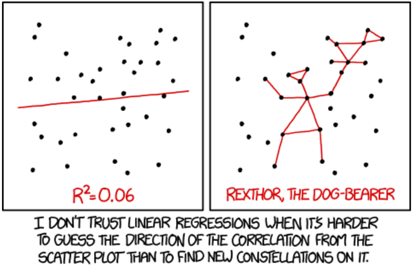
Generalised linear models

Bert van der Veen

Department of Mathematical Sciences, NTNU

So far: linear models

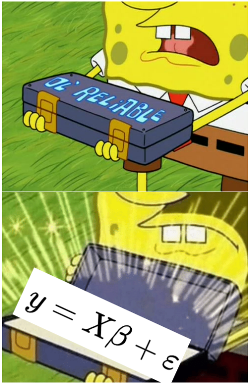
$$y_i = \alpha + \sum_{k=1}^{p} x_{ik}\beta + \epsilon_i \qquad (1)$$


R²=0.06
REXTHOR, THE DOG-BEARER
I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

So far: linear models

$$y_i = \alpha + x_i\beta + \epsilon_i \tag{2}$$

- y_i the data
- $\alpha + x_i\beta$ the systematic component: "linear predictor"
- ϵ_i the random component: "error"



Key assumptions

- Linearity (straight line)
- Independence of errors
- Homoscedasticity (same variance for all errors)
- Normality (distribution of errors)

Real data do not usually follow these assumptions

and don't range $(-\infty, \infty)$

Generalised linear models (GLMs)

GLMs as a framework were introduced by Nelder and Wedderburn (1972) uniting many different models. With a special focus on teaching statistics.

- Linear regression
- Logistic regression
- Probit regression
- Complementary log-log regression
- Log-linear regression
- Gamma regression

Generalised linear models (2)

GLMs extend the linear model framework to address:

- Variance changes with the mean
- Range of **y** is bounded

The basis of many statistical models

Many results are now asymptotic (via the normal distribution)

Components of a GLM

- Systematic component: η
- Random component: data/distribution
- The link function: connects these components
  - This is not a data transformation
- The variance function

No explicit error term

GLM Likelihood

- We still use maximum likelihood for estimation
- But now a different likelihood function (in EF for fixed ϕ)

All GLMs can be formulated in terms of EF:

$$\mathcal{L}(y_i;\Theta) = \exp\{\frac{y_i\eta_i - b(\eta_i)}{a(\phi)} + c(y_i,\phi)\} \tag{3}$$

Linear regression as EF

Previously:

$$\mathcal{L}(y_i;\Theta) = \frac{1}{\sigma\sqrt{2\pi}}\exp\{-\frac{1}{2}\frac{(y_i-\mu_i)^2}{\sigma^2}\} \tag{4}$$

Linear regression as EF

Previously:

$$\mathcal{L}(y_i;\Theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2}\frac{(y_i - \mu_i)^2}{\sigma^2}\} \tag{4}$$

Now:

$$\mathcal{L}(y_i;\Theta) = \mathsf{exp}\{\frac{y_i\eta_i - b(\eta_i)}{a(\phi)} + c(y_i,\phi)\} \tag{5}$$

Linear regression as EF

Previously:

$$\mathcal{L}(y_i;\Theta) = \frac{1}{\sigma\sqrt{2\pi}}\exp\{-\frac{1}{2}\frac{(y_i - \mu_i)^2}{\sigma^2}\} \qquad (4)$$

Now:

$$\mathcal{L}(y_i;\Theta) = \text{exp}\{\frac{y_i\eta_i - b(\eta_i)}{a(\phi)} + c(y_i,\phi)\} \qquad (5)$$

- for some known functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$
- for normal distribution: $\eta_i = \mu_i$, $a(\phi) = \sigma^2$, $b(\eta_i) = -\mu^2/2$, $c(y_i,\phi) = -\frac{1}{2}\{y^2/\sigma^2 + \log(2\pi\sigma^2)\}$

The linear model

Previously we wrote the linear model as:

$$y_i = \alpha + \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{6}$$

The linear model

Previously we wrote the linear model as:

$$y_i = \alpha + \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{6}$$

This is the same as:

$$\mathbb{E}(y_i | \mathbf{x}_i) = \alpha + \mathbf{x}_i \boldsymbol{\beta} \tag{7}$$

The linear model

Previously we wrote the linear model as:

$$y_i = \alpha + \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{6}$$

This is the same as:

$$\mathbb{E}(y_i | \mathbf{x}_i) = \alpha + \mathbf{x}_i \boldsymbol{\beta} \tag{7}$$

when $\mathbb{E}(\epsilon_i) = 0$.

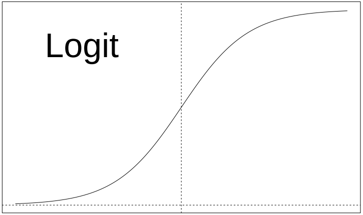
Generalised linear model

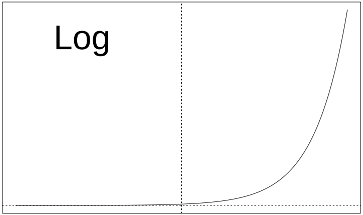
$$\begin{aligned} g\{\mathbb{E}(y_i|x_i)\} &= \eta_i = \alpha + x_i\beta \\ \mathbb{E}(y_i|x_i) &= g^{-1}(\eta_i) = g^{-1}(\alpha + x_i\beta) \end{aligned} \tag{8}$$

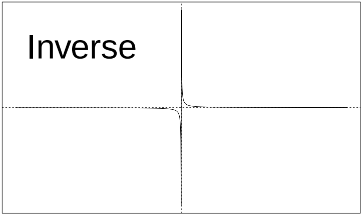
g$(\cdot)$ is the **link function**
g$^{-1}(\cdot)$ is the **inverse link function**

The link function

- Is a smooth/monotone function
- Has an inverse $g^{-1}(\cdot)$
- Restricts the scale
- g$(\cdot)$ can be e.g.







Variance function

Variance changes with the mean:

$$\mathsf{var}(y_i; \mu_i, \phi) = \frac{\partial^2 g(\eta_i)}{\partial \eta_i^2} a(\phi)$$

- ϕ: the dispersion parameter, constant over observations
  - Fixed for some response distributions
- $a(\phi)$ is a function of the form ϕ / w_i (McCullagh and Nelder 1989)

Assumptions

- No outliers
- Independence
- Correct distribution
- Correct link function
- Correct variance function (implied by previous two)

More on checking assumptions in GLMs tomorrow.

Fitting GLMs

Unlike LMs, parameters in GLMs need to be estimated **iteratively**.

- More difficult to fit
- Requires numerical *optimisation*
- Susceptible to local convergence

Popularisation of GLMs

Nelder and Wedderburn (1972) proposed GLMs as a class to unify different forms of regression.

- Linear regression
- Probit regression
- Logistic regression
- Log-linear regression
- Gamma regression
- Inverse Gaussian regression

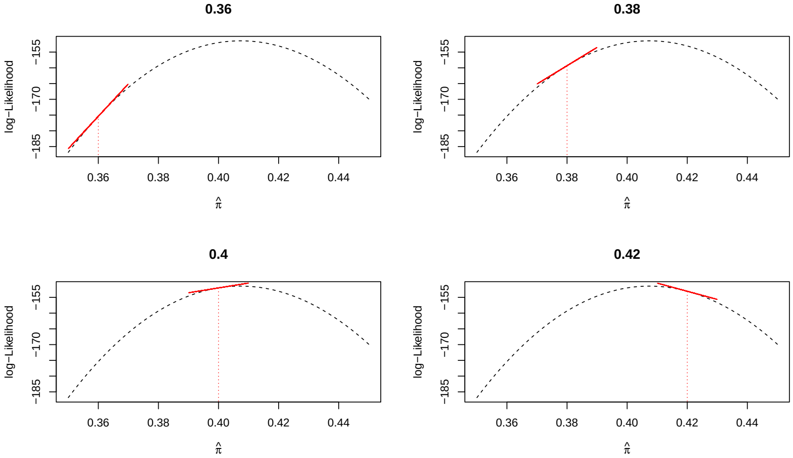
McCullagh and Nelder (1989) wrote a book that popularised the class.

Fitting GLMs

Easy and quick in `R`.

Mathematically more involved than LMs.

Finding the maximum (from day 1)



We need a good algorithm to find the maximum!

Finding the maximum: GLMs

$$\Theta^{t+1} = \Theta^t + \frac{\partial^2 \log\{\mathcal{L}(\mathbf{y}_i; \Theta^t)\}}{\partial\Theta\partial\Theta^\top}^{-1} \frac{\partial\mathcal{L}(\mathbf{y}_i; \Theta^t)}{\partial\Theta}$$

- (Newton-Rhapson) Can get quite expensive to evaluate.
- Nelder and Wedderburn (1972) instead suggested an algorithm that fits a LM repeatedly.

Iteratively reweighted least squares (IRLs)

1) Start at some (decent) guess (e.g., $\hat{\boldsymbol{\eta}} = \mathbf{y} + \boldsymbol{\epsilon}$) (Wood 2017)

Iteratively reweighted least squares (IRLs)

1) Start at some (decent) guess (e.g., $\hat{\boldsymbol{\eta}} = \mathbf{y} + \boldsymbol{\epsilon}$) (Wood 2017)
2) Set $\mathbf{z}^t = \boldsymbol{\eta} + \frac{\partial \hat{\boldsymbol{\eta}}}{\partial g^{-1}(\hat{\boldsymbol{\eta}})}\{\mathbf{y} - g^{-1}(\hat{\boldsymbol{\eta}})\}/a(g^{-1}(\hat{\boldsymbol{\eta}}))$

Iteratively reweighted least squares (IRLs)

1) Start at some (decent) guess (e.g., $\hat{\boldsymbol{\eta}} = \mathbf{y} + \boldsymbol{\epsilon}$) (Wood 2017)
2) Set $\mathbf{z}^t = \boldsymbol{\eta} + \frac{\partial \hat{\boldsymbol{\eta}}}{\partial g^{-1}(\hat{\boldsymbol{\eta}})}\{\mathbf{y} - g^{-1}(\hat{\boldsymbol{\eta}})\}/a(g^{-1}(\hat{\boldsymbol{\eta}}))$
3) Calculate weights $\mathbf{w} = a(g^{-1}(\hat{\boldsymbol{\eta}}))/\{\frac{\partial \hat{\boldsymbol{\eta}}}{\partial g^{-1}(\hat{\boldsymbol{\eta}})}^2 \text{var}(\mathbf{y}_i)\}$

Iteratively reweighted least squares (IRLs)

1) Start at some (decent) guess (e.g., $\hat{\boldsymbol{\eta}} = \mathbf{y} + \boldsymbol{\epsilon}$) (Wood 2017)
2) Set $\mathbf{z}^t = \boldsymbol{\eta} + \frac{\partial \hat{\boldsymbol{\eta}}}{\partial g^{-1}(\hat{\boldsymbol{\eta}})}\{\mathbf{y} - g^{-1}(\hat{\boldsymbol{\eta}})\}/a(g^{-1}(\hat{\boldsymbol{\eta}}))$
3) Calculate weights $\mathbf{w} = a(g^{-1}(\hat{\boldsymbol{\eta}}))/\{\frac{\partial \hat{\boldsymbol{\eta}}}{\partial g^{-1}(\hat{\boldsymbol{\eta}})}^2 \text{var}(\mathbf{y}_i)\}$
4) Fit weighted LM with $\mathbf{z}$ as "data" and $\mathbf{w}$ as weights

Iteratively reweighted least squares (IRLs)

1) Start at some (decent) guess (e.g., $\hat{\boldsymbol{\eta}} = \mathbf{y} + \boldsymbol{\epsilon}$) (Wood 2017)
2) Set $\mathbf{z}^t = \boldsymbol{\eta} + \frac{\partial \hat{\boldsymbol{\eta}}}{\partial g^{-1}(\hat{\boldsymbol{\eta}})}\{\mathbf{y} - g^{-1}(\hat{\boldsymbol{\eta}})\}/a(g^{-1}(\hat{\boldsymbol{\eta}}))$
3) Calculate weights $\mathbf{w} = a(g^{-1}(\hat{\boldsymbol{\eta}}))/\{\frac{\partial \hat{\boldsymbol{\eta}}}{\partial g^{-1}(\hat{\boldsymbol{\eta}})}^2 \mathrm{var}(\mathbf{y}_i)\}$
4) Fit weighted LM with $\mathbf{z}$ as “data” and $\mathbf{w}$ as weights
5) Repeat until convergence

Iteratively reweighted least squares (IRLs)

1) Start at some (decent) guess (e.g., $\hat{\boldsymbol{\eta}} = \mathbf{y} + \boldsymbol{\epsilon}$) (Wood 2017)
2) Set $\mathbf{z}^t = \boldsymbol{\eta} + \frac{\partial\hat{\boldsymbol{\eta}}}{\partial g^{-1}(\hat{\boldsymbol{\eta}})}\{\mathbf{y} - g^{-1}(\hat{\boldsymbol{\eta}})\}/a(g^{-1}(\hat{\boldsymbol{\eta}}))$
3) Calculate weights $\mathbf{w} = a(g^{-1}(\hat{\boldsymbol{\eta}}))/\{\frac{\partial\hat{\boldsymbol{\eta}}}{\partial g^{-1}(\hat{\boldsymbol{\eta}})}^2 \text{var}(\mathbf{y}_i)\}$
4) Fit weighted LM with $\mathbf{z}$ as "data" and $\mathbf{w}$ as weights
5) Repeat until convergence

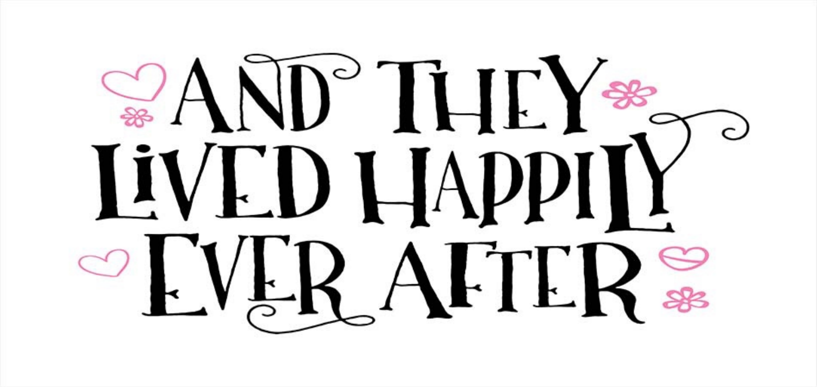
(details omitted)

And that is the day researchers started liking GLMs.

Iteratively reweighted least squares (IRLs)

Prevents having to do numerical optimisation.

The end?



Why is stuff this important?

1) A basic (mathematical) understanding helps apply methods correctly.
2) GLMs may not always converge to the MLE. Then, you will get warnings/errors.
3) If you understand them, you might know what to do! Similar problems in a lot of more complex models (e.g., GLMMs).

Commonly used GLMs

- Normal
- Binomial: occurrence/counts. Prevalence of diseases, number of germinated seeds out of a total
- Poisson: counts. Number of fish caught, stars in the night sky
- Negative binomial (fixed dispersion): still counts
- Gamma: (positive) continuous. Waiting time, body size
- Ordinal (cumulative link). Survey responses

in R

- Similar to the lm() function!
- Now the glm() function

A linear regression:

```r
model <- glm(y ~ x, family = gaussian(link = identity), data = data)
```

A glm:

```r
model <- glm(y ~ x, family = poisson(link = log), data = data)
```

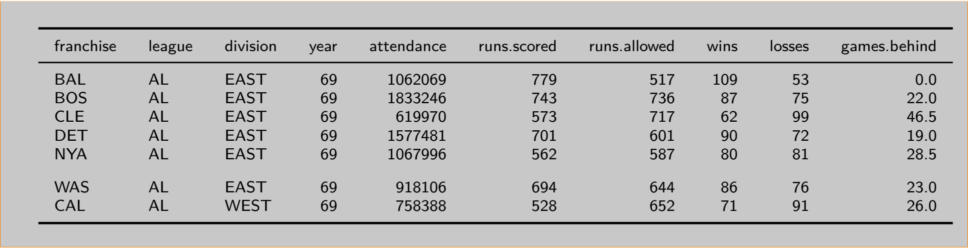
Example: Baseball game wins

Wins of baseball games

Baseball wins: the data

Data from Cochran (2002)

- 838 games
- Many variables: team, league, division, year, runs scores, wins
- Response variable: attendance



franchise	league	division	year	attendance	runs.scored	runs.allowed	wins	losses	games.behind
BAL	AL	EAST	69	1062069	779	517	109	53	0.0
BOS	AL	EAST	69	1833246	743	736	87	75	22.0
CLE	AL	EAST	69	619970	573	717	62	99	46.5
DET	AL	EAST	69	1577481	701	601	90	72	19.0
NYA	AL	EAST	69	1067996	562	587	80	81	28.5
WAS	AL	EAST	69	918106	694	644	86	76	23.0
CAL	AL	WEST	69	758388	528	652	71	91	26.0

Baseball wins: fitting the linear model

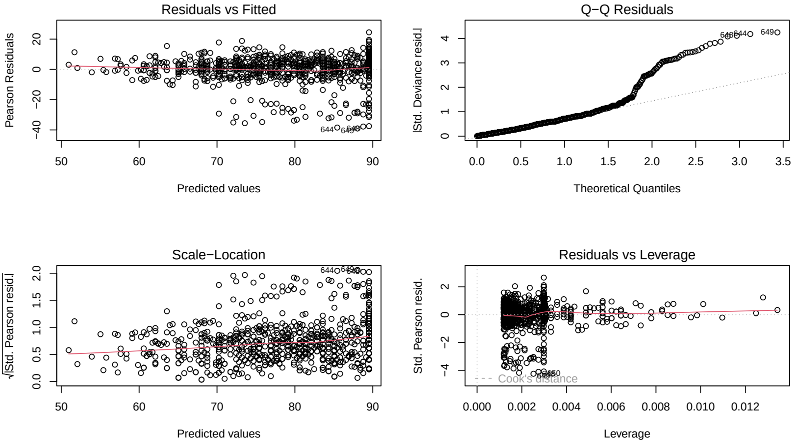
```
model <- glm(wins ~ games.behind,
             family = gaussian(link=identity),
             data = MLBattend)
```

Baseball wins: model summary

```
summary(model)
```

```
##
## Call:
## glm(formula = wins ~ games.behind, family = gaussian(link = identity),
##     data = MLBattend)
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  89.52380    0.50282   178.0   <2e-16 ***
## games.behind -0.74186    0.02707   -27.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 84.71544)
##
##     Null deviance: 134433  on 837  degrees of freedom
## Residual deviance:  70822  on 836  degrees of freedom
## AIC: 6102.3
##
## Number of Fisher Scoring iterations: 2
```

Baseball wins: residuals

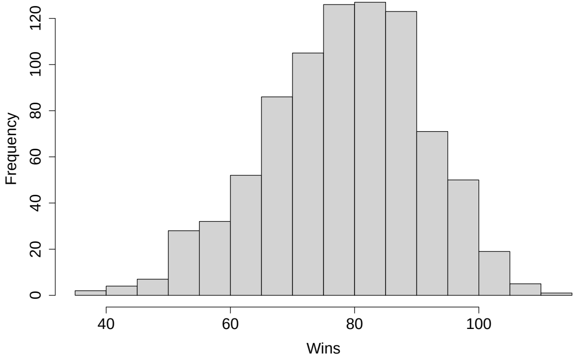


Baseball wins: response distribution

Should we fit a

a) Binomial
b) Poisson
c) Gamma

regression?



```
model <- glm(wins ~ games.behind, family = ...,
             data = MLBattend)
```

Post-fitting

OK, let's say we have a “final” model. Now, the real work begins! (interpreting and presenting the results)

Back-transformation

- Confidence intervals can be back-transformed to the response scale
  - as long as we have "monotonicity" of the link function
- Parameter estimates can sometimes be back-transformed
- Standard errors cannot be back-transformed!
  - Interpretation happens on the **link** scale

Prediction

$$\hat{\mu}_i = g^{-1}(\alpha + \mathbf{x}_i \beta) \tag{9}$$

```
predict(model, type = "response")#Type = link alternatively
```

Newdata

```
predict(model, newdata = X, type = "response")
```

Intervals for $\widehat{\mu}_i$

- Confidence intervals
- Prediction intervals

More on predicting tomorrow in the practical

Deviance

- LMs used RSS to quantify fit
- GLMs use deviance to quantify **lack of fit**
- Deviance $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$ is different for every GLM.

Deviance

- LMs used RSS to quantify fit
- GLMs use deviance to quantify **lack of fit**
- Deviance $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$ is different for every GLM.

Normal: $\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$ Poisson: $2\sum_{i=1}^{n} y_i \log(y_i/\hat{\mu}_i) - y_i + \hat{\mu}_i)$

binomial: $2\sum_{i=1}^{n} y_i \log(y_i/\hat{\mu}_i) + (N - y_i)\log\{(N - y_i)/(N - y_i)\}$

gamma: $2\sum_{i=1}^{n} -\log\{y_i/\hat{\mu}_i\} + (y_i - \hat{\mu}_i)/\hat{\mu}_i$

and so on.

Deviance

- Measures (twice) difference to a model that perfectly fits the data (“saturated model”)
  - equal to RSS for linear regression
- Can be used for hypothesis testing, calculate $R^2_{deviance}$, and “deviance residuals”

Asymptotic requirements

- Binomial responses: $N\pi \geq 3$ and $N(1-\pi) \geq 3$
  - Deviance for $N = 1$ has no concept of residual variability
- Poisson responses: $\lambda \geq 3$
- Gamma distribution: $\phi \leq 3$

(Dunn and Smyth 2021)

Residuals

- GLMs lack an explicit error term, but it is there!
- So we can still check assumptions by residuals, though they are differently defined
- There are different types; Pearson, Deviance, Anscombe, Quantile residuals..
- We usually -hope- that they are approximately normally distributed
- Residual checking in GLMs can be difficult

(omitted details)

Recap

- Remember to bring all components together
- Parameter estimates, uncertainty, multiple predictors, interaction, model selection
- GLMs for when assumptions of LMs fail (which is very often)
- We covered components of GLMs here
- And how they are fitted in R
- Deviance and residuals