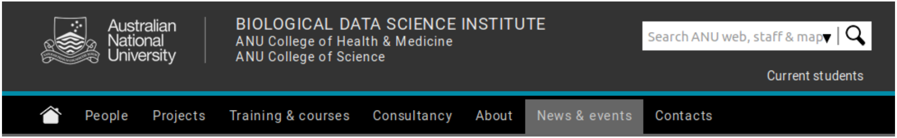
Where to find help?

- The Internet
- Colleagues
- (books: learn BEFORE you have an issue)
- Courses, workshops, consulting

Where to find help?

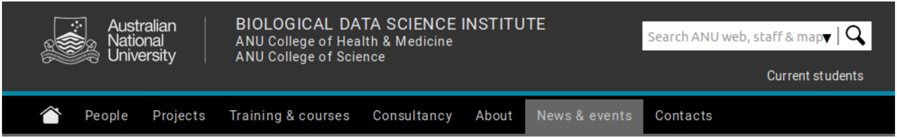
- The Internet
- Colleagues
- (books: learn BEFORE you have an issue)
- Courses, workshops, consulting



BDSI: Terry Neeman, Marcin Adamski, Cameron Jack, myself . . .

Where to find help?

- The Internet
- Colleagues
- (books: learn BEFORE you have an issue)
- Courses, workshops, consulting



BDSI: Terry Neeman, Marcin Adamski, Cameron Jack, myself . . .

+ R-Ladies, Coding Club

Statistical Modelling: Beyond Linear Models, Generalized Linear Models

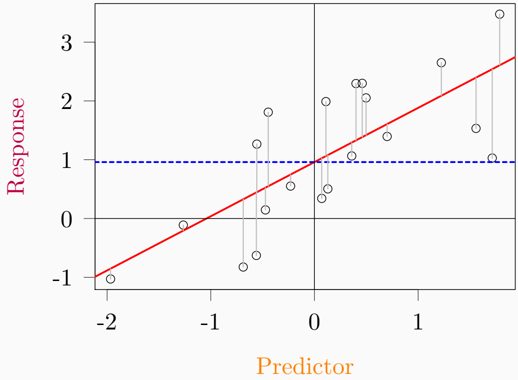
Chapter 5

Timothée Bonnet

May 15, 2019

Simple linear models

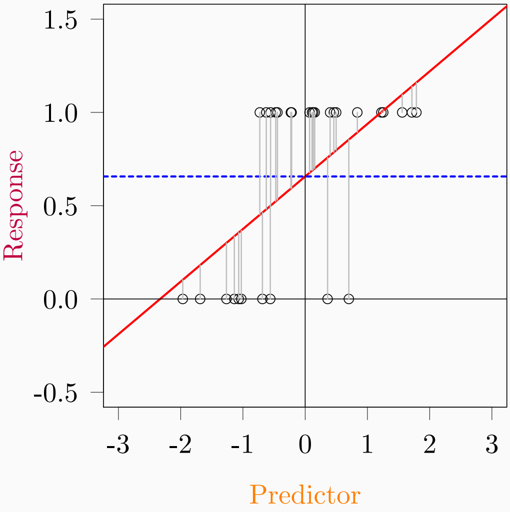
Response = Intercept + Slope × Predictor + Error



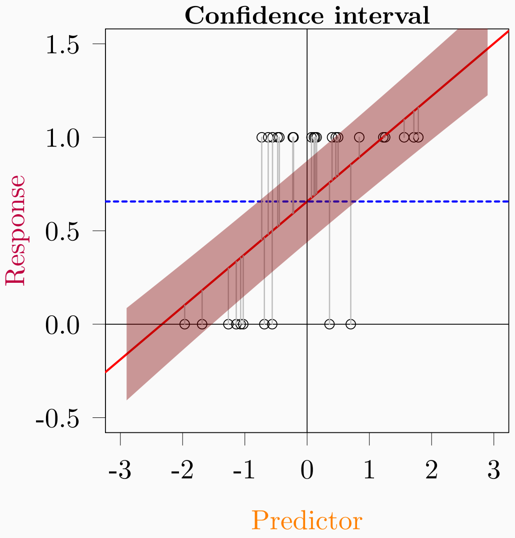
Linear model basic assumptions

- Predictor not perfectly correlated
  Risk: Model won't run, unstable convergence, or huge SE
- Little error in predictors
  Risk: bias estimates (underestimate with Gaussian error)
- Gaussian error distribution
  Risk: Poor predictions
- Homoscedasticity (constant error variance)
  Risk: Over-optimistic uncertainty, unreliable predictions
- Independence of error
  Risk: Bias and over-optimistic uncertainty

A simple linear model failure: binary data

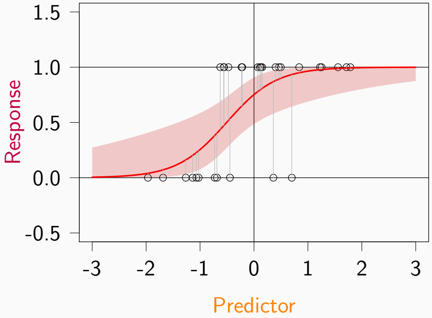


A simple linear model failure: binary data



Assumptions violated:
Non-Gaussian errors, non-constant error variance, correlated errors

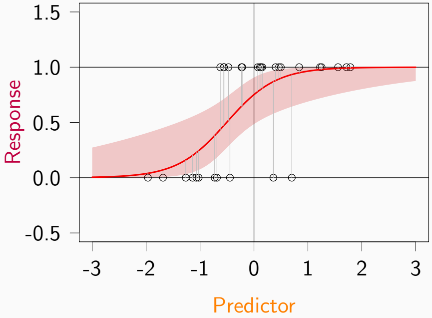
What we want our model to do



What we need:

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

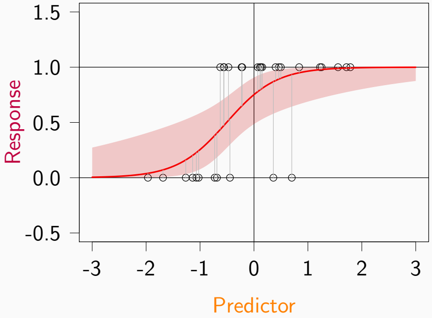
What we want our model to do



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 $\times$ predictor $\ldots$)

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 $\times$ 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 $\times$ 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 $\times$ 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

Binary data

Count data

Logistic regression

- Binary or proportion data (survival, presence/absence...)

Logistic regression

- Binary or proportion data (survival, presence/absence...)
- Binomial probability distribution ($=$ Bernouilly if binary data)

Logistic regression

- Binary or proportion data (survival, presence/absence...)
- Binomial probability distribution (= Bernouilly if binary data)
- Link function often logit: $y = \log(\frac{probability}{1-probability})$

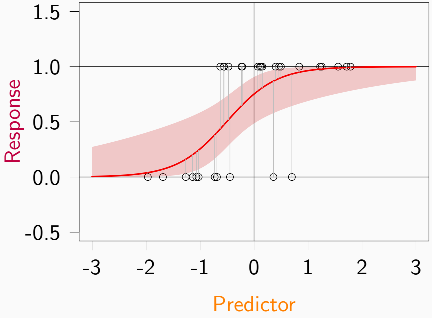
Logistic regression

- Binary or proportion data (survival, presence/absence...)
- Binomial probability distribution (= Bernouilly if binary data)
- Link function often logit: $y = \log(\frac{probability}{1 - probability})$
- Linear function $y = intercept + slope_1 predictor_1 + slope_2 predictor_2 +$

What is the Bernouilli distribution?

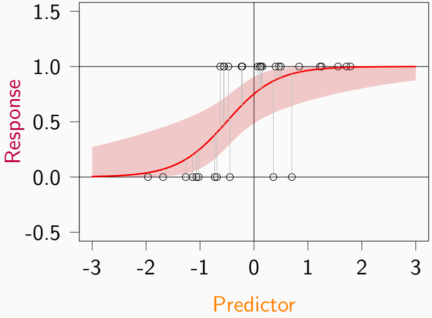
```
bernouilli_random_sample <- rbinom(n = 10000, size = 1, prob = 0.3)
hist(bernouilli_random_sample)
mean(bernouilli_random_sample); 0.3
var(bernouilli_random_sample); 0.3*(1-0.3)
```

What to do with logistic regression



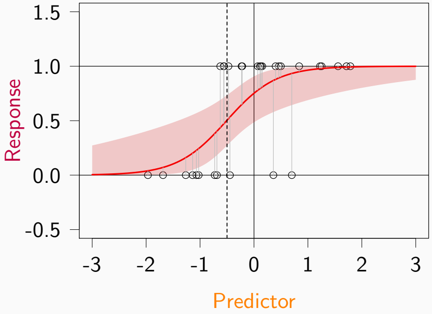
1. Response increase/decrease with increasing predictor?

What to do with logistic regression



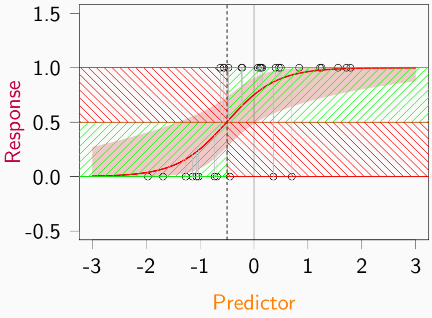
1. Response increase/decrease with increasing predictor?
2. Estimate probability of 0/1 given a predictor value

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 ($\rightarrow$ 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 ($\rightarrow$ Machine Learning)

Logistic regression in R

```
glm(formula = obs ~ 1 + x, family = "binomial", data=data)
```

Does survival probability depend on size?

Exercise, part 1

1. Load `survivalsize.csv`
2. Plot survival data. What kind of distribution is it?
3. Logistic GLM of survival as a function of size. How does size correlates with survival?
4. What is the unit of coefficients?

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:

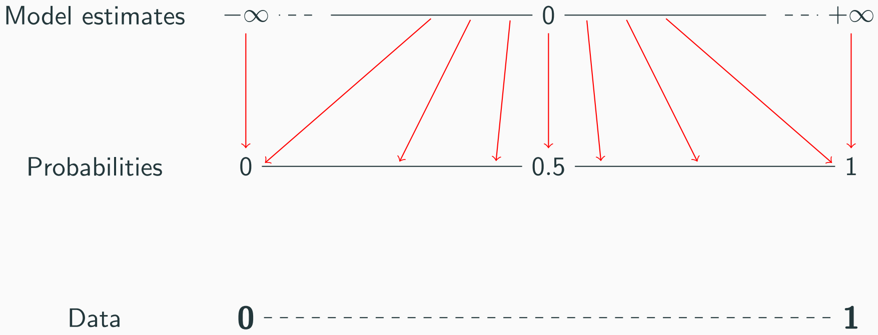
Model estimates $-\infty$ ---- 0 ---- $+\infty$

Probabilities 0 ---- 0.5 ---- 1

Data $\mathbf{0}$ ---- $\mathbf{1}$

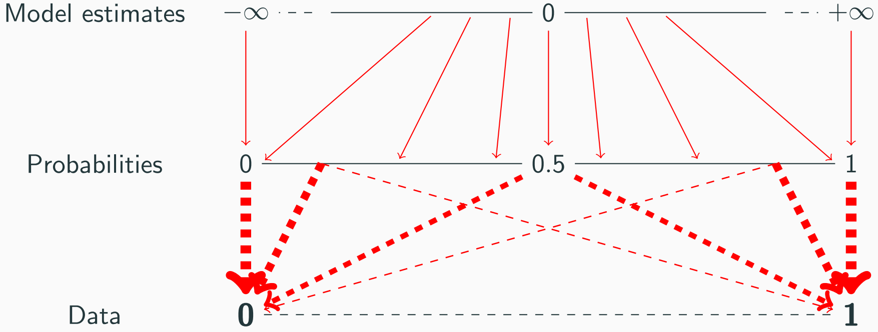
Back-transformation

Scales:



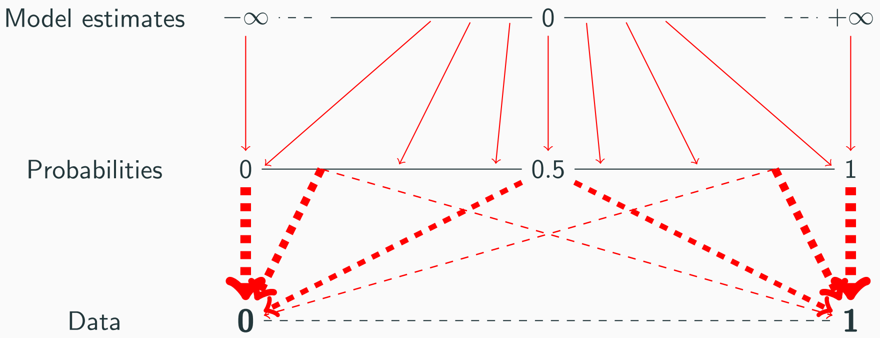
Back-transformation

Scales:



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

Does survival probability depend on size?

Exercise, part 2

1. Load `survivalsize.csv`
2. Fit a linear model and a logistic model with intercept only. How to interpret the estimates?

Does survival probability depend on size?

Exercise, part 2

1. Load `survivalsize.csv`
2. Fit a linear model and a logistic model with intercept only. How to interpret the estimates?

hints:

1. For a given predicted y, $\exp(y)$ is the odd ratio: probability success / probability failure
2. Back-transformation inverse-logit: $\mathit{probability} = \frac{1}{1+exp(-y)}$

Solutions part 2

```
surv <- read.csv("Data/survival.csv")
plot(surv$survival)
lmsurv <- glm(survival~1, data=surv, family=gaussian)
lregsurv <- glm(survival~1, data=surv, family=binomial)

#linear model prediction:
coefficients(lmsurv)

#logistic reg prediction:
plogis(coefficients(lregsurv))
1/(1+exp(-coefficients(lregsurv)))
exp(coefficients(lregsurv))

#observed mean survival:
mean(surv$survival)
#mean odd-ratio:
mean(surv$survival)/(1-mean(surv$survival))
```

Does survival probability depend on size?

Exercise, part 3

1. Fit a linear regression and a logistic regression of survival on relative size, compare the outputs
2. Check the diagnostic plots for both models. Should you be worried?
3. Extract and visualize a model prediction from both models (use the function predict(), and/or do it by hand to practice link-function back-transformation)

Solutions part 3

```
lmsurvS <- glm(survival~1 + relative_size, data=surv, family=gaussian)
lregsurvS <- glm(survival~1 + relative_size, data=surv, family=binomial)

summary(lmsurvS)
summary(lregsurvS)

plot(lmsurvS)
plot(lregsurvS)

plot(surv$relative_size, surv$survival, ylim=c(-0.2,1.2))
abline(lmsurv, col="red")

plot(surv$relative_size, surv$survival, ylim=c(-0.2,1.2))
datforpred <- data.frame(relative_size=seq(from=-3,to=4, by=0.1))
datforpred$prob <- predict(lregsurvS, newdata = datforpred,
type = "response")
lines(datforpred$relative_size, datforpred$prob, col="red")

ggplot(surv, aes(x = relative_size, y=survival))+geom_point()+
stat_smooth(method = "glm", method.args = list(family = "binomial"))
```

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)!!

More practice: does survival probability depend on weight? does the relationship depend on sex?

Exercise

1. Load `survivalweight.csv`
2. Plot data
3. Fit a logistic model to address these questions
4. Plot the results

Binary data

Count data

Poisson regression

- Count data
- Poisson distribution
- Link function: logarithm
- Inverse link function: exponential
- Linear function $y = intercept + slope_1 predictor_1 + slope_2 predictor_2 + \ldots$

What is the Poisson distribution?

```
poisson_random_sample <- rpois(n = 10000, lambda = 4)
hist(poisson_random_sample)
mean(poisson_random_sample)
var(poisson_random_sample)
```

Poisson regression in R

```
glm(formula = obs ~1 + x, family = "poisson", data=data)

glm(formula = obs ~1 + x, family = "quasipoisson", data=data)
```

family = "poisson" is dangerous

- A true Poisson distribution has $E(\exp(Y)) = V(\exp(Y))$
- Assumes no unexplained variation in Y
- `glm()` follows this assumption
- In nature, $E(\exp(Y)) < V(\exp(Y))$ most of the time
- SE and p-value to small
- family = "quasipoisson" correct the uncertainty in `glm()`
- or mixed model with `(1|obs)`
- other packages never follow the assumption (`MCMCglmm`)

Practice with Poisson glm

Exercise

1. Load the data reproduction.csv
2. Plot reproduction data, calculate the mean and variance.
3. Overlay a Gaussian distribution of same mean and variance, does it fit?
4. Fit an compare a lm and a Poisson glm of reproduction on size
5. Check the diagnostic plots for both models. Should you be worried?
6. Extract and visualize a model prediction from both models (use the function predict, and/or do it by hand to practice link-function back-transformation)
7. Before GLMs, researchers used to log-transform the data and fit linear models. What are the problems with this approach?

Can we decrease aggressive behavior in noisy miners?

Context

- "Harassment.Data.csv"
- Outcome measure: number of attacks
- Experimental factor: Removal of noisy miners (Control/Treatment); Just-After Treatment / long-term ("Phase")
- Data: 6 farms, 8 one-hour surveys for each combination