logiwhat?

From Wikipedia:

> *The function was named in 1844 by Pierre François Verhulst, who studied it in relation to population growth. The initial stage of growth is approximately exponential (geometric); then, as saturation begins, the growth slows to linear (arithmetic), and at maturity, growth stops. Verhulst did not explain the choice of the term "logistic", but it is presumably in contrast to the logarithmic curve and by analogy with arithmetic and geometric. His growth model is preceded by a discussion of arithmetic growth and geometric growth, and thus "logistic growth" is presumably named by analogy, logistic being from Ancient Greek logistikós, a traditional division of Greek mathematics. The term is unrelated to the military and management term logistics.*

The "logit" function/unit/transform comes from the contraction of **log**istic un**it**.

Generalized linear models and the logistic regression

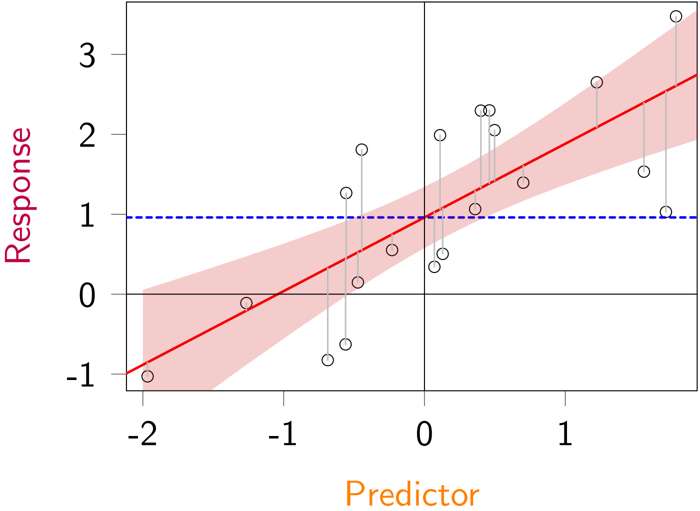
Timothée Bonnet

Thanks to BDSI for support!

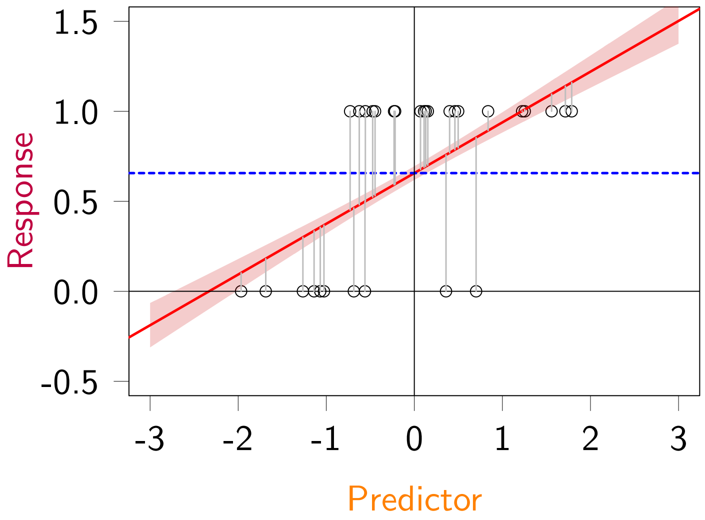
April 11, 2019

A simple linear model

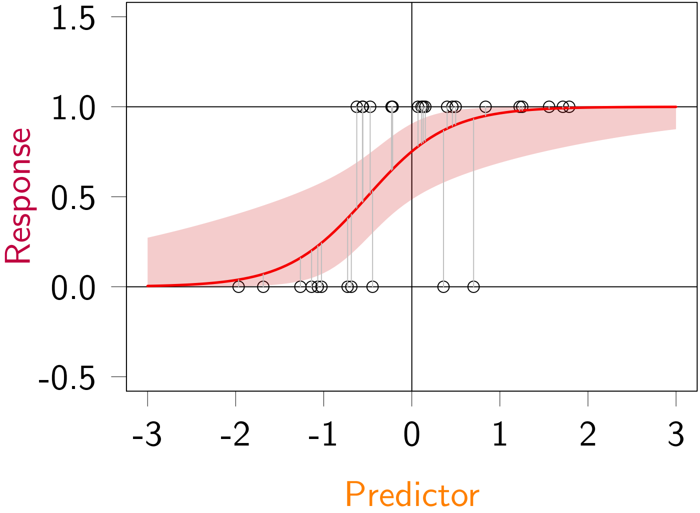
Response = **Intercept** + **Slope** × **Predictor** + **Error**



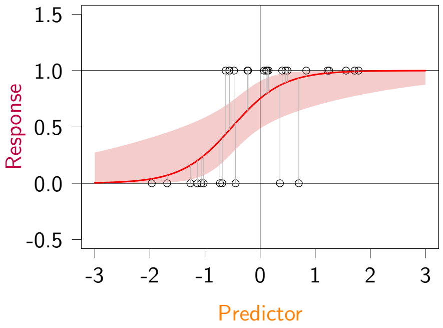
A simple linear model failure: binary data



What we want our model to do



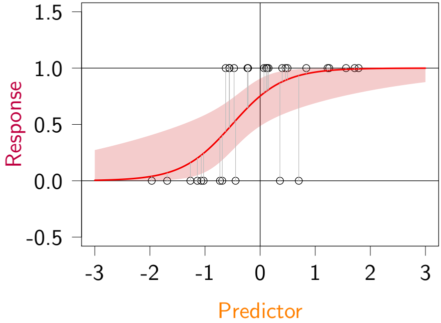
What we want our model to do



What we need:

1. Convert the predictor open scale ($-\infty$ to $+\infty$) to a bounded scale (0 to 1)

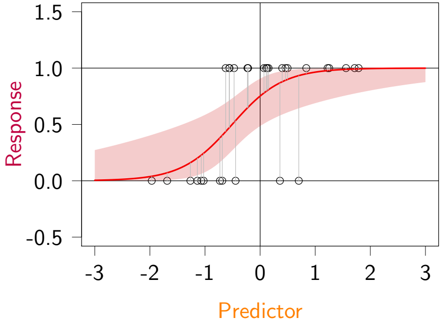
What we want our model to do



What we need:

1. Convert the predictor open scale ($-\infty$ to $+\infty$) to a bounded scale (0 to 1)
2. Acknowledge discrete data

What we want our model to do



What we need:

1. Convert the predictor open scale ($-\infty$ to $+\infty$) to a bounded scale (0 to 1)
2. Acknowledge discrete data
3. Response variability depends on expected value

That is what a Generalized Linear Model does

Vocabulary warning

- General Linear Model (=linear model with several responses, multivariate)
- **Generalized Linear Model (=non-normal errors, and uncertainty dependent on the mean)**

That is what a Generalized Linear Model does

Vocabulary warning

- General Linear Model (=linear model with several responses, multivariate)
- **Generalized Linear Model (=non-normal errors, and uncertainty dependent on the mean)**

What a GLM is:

1. **Linear function** (reponse = intercept + slope × predictor . . .)

That is what a Generalized Linear Model does

Vocabulary warning

- General Linear Model (=linear model with several responses, multivariate)
- **Generalized Linear Model (=non-normal errors, and uncertainty dependent on the mean)**

What a GLM is:

1. **Linear function** (reponse = intercept + slope × predictor . . .)
2. "**Link function**" = a map between the linear function ($-\infty$ to $+\infty$) and a probability distribution (from 0 to 1 for Bernouilli)

That is what a Generalized Linear Model does

Vocabulary warning

- General Linear Model (=linear model with several responses, multivariate)
- **Generalized Linear Model (=non-normal errors, and uncertainty dependent on the mean)**

What a GLM is:

1. **Linear function** (reponse = intercept + slope × predictor . . .)
2. "**Link function**" = a map between the linear function ($-\infty$ to $+\infty$) and a probability distribution (from 0 to 1 for Bernouilli)
3. **Probability distribution** (Bernouilli, Binomial, Poisson. . .) thought to generate the data (either 0 or 1 for Bernouilli)

That is what a Generalized Linear Model does

Vocabulary warning

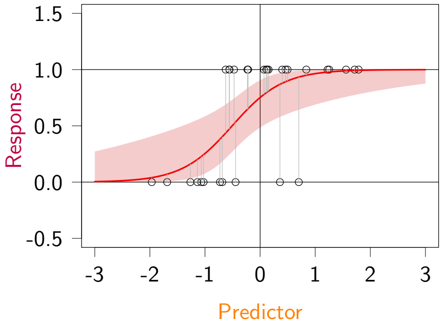
- General Linear Model (=linear model with several responses, multivariate)
- **Generalized Linear Model (=non-normal errors, and uncertainty dependent on the mean)**

What a GLM is:

1. **Linear function** (reponse = intercept + slope × predictor . . .)
2. "**Link function**" = a map between the linear function ($-\infty$ to $+\infty$) and a probability distribution (from 0 to 1 for Bernouilli)
3. **Probability distribution** (Bernouilli, Binomial, Poisson. . .) thought to generate the data (either 0 or 1 for Bernouilli)

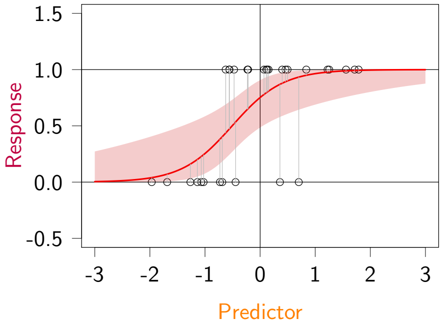
GLMs fit continuous expected response; we observe discrete realizations

What to do with logistic regression



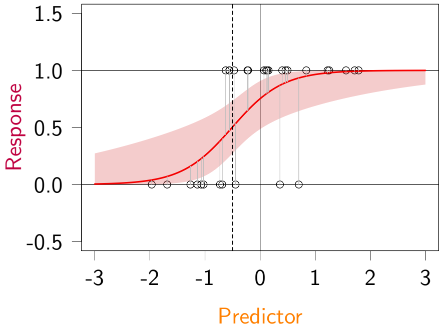
1. Response increase/decrease with increasing predictor?

What to do with logistic regression



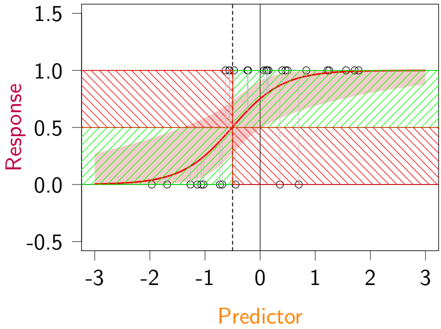
1. Response increase/decrease with increasing predictor?
2. Estimate probability of 0/1 given a predictor value

What to do with logistic regression



1. Response increase/decrease with increasing predictor?
2. Estimate probability of 0/1 given a predictor value
3. Predict 0/1 and classify predictor values (→ Machine Learning)

What to do with logistic regression



1. Response increase/decrease with increasing predictor?
2. Estimate probability of 0/1 given a predictor value
3. Predict 0/1 and classify predictor values (→ Machine Learning)

Exercise 1, fit a glm

Load the dataset `survivalsize.csv`. It contains fake data of individual-based measurements of body size and of survival from the time of measurement to the next year. Look at a summary of the data and plot them. Do you think size affects survival? Use the function glm() to fit a logistic regression. What should the `family` argument be? What is the direction of the effect of size on survival?

Exercise 2, Model assumptions?

In R some model assumptions of linear models are routinely checked using plot(lm()): residual normality, independance and homogeneous variance, and legerage. If you know about these diagnostics (and what the plots should ideally look like) check them for your glm. Should you worry?

Model assumptions

Logistic regression assumes:

- **Binary data**

Model assumptions

Logistic regression assumes:

- **Binary data**
- No unaccounted source of correlations in the date (e.g., pseudo-replication, spatial autocorrelations, phylogenetic signal. . .)

Model assumptions

Logistic regression assumes:

- **Binary data**
- No unaccounted source of correlations in the date (e.g., pseudo-replication, spatial autocorrelations, phylogenetic signal. . .)
- (no error in the predictors)

Model assumptions

Logistic regression assumes:

- **Binary data**
- No unaccounted source of correlations in the date (e.g., pseudo-replication, spatial autocorrelations, phylogenetic signal. . .)
- (no error in the predictors)
- (no complete separation = only 0s or only 1s for some predictor level)

Model assumptions

Logistic regression assumes:

- **Binary data**
- No unaccounted source of correlations in the date (e.g., pseudo-replication, spatial autocorrelations, phylogenetic signal. . .)
- (no error in the predictors)
- (no complete separation = only 0s or only 1s for some predictor level)

Model assumptions

Logistic regression assumes:

- **Binary data**
- No unaccounted source of correlations in the date (e.g., pseudo-replication, spatial autocorrelations, phylogenetic signal. . .)
- (no error in the predictors)
- (no complete separation = only 0s or only 1s for some predictor level)

NO assumptions about the distribution of residuals (Normality, homoscedasticity).
BUT more assumptions in non-binary GLMs (proportions and count data)!!

Back-transformation

```
summary(glm(survival ~ 1 + size, data = survdat, family = "binomial"))


Call:
glm(formula = survival ~ 1 + size, family = "binomial", data = survdat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6020  -0.6057   0.1078   0.6412   2.1218

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -13.4781     1.9485  -6.917 4.61e-12 ***
size          2.8078     0.4015   6.993 2.70e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 274.83  on 199  degrees of freedom
```

Back-transformation

Scales:

Back-transformation

Scales:

Model estimates $-\infty$ - - ———————— 0 ———————— - - $+\infty$

Back-transformation

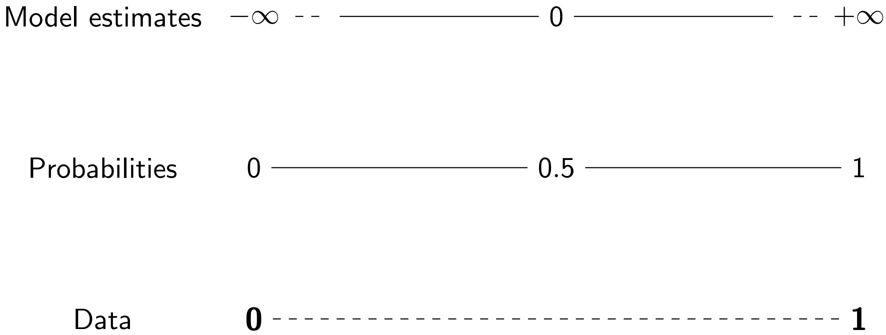
Scales:

Model estimates $-\infty$ - - ——— 0 ——— - - $+\infty$

Probabilities 0 ——— 0.5 ——— 1

Back-transformation

Scales:



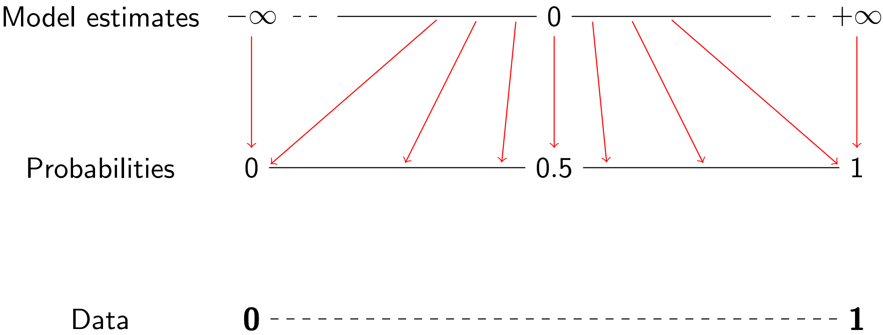
Model estimates $-\infty$ - - —— 0 —— - - $+\infty$

Probabilities 0 —— 0.5 —— 1

Data **0** - - - - - - - - - - **1**

Back-transformation

Scales:



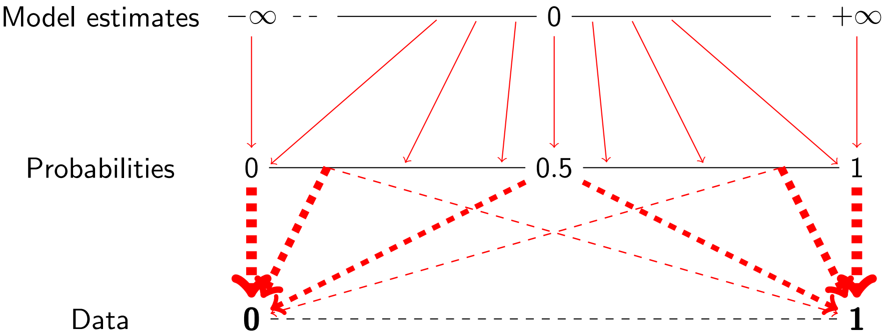
Model estimates $-\infty$ - - 0 - - $+\infty$

Probabilities 0 0.5 1

Data **0** - - - - - - - - - - - - - - - - - - - **1**

Back-transformation

Scales:



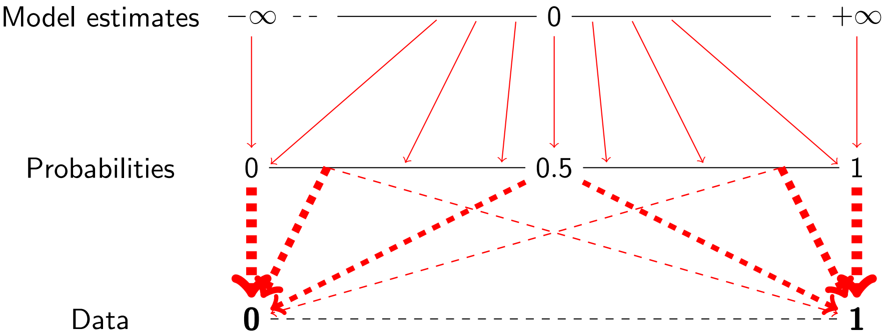
Model estimates $-\infty$ 0 $+\infty$

Probabilities 0 0.5 1

Data **0** **1**

Back-transformation

Scales:



Conversion:

- from model to probability: $p = \frac{1}{1+\exp(-x)}$ or `plogis(x)`
- probability and data on same scale, but continuous/discrete
- $\exp(\textit{slope}) =$ odd-ratio

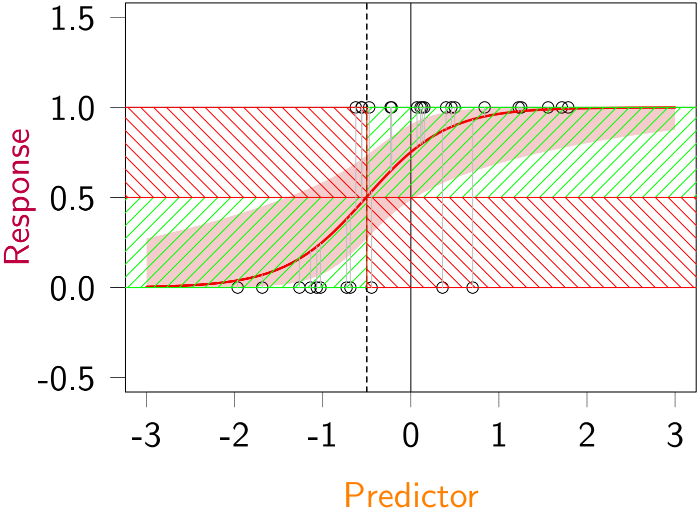
Exercise 3, interpretation

Exercise 4, visualize model

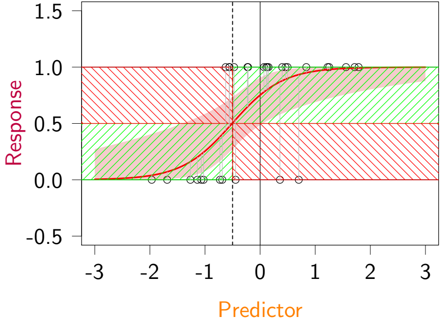
Take pictures!

BDSI would like some pictures to illustrate their website. Is everybody okay with pictures?

Clustering

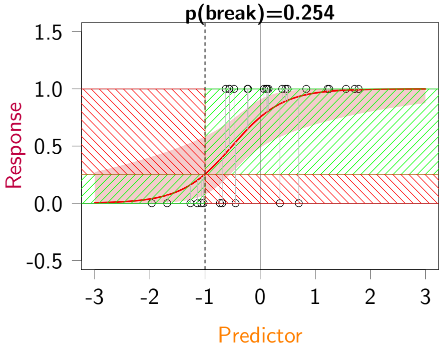


Clustering: what threshold?



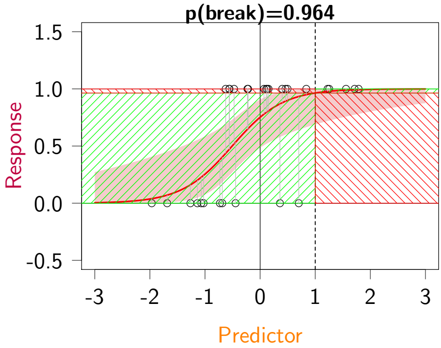
- Break at 50% probability?

Clustering: what threshold?



- Break at 50% probability?
- Never miss a case? (but false positives!) e.g., Epidemic prevention

Clustering: what threshold?



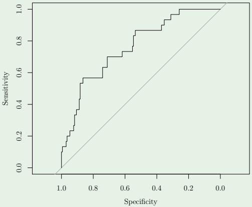
- Break at 50% probability?
- Never miss a case? (but false positives!) e.g., Epidemic prevention
- Never false positive? (but miss some positives!) e.g., Presumption of innocence

Exercise 5

Exercise 5

How good is the classification? What threshold to use?

- Receiver Operating Characteristic and Area Under the Curve (ROC / AUC)
- E.g., `library(pROC)`



Mixed effect logistic regression

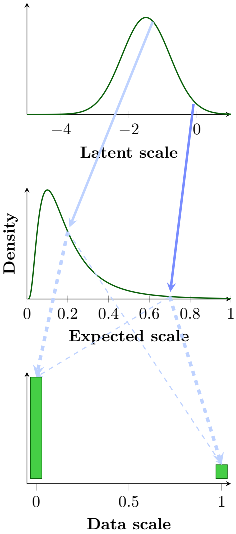
GLM + random effect = GLMM

Mixed effect logistic regression

GLM + random effect = GLMM

```r
library(lme4)
glmer(response ~ 1 + predictor + (1|group), family="binomial",
          data=dat)
```

Mixed effect logistic regression



- There is no meaningful residual variance
- Random effect variance distorted by logit
- Lots of variance from random process on top of expected values

Repeatability in GLMM

In LMM: repeatability = variance random effect / (variance random effect + residual variance)

Repeatability in GLMM

In LMM: repeatability = variance random effect / (variance random effect + residual variance)

but in GLMM, especially logistic:

- On what scale to take the random effect?
- No residual variance! Is repeatability always 1??

Exercise 6