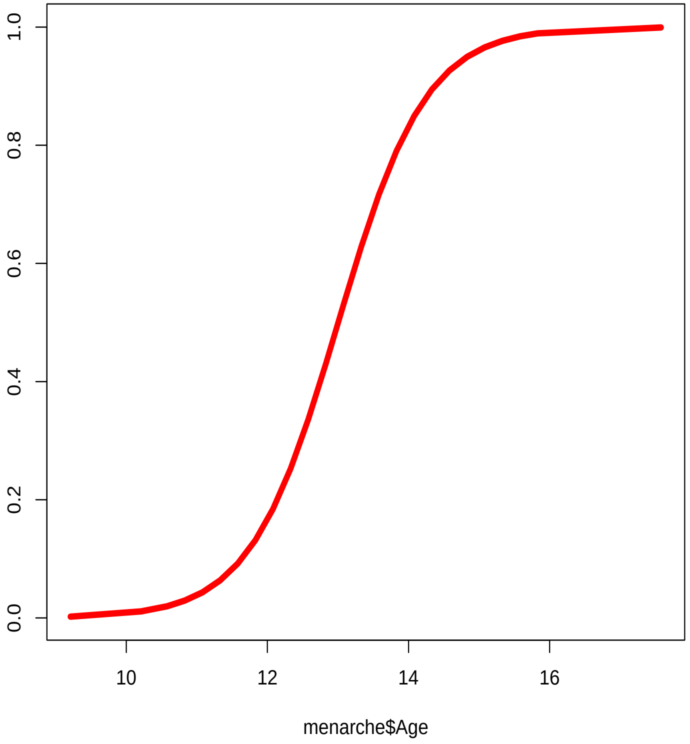
- E.g. logit transform changes binomial data to a logistic form

- X-range is (-∞ , ∞)

(unbounded, continuous)

- Y-range is [0,1]

(bounded like binomial)



Link function

- Relates *y* to the linear predictor *η*

$$\eta = g(y) = g\left[\sum x_j \beta_j\right]$$

- To calculate a predicted value of y, just use the inverse (reciprocal) of the link function.

$$\hat{y} = g^{-1}(\hat{\eta}) = g^{-1}\left[\sum x_j \beta_j\right]$$

- Models are fit using maximum likelihood estimation

Common Link functions

- The usual (canonical) link functions are:

Error	Link	Link function
Gaussian	Identity	$\eta = y$
Poisson	Log	$\eta = \ln(y)$
Binomial	Logit	$\eta = \ln(p / (1-p))$

Unlike μ, η is always unbounded (i.e. it can take any value from + ∞ to - ∞).

Example 2.1. Seal problem (Binomial data)

- A behavioural ecologist approached 50 nursing seals several times over a 2 day period and recorded the number of aggressive responses.
- She wanted to check if the age of cubs influenced the probability of an aggressive response.

Prediction

- How to use chosen models for prediction?

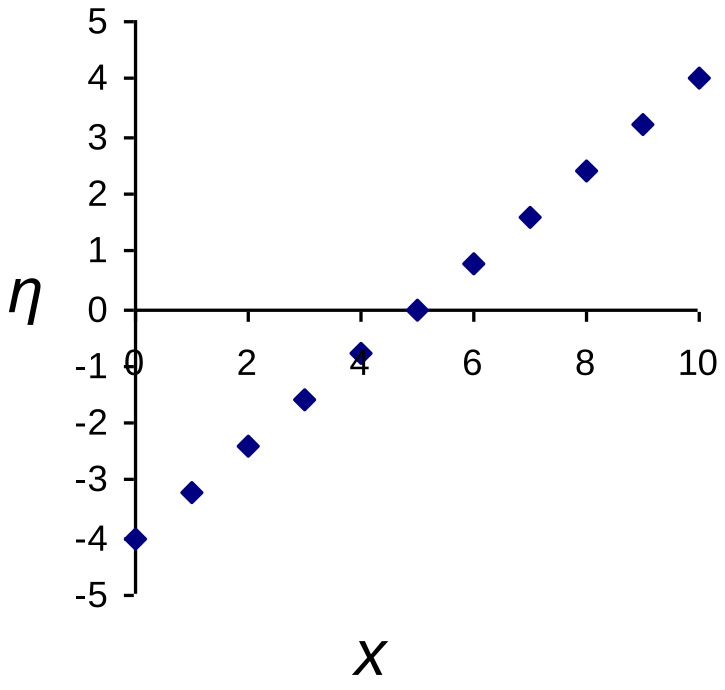
Inverse link functions

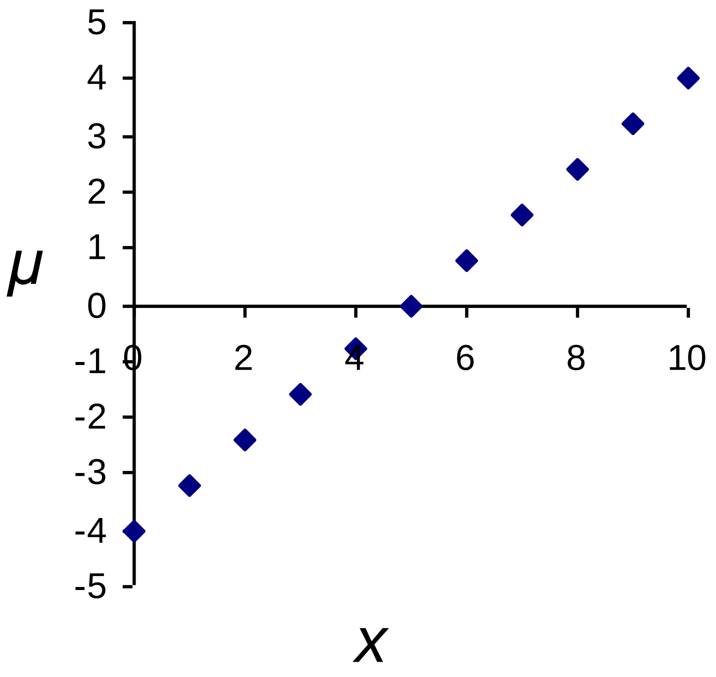
- Since the GLIM predicts η rather than y, we need an **"inverse link" function** to make quantitative predictions of μ using our model

Error	Link	Inverse Link function
Gaussian	$\eta = y$	$\hat{y} = \eta$
Poisson	$\eta = \ln(y)$	$\hat{y} = e^{\eta}$
Binomial	$\eta = \ln(p / (1-p))$	$p = e^{\eta} / (1+e^{\eta})$

Gaussian model

$\eta = \hat{y} = a + bX$ $\hat{y} = a + bX$

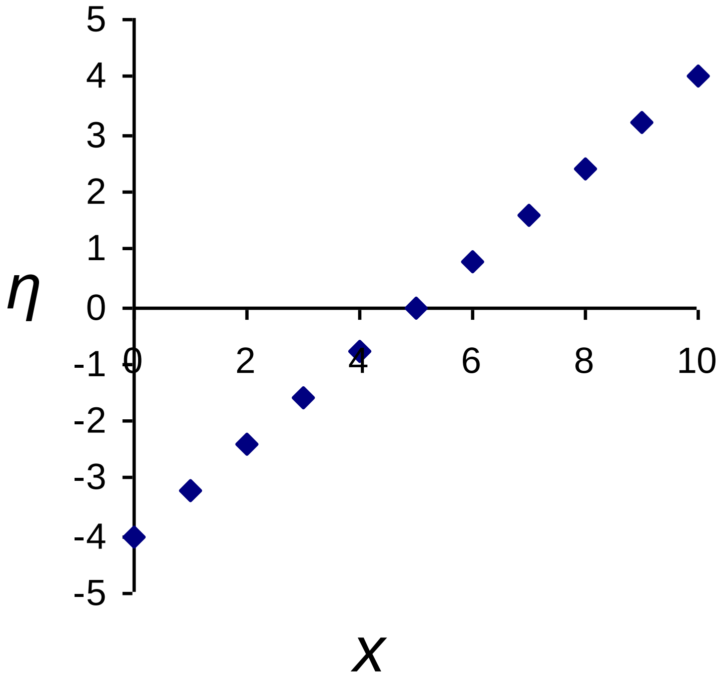


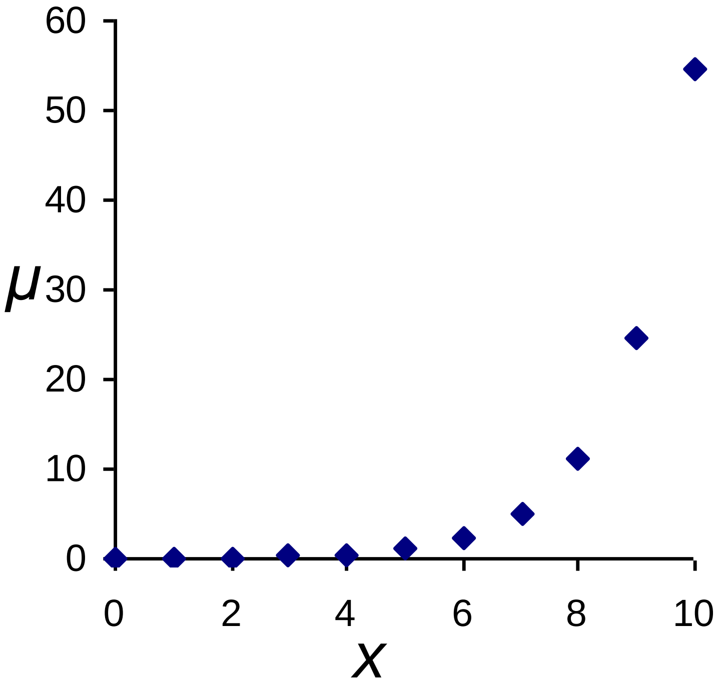


Poisson model

$\eta = \ln(y) = a + bX$

$\hat{y} = e^{a+bX}$

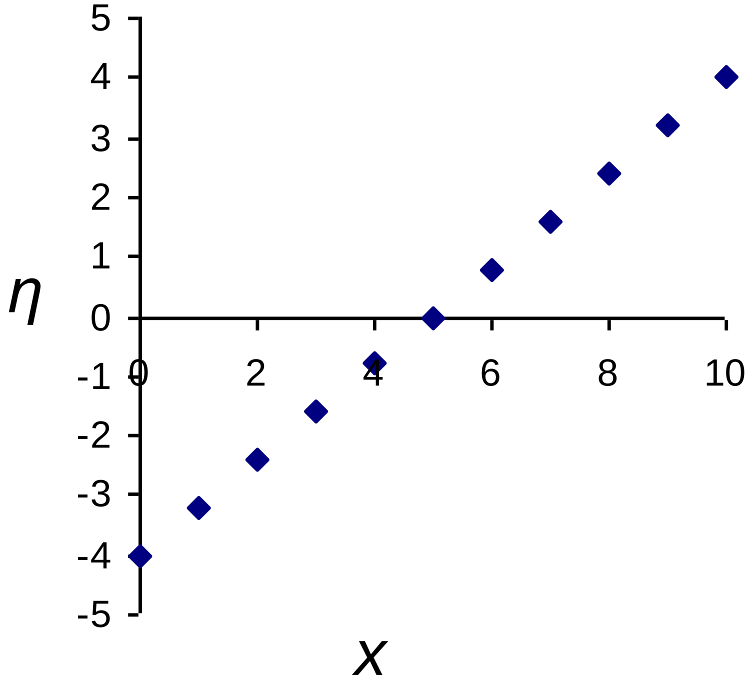


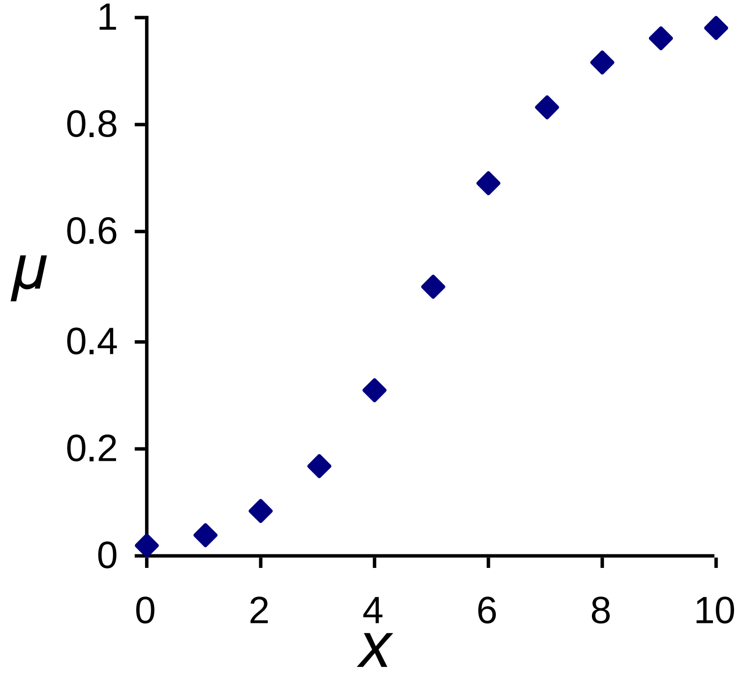


Logistic (Binomial) model

$\eta = \ln(p / (1-p)) = a + bX$

$p = e^{a+bX} / (1+e^{a+bX})$





Example 2.1 Continued.. Predictions

Maximum likelihood estimation

- Understanding how parameters are found in GLMs.

Maximum Likelihood Estimation (MLE)

- Before going further we need to introduce the concept of maximum likelihood estimation
- Maximum likelihood estimation is an alternative way of estimating the β_i's and σ^2
- In the first lecture we looked at least squares estimation:

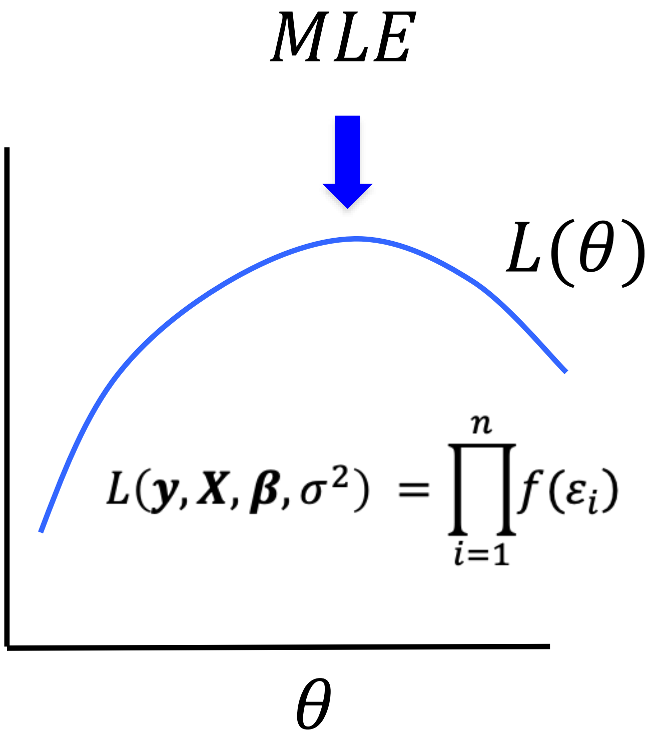
$$Min \sum_{1}^{n} (y_i - \hat{y}_i)^2 = Min \sum_{1}^{n} \varepsilon^2$$

- Least squares works for linear models with normally-distributed error; it's not efficient for alternative distributions

Maximum Likelihood Estimation (MLE)

The MLE is the most likely value of the parameters given the data

- MLE uses:
  - the data information (**y**, **X**),
  - the model formulation ($\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$)
  - a probability density function $f(e_i)$ (e.g. Normal, Binomial, Poisson)
- Then it constructs a product function for that data called a likelihood function L(θ), where θ are the parameters to be solved (e.g. β_0, β_1, σ)
- Then to solve for the L(θ), it looks for the maximum value of L(θ) (the turning point), using <u>derivatives</u> and numeric solutions

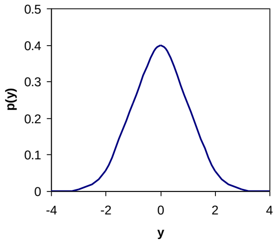


$$L(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} f(\varepsilon_i)$$

MLE example: normal data

- Choose the error distribution formula:

$$f(\varepsilon_i) = \frac{1}{\sigma\sqrt{2\pi}}\, exp\left(-\frac{1}{2\sigma^2}{\varepsilon_i}^2\right)$$



- Next, multiply this formula n times, to get the likelihood function (n is for the number of data points):

$$L(\boldsymbol{y},\boldsymbol{X},\boldsymbol{\beta},\sigma^2) = \prod_{i=1}^{n} f(\varepsilon_i) = \frac{1}{(2\pi)^{n/2}\sigma^n}\, exp\left(-\frac{1}{2\sigma^2}\boldsymbol{\varepsilon'}_{\boldsymbol{1xn}}\,\boldsymbol{\varepsilon}_{\boldsymbol{nx1}}\right)$$

- Now substitute vector ε = y – Xβ to add the actual data

$$L(\boldsymbol{y},\boldsymbol{X},\boldsymbol{\beta},\sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n}\, exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})\right)$$

MLE example: normal data

- Products are difficult to solve, so take log of L(θ)) to make additive:

$$\ln\ L(\boldsymbol{y},\boldsymbol{X},\boldsymbol{\beta},\sigma^2) = -\frac{n}{2}\ln(2\pi) - \ n\ln\sigma - \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})$$

- Now take the derivatives of unknowns to find the turning points:

$$\frac{\partial}{\partial\theta}\left(\ln\ L(\boldsymbol{y},\boldsymbol{X},\boldsymbol{\beta},\sigma^2)\right) = 0$$

for $\theta = \beta_i$, σ^2

Turning points represent the maximum likelihoods, and thus the best estimates of β_i , σ^2

MLE example: normal data

$$\ln\ L(\boldsymbol{y},\boldsymbol{X},\boldsymbol{\beta},\sigma^2) = -\frac{n}{2}\ln(2\pi) - n\ln\sigma - \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})$$

- Now solve for β_i :

$$0 = \frac{\partial}{\partial\boldsymbol{\beta}}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})$$

$$0 = \frac{\partial}{\partial\boldsymbol{\beta}}(\boldsymbol{y}'\boldsymbol{y}-\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y}-\boldsymbol{X\beta}\boldsymbol{y}' + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X\beta})$$

$$0 = \frac{\partial}{\partial\boldsymbol{\beta}}\left(\boldsymbol{y}'\boldsymbol{y}-\boldsymbol{2\beta}'\boldsymbol{X}'\boldsymbol{y} \quad + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X\beta}\right)$$

$$0 = -\boldsymbol{2X}'\boldsymbol{y} \quad + \boldsymbol{2\beta}'\boldsymbol{X}'\boldsymbol{X}$$

- Thus:

$$\widehat{\boldsymbol{\beta}} = \ (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

(this is the same as the least squares estimate in Lesson 1)

MLE example: normal data

$$\ln \ L(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \ n \ln \sigma - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})$$

- For σ^2, the MLE generates a downward biased estimate

$$\tilde{\sigma}^2 = \frac{(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})}{n}$$

- For normal data (a closed form likelihood), this is easily converted to an unbiased form using algebra:

$$\hat{\sigma}^2 \ = \ \tilde{\sigma}^2 \ [n/(n-p)]$$

- **Note**: Restricted maximum likelihood (REML) generates unbiased estimates of σ^2 for more difficult likelihoods that are not closed form.

Further points about MLEs:

MLEs have better statistical properties than LSEs or other estimators.

1. MLEs are BLUP – Best Linear Unbiased Predictors of β_i

2. MLEs have lower variance when compared to other unbiased predictors (e.g. RMLEs) i.e. can underestimate σ^2

3. MLEs are consistent => they converge to true population parameters with large n

4. MLEs are sufficient => they contain all the information in the sample

MLEs :

We demonstrated how to derive matrix-based formulae for MLE estimates for models with normal errors.

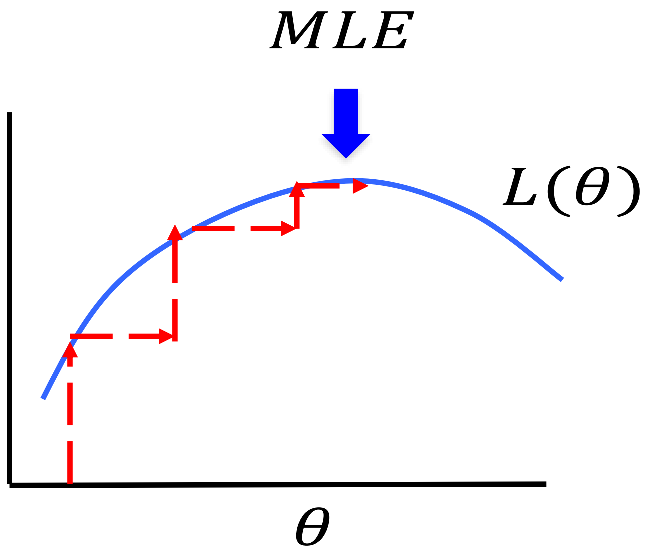
Of course there exist similar formulae for MLE estimates for other types of distributions (Binomial, Poisson, Gamma etc)

MLEs are the standard ways that more complex error distributions get handled.

They are also important to us for evaluating model performance (coming soon)

MLEs :

For simple models, exact differential solutions are possible. For more complex models, this is not possible, and turning points are found using numerical searching.



Inference

- How to evaluate whether our model fits the data well?
- How to evaluate whether all our predictors are useful for the model?
- Because the residuals ($e_i = y_i - \hat{y}_i$) are not normal,
  - we cannot use R^2 as a goodness-of-fit measure,
  - we cannot apply the usual diagnostic tests to evaluate residuals
  - We cannot use t-tests or ANOVA to evaluate model parameters
- We must use other tests that don’t assume this

How to make inferences?

- Mean comparisons (coefficient significance)
  - Wald Z-statistic tests (glm equivalent of t-test)
- Variance analysis (model significance)
  - Deviance tests
  - AIC comparisons

Wald Z test

- Test based on the ratio of the mean estimate divided by its standard error

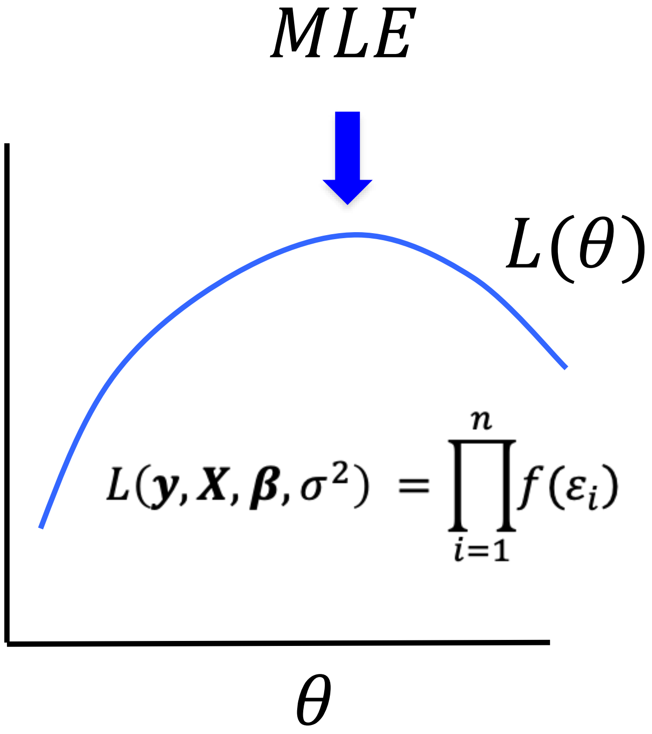
```
Coefficients:
            Estimate  Std. Error   z value   Pr(>|z|)
(Intercept)  3.51254   0.03153    111.41    <2e-16 ***
trtI        -2.04620   0.09320.   -21.95    <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- This ratio is ~ normal and follows a fixed normal distribution known as the Z-distribution (recall t-distributions change shape based on error df)

Deviance

- The method uses maximum likelihood estimation

- IMPORTANTLY: if a predictor explains data well, then it INCREASES the LIKELIHOOD



$$L(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} f(\varepsilon_i)$$

Deviance

Three models:

- Null model: minimal information explained

  y ~ 1 (no predictors)

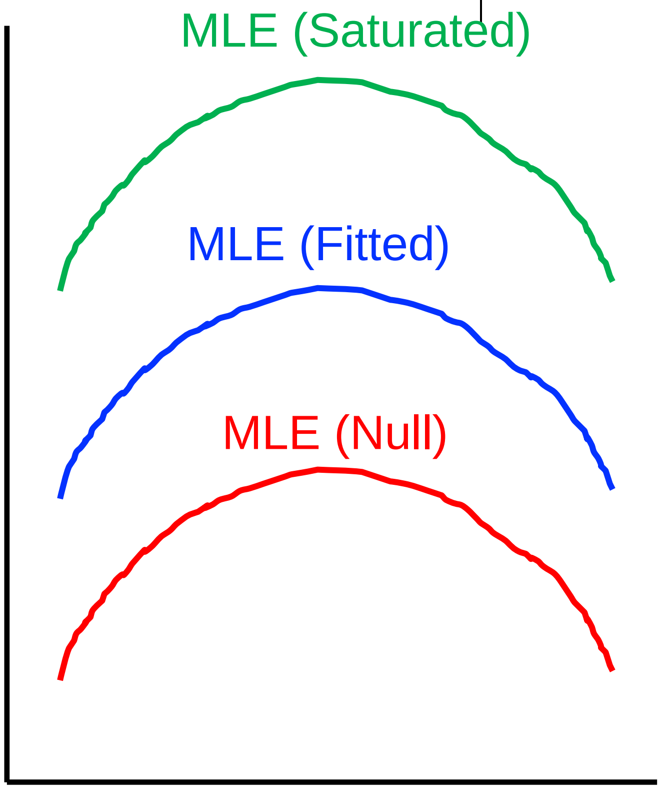
- Fitted model: our prediction model

  y ~ x1 + x2

- Saturated model: maximal information explained

  y~ x1 + x2 + x3 +….+ xn

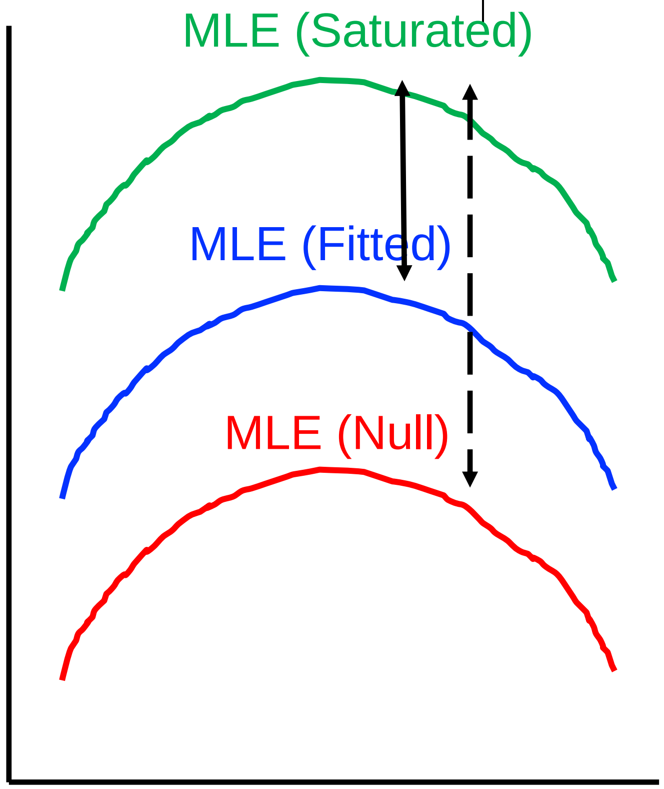
  (one parameter for each data pt)



Deviance

- Deviance is difference between likelihoods and the saturated Lik.
    - E.g. fitted model deviance

$$Deviance = -2\,ln\frac{L(saturated\ model)}{L(fitted\ model)}$$



- To test significance we compare nested deviances using an LRT

$$Likelihood\ ratio\ \ test = \frac{Residual\ deviance\ (tested\ model)}{Null\ deviance\ (null\ model)}, \qquad \sim \chi^2_{p-1}$$

Deviance in summary()

```
> summary(mod1)

Call:  glm(formula = n.aphids ~ trt, family = poisson, data = aphid2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.6185  -1.4240  -0.1273   1.4063   4.5301

Coefficients:
                Estimate  Std. Error   z value   Pr(>|z|)
(Intercept)  3.51254    0.03153    111.41   <2e-16 ***
trtI            -2.04620    0.09320.    -21.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1178.83  on 59  degrees of freedom
Residual deviance:  412.14  on 58  degrees of freedom
AIC: 653.93
Number of Fisher Scoring iterations: 5
```

Likelihood ratio tests on nested models

```
> #fitted model
> mod1 <- glm(n.aphids~trt, data=aphid2, family=poisson)
> #null model
> mod0 <- glm(n.aphids~1, data=aphid2, family=poisson)
>
> #perform likelihood ratio test
> anova(mod1,mod0,test='Chisq')
Analysis of Deviance Table

Model 1: n.aphids ~ trt
Model 2: n.aphids ~ 1
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        58     412.14
2        59    1178.83 -1   -766.7 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Example 2.1 Continued.. Deviance

Residual properties:

- Using GLMs means that we assume certain residual properties no longer apply:
- Most importantly, the residuals are not necessarily normal
- We cannot use the usual residual tests applied to LMs
- What we need to test is whether our models predict the response data well

Residual properties: Over- and Under-dispersion

- Definition: variance is greater / less than expected based on a particular statistical model

  (found in binomial and Poisson models)

- Diagnosis:

$$\text{Underdispersed} \quad \frac{Deviance}{Error\ df} \ll 1 \ll \frac{Deviance}{Error\ df} \quad \text{Overdispersed}$$

- Can result from:
  - Patchiness (heterogeneity) in the population variable
  - An unmeasured factor
  - A misspecified error distribution

Over- and Under-dispersion: Why do we care?

- Recall a Wald test:

$$Wald\ test: \frac{\beta_p}{s.e.} \quad \sim Z(0, \sigma)$$

- With overdispersion, fitted model generates overly small standard errors (SEs) for estimated coefficients, so things become a lot more significant thatn they should be

  -> type I error increases

- With under dispersion, fitted model generates overly large SEs, so things become a lot less significant than they should be

  -> type II error increases

https://www.r-bloggers.com/2017/06/what-does-under-or-over-dispersion-look-like/

Dispersion test code

- Very simple to apply in R

```
> chisq <- sum(resid(mod1, type='pearson')^2)
➢ chisq/df.residual(mod1)
➢ [1] 5.85938
➢ ## much greater than 1
```

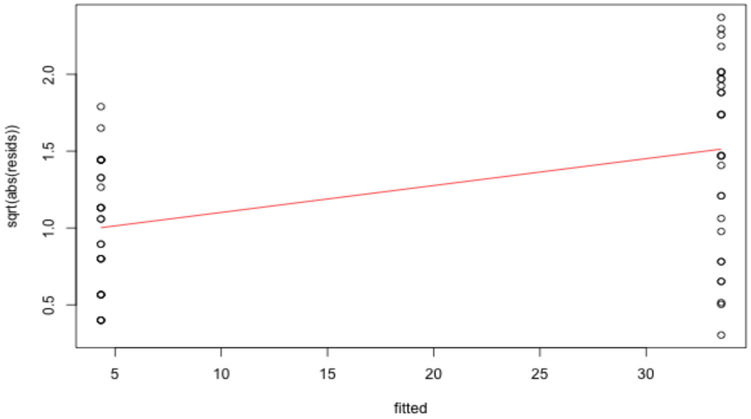
- Alternatively, use ‘dispersiontest’ from the package ‘AER’ for a formal probability test

```
➢  dispersiontest(glm1,alternative = "less")  #is underdispersed?
```

Dispersion test visually

- We can also check the homogeneity plot

```
resids <- resid(mod1, type='pearson')
fitted <- fitted(mod1)
plot(sqrt(abs(resids))~ fitted)
lines(lowess(sqrt(abs(resids))~fitted), col='red')
```



- Here you can see that the average standardised residual is >1 for the one group

Dealing with dispersion problems

- 1. Use Quasi distributions (hatchet fix)
  - Quasipoisson and Quasibinomial
  - [these don't generate real likelihoods]

- 2. Try alternative proper distributions [likelihoods]
  - Think more about the properties of the model?
  - Pure variance inflation: Generalised Poisson models, Negative binomial models; GLMMs with observation level random effects
  - Mixed process distributions: zero-inflation models; hurdle models

Quasi-likelihood

- **Solution:** allow the variance to have a multiplicative dispersion factor, ϕ

1. Binomial: $\mathrm{Var}(y) = \phi\mu(1-\mu)$
2. Poisson: $\mathrm{Var}(y) = \phi\mu$

- Models fit using maximum-likelihood approach & parameter estimates don’t change; residual errors are posthoc rescaled

- In R, we specify quasi likelihood in the family command, e.g., `family=quasipoisson` and then when we conduct the analysis of deviance using an F ratio instead of χ^2 .

Negative binomial

- Recommended for overdispersion
- Recall that in a Poisson model, var = mean.
- This seriously constrains the shape of the distribution

- NB is a compound distribution that allows the value of lambda (and thus the poisson rate model) to change according to a gamma distribution

$$\mathrm{f}(x, \lambda) = \int_0^\infty \frac{\lambda^x\, e^{-\lambda}}{x!}\ .f_{Gamma\left(r, \frac{1-p}{p}\right)}(\lambda)\, d\lambda$$

- This model allows the variance estimate to have a more complex relationship with the mean: $\sigma^2 = \mu + \mu^2/k$

Conway-Maxwell Poisson

- CMP distribution recommended for under-dispersion
- https://en.wikipedia.org/wiki/Conway%E2%80%93Maxwell%E2%80%93Poisson_distribution
- Poisson pdf: $$f(x, \lambda) = \frac{\lambda^{x} e^{-\lambda}}{x!} = \frac{\lambda^{x}}{x!} \boxed{\frac{1}{e^{\lambda}}}$$ ⍰ Normalising constant
- CMP pdf: $$f(x, \quad , v) = \frac{\lambda^{x}}{(x!)^{v}} \boxed{\frac{1}{Z(\lambda, v)}}$$

  $$\text{where: } Z(\lambda, v) = \sum_{j=0}^{\infty} \frac{\lambda^{j}}{(j!)^{v}}$$
- The v parameter adjusts the rate of decay

NegBin and CMP

- NB and CMP are parametric and can be solved with numerical ML
- CMP is directly related to Poisson because when v=1, the model reduces to the Poisson.

  (because $\mathrm{Z}(\lambda, v) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^v} = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^1} = e^{\lambda}$)
- CMP can be applied using the glm.cmp function in the ‘COMPoissonReg’ package
- negbin can be applied using the glm.nb function in the ‘MASS’ package

Example 2.2. Aphid data (Poisson with overdispersion)

A researcher is interested whether an insecticide is useful in reducing aphid infestations. He applies two treatments (control, insecticide). Then, after a fixed number of days, he counts the number of aphids per leaf area. He also records how long it takes him to check the leaves for aphids, which differs per individual

- Open AphidData2.csv
- Model aphid counts as a function of treatment only
- Plot diagnostics and test for overdispersion
- Deal with the overdispersion using pseudopoisson distribution

Comparing models with different distributions – can we use AIC?

- This is a question that causes some confusion:
- If I think my model is under-/over-dispersed, can I use AIC to compare the model efficiency between the simple model (e.g. Poisson) and the more complex model (e.g. NB)?

- My reading of the literature is that this could be taken in a philosophical sense:
- EITHER: Yes, its possible for closely related models
- OR: No, don’t do it because by definition the model which fixes overdispersion must be preferred.

Comparing models with different distributions – can we use AIC?

- certainly true that many statisticians seem to have no problem doing this:
- https://math.usu.edu/jrstevens/biostat/PoissonNB.pdf # compares Negative Binomial and Poisson
- https://stats.idre.ucla.edu/stata/faq/how-can-i-use-countfit-in-choosing-a-count-model/ #compares a whole bunch of count data variations on the basic Poisson model
-
- I myself have used AIC comparisons between closely related models before: https://www.nature.com/articles/s41598-017-09768-z

Argument against AIC comparisons

- Intuitively, the argument against doing AIC comparisons between different models is that the AIC depends on the loglikelihoods of the models [recall: AIC = -2 logLik + 2(k+1)], and the loglikelihoods themselves will by definition differ because of the differences in the model functions, regardless of the data, so the AIC values generated cannot be meaningfully compared.

- The two links below follow this train of thought.
- https://stats.stackexchange.com/questions/139201/model-selection-can-i-compare-the-aic-from-models-of-count-data-between-linear

- https://stats.stackexchange.com/questions/345069/likelihood-comparable-across-different-distributions

Argument for AIC comparisons

- Closely related models are all compound functions that are products of a basic model function and anither function that adjusts the shape. By definition, the compound function will only yield different (and strictly greater) logLik than the basic function when the additional terms significantly improve the fit of the model. If the additional terms do not improve the fit, then the logLik of the basic and compound functions will be the same.
- Consider the relationship between the Negative Binomial model and the Poisson model. The Negative Binomial consists of the basic Poisson model multiplied by a gamma distribution.

$$f(k;r,p)=\int_0^\infty f_{\mathrm{Poisson}(\lambda)}(k)\cdot f_{\mathrm{Gamma}\left(r,\,\frac{1-p}{p}\right)}(\lambda)\;d\lambda$$

Other methods to compare?

- Some statisticians simply compare models using the Pearson chisq dispersion statistic for each model (this is the value we calculated to evaluate overdispersion) and a visual evaluation of the ‘homogeneity plot” (scaled abs residuals vs fitted values). See:
- https://www.theanalysisfactor.com/poisson-or-negative-binomial-using-count-model-diagnostics-to-select-a-model/

- You will see that the residuals themselves are not homogeneous, but they are all much reduced in magnitude with the Negative binomial model.

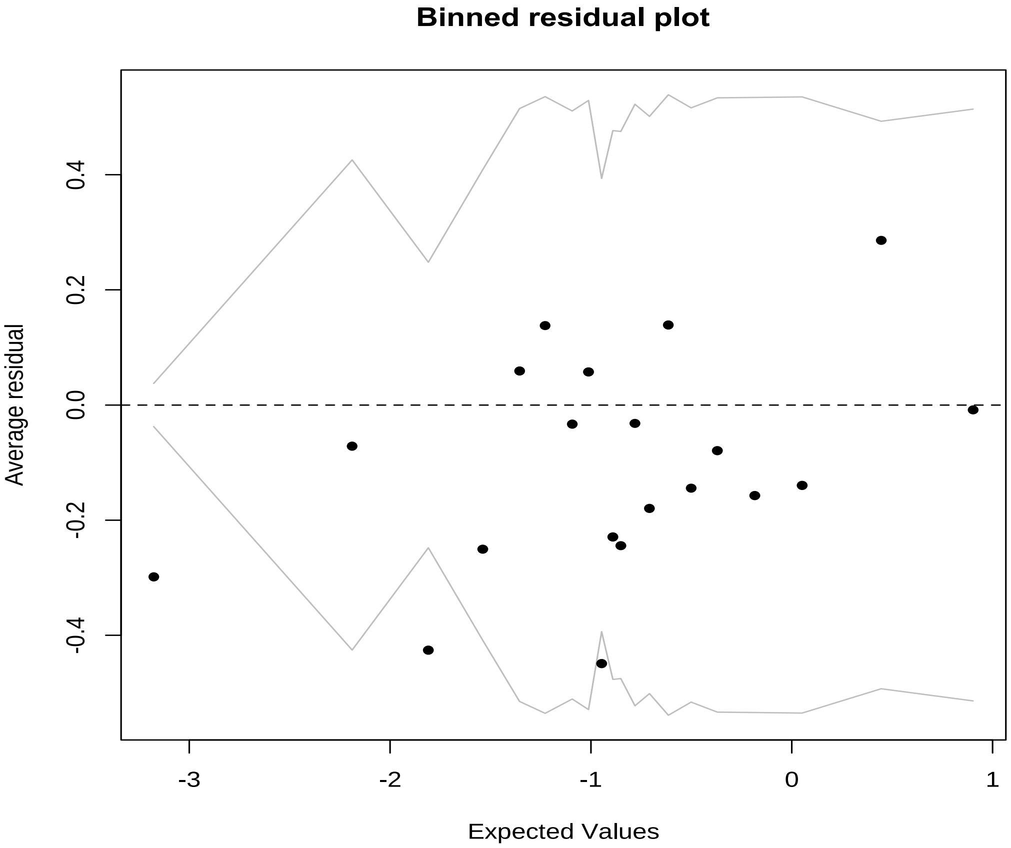
Testing for overdispersion in binary data

This is not a simple problem because the data are 0’s and 1’s.

The overdispersion test can be used, but we can also use something called a binned plot.

Residuals divided into ranges of expected values (bins), then average resids created and compared to 95%CIs created per bin.

If model is true, then 95% of the mean residuals should lie in the 95% CIs



Exercise 2.3.
Graduate admissions

A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and rank (prestige of the undergraduate institution) affect admission into graduate school. The response variable, "admit" is a binary variable (admit/don't admit).

- Open binary.csv
- Make an appropriate model and run it
- Plot residual diagnostics and test for overdispersion
- Test whether it explains significant variation (you will need to run the anova() statement comparing fitted and null models

Assessing predictors/ model selection

When you have a lot of predictors in a GLM

$$y_{ijk} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon_{ijk}$$

How do you decide which predictors are important?

[i.e. How to choose best models?]

Model selection Methods

1. Consider the significance of the regression estimate in the linear model

- These are based on Wald tests $\sim C^{2}_{1df}$

2. Compare nested models using the deviance ratio tests

$$LRT = \frac{Residual\ deviance\ (tested\ model)}{Null\ deviance\ (null\ model)}, \qquad \sim \chi^{2}_{p-a}$$

3. Compare a larger group of subset models using AICc, like in information theoretic approaches

Exercise: Return to Admissions problem

Assumptions of a GLM

Linear models assumptions:

1. The model makes biological sense
2. Additivity (terms are added together)
3. Linearity
4. Independence of errors
5. Homoscedasticity – equal variance of errors
6. ~~Normality of errors.~~

End of Lecture 2.

Optional: Exercise 2.4. Children ever born

Question: what factors affect the number of children born to women in Fiji? The dataset ceb.csv contains:

-> Predictors: marriage duration (fact.), residence (fact.), education (fact.)

-> Response: mean number of children ever born per categorical group

-> Number of women sampled per group

- Make an additive model using all predictors and test whether they are all necessary using model selection
- Plot residual diagnostics and test for overdispersion
- Write out the back-transformed model