

Logistic regression - an introduction

Ian Handel

June 2019

Learning Outcomes

- **Refresher** on logs, odds, probability and linear regression
- Understand why linear regression not sensible for **binary data**
- Explain how **logit** and binomial model let us **extend linear regression**
- Be able to run a **simple logistic regression in R**
- Be able to explain basic R glm **output**
- Be able to explain **estimates** with categorical and continuous variables
- Explain **significance test results** on variables
- Things to watch out for!
- **Know where to go next!**

But first - some R

```
install.packages("usethis")  
usethis::use_course("bit.ly/ian-logreg")
```

Logarithms ('logs')

Can we skip this bit?

$$\log_{10}(10) = 1$$

$$\log_{10}(1000) = 3$$

$$\log_{10}(0.01) = -2$$

We can have other bases e.g. e

$$\log_e(2.718) \simeq 1$$

And reversing this...

$$10^3 = 1000$$

$$e^2 \simeq 7.389$$

Odds and probability

Probabilities have values from 0 ('never happens') to 1 ('always happens')

'events of interest' ÷ 'all events'

What is the probability that a fair coin lands on heads?

$$1/2 = 0.5$$

What is the probability that a 6 sided die 🎲 lands on 4?

$$1/6 \simeq 0.166$$

Odds have values from 0 ('never happen') to infinity ('always happens')

'events of interest' ÷ 'other events'

What is the odds that a fair coin lands on heads?

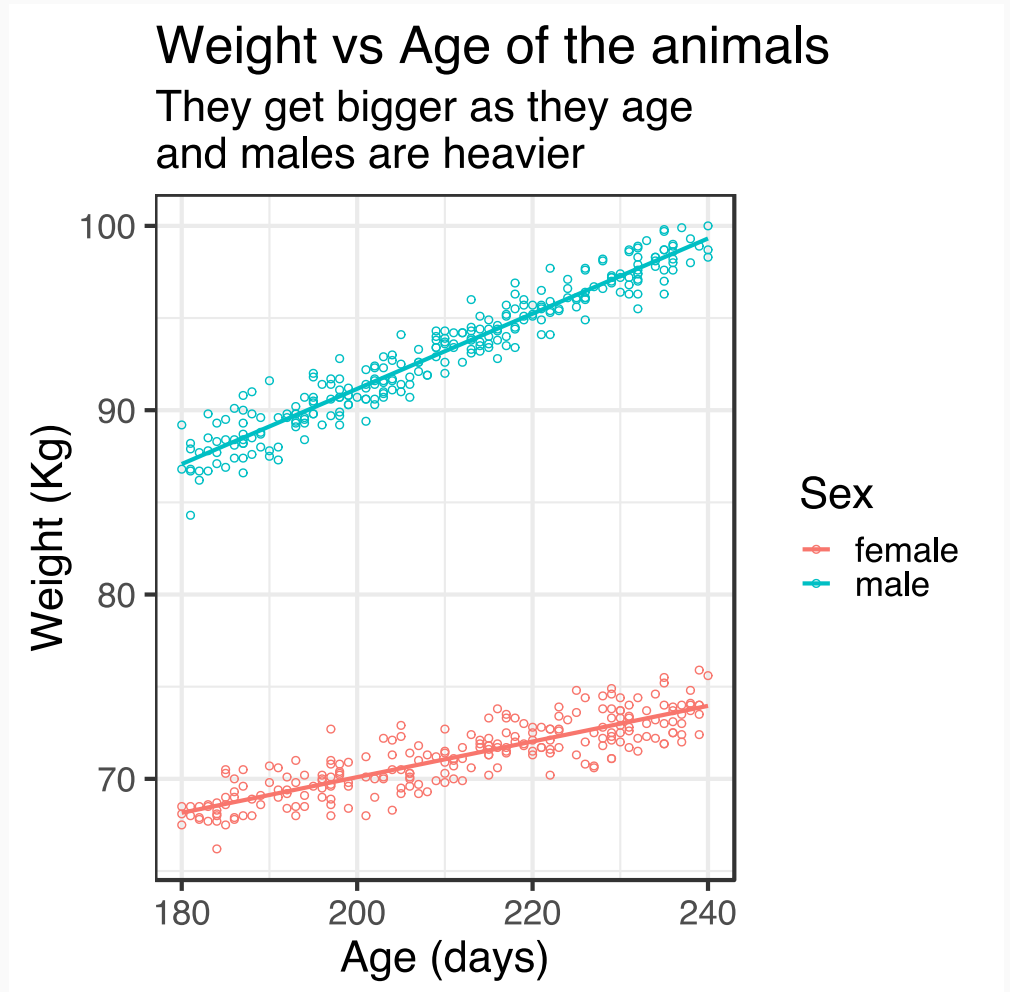
$$1/1 = 1$$

What is the odds that a 6 sided die 🎲 lands on 4?

$$1/5 = 0.2$$

Linear regression

- numerical outcome
- numerical / categorical predictors
- linear relationship



Linear regression in R

```
mod ← lm(weight ~ age + sex, data = dat)
```

Call:

```
lm(formula = weight ~ age + sex, data = dat)
```

Residuals:

Min	1Q	Median	3Q	Max
-4.49	-0.98	0.09	0.95	3.70

Coefficients:

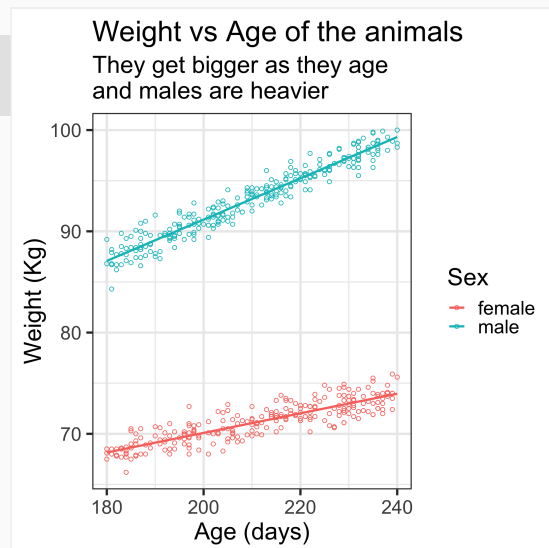
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	39.2331	0.7430	53	<2e-16	***
age	0.1511	0.0035	43	<2e-16	***
sexmale	22.2003	0.1196	186	<2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.3 on 497 degrees of freedom

Multiple R-squared: 0.99, Adjusted R-squared: 0.99

F-statistic: 1.8e+04 on 2 and 497 DF, p-value: <2e-16



Analysing binary data

Binary data common in epidemiology e.g.

- alive/dead
- healthy/diseased

Example data

ID	treatment	age	region	supp	sex	weight	status
A0458	control	209	C	7.969148	male	93.4	diseased
A0468	treated	190	C	5.740540	male	87.8	healthy
A0143	control	239	B	5.864910	female	74.0	healthy
A0413	control	235	D	3.075370	male	97.0	diseased
A0319	control	197	B	4.013347	male	89.7	healthy
A0257	control	194	B	7.821795	female	69.1	diseased

Univariable analysis

Status vs treatment

treatment	diseased	healthy
control	171	43
treated	200	86

```
with(dat,  
      fisher.test(status, treatment))
```

Fisher's Exact Test for Count Data

data: status and treatment

p-value = 0.01312

alternative hypothesis: true odds ratio is not equal to 1

95 percent confidence interval:

1.104065 2.669617

sample estimates:

odds ratio

1.708177

Multivariable analysis

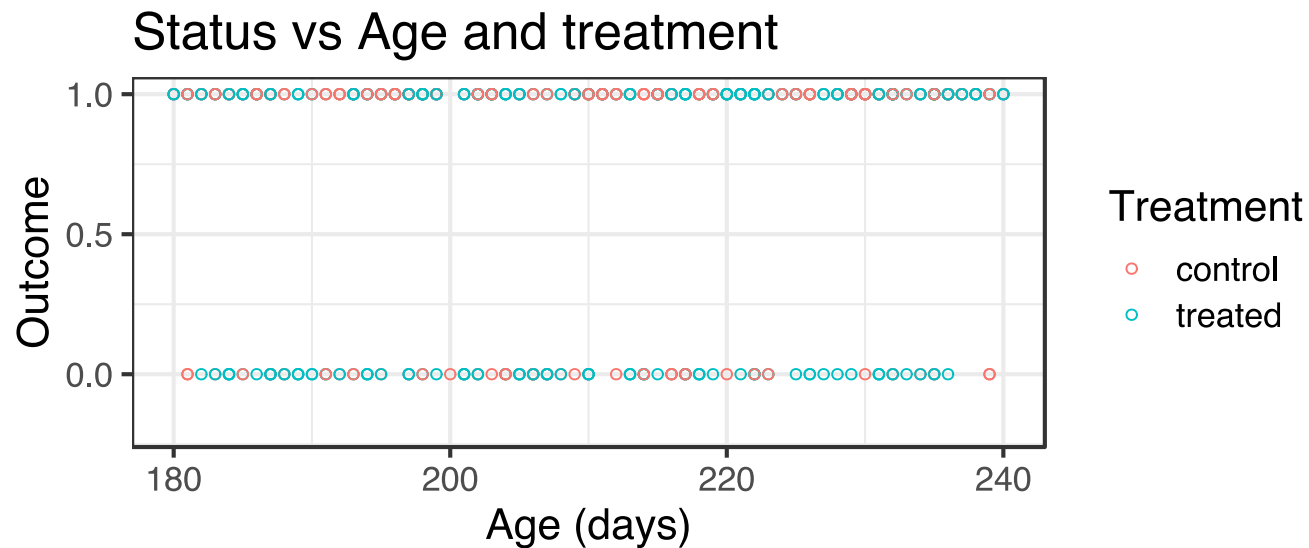
How about recoding the outcome as 0/1?

Example data

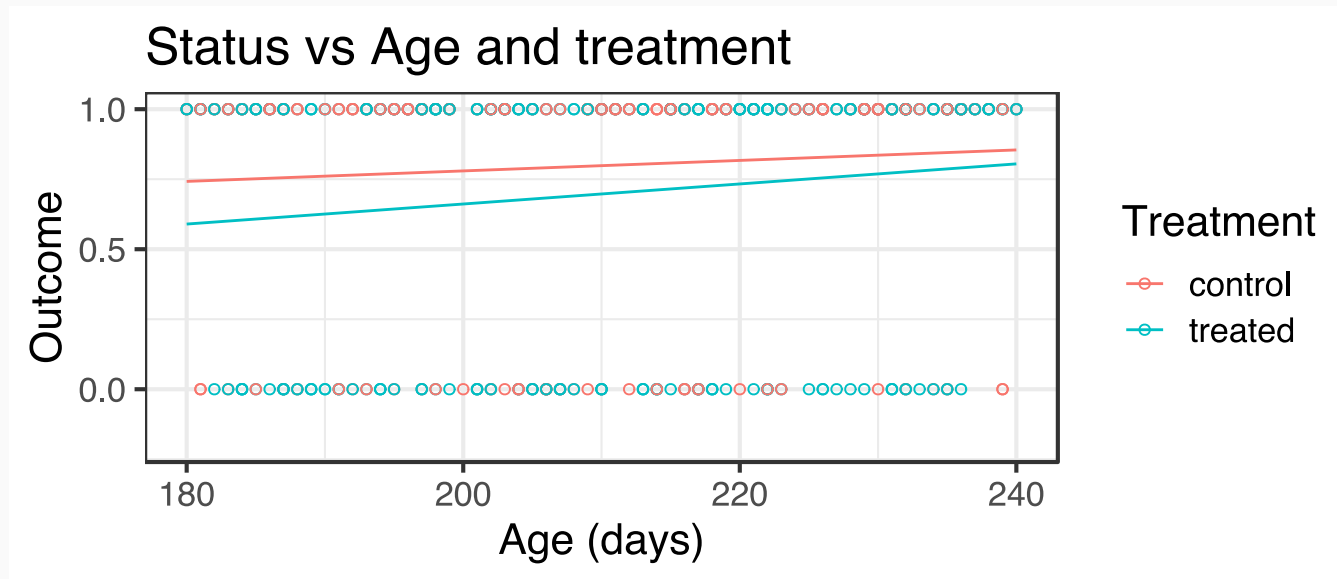
ID	treatment	age	region	supp	sex	weight	status	status01
A0458	control	209	C	7.969148	male	93.4	diseased	1
A0468	treated	190	C	5.740540	male	87.8	healthy	0
A0143	control	239	B	5.864910	female	74.0	healthy	0
A0413	control	235	D	3.075370	male	97.0	diseased	1
A0319	control	197	B	4.013347	male	89.7	healthy	0
A0257	control	194	B	7.821795	female	69.1	diseased	1

Then use linear regression...

Linear regression 1



Linear regression 2



Problems

- predicts (impossible) intermediate values
- can predict <0 and >1

So how do we fix this?

Linear regression does this...

$$weight \sim \beta_0 + \beta_1 age + \beta_2 sex + \epsilon$$

or in english...

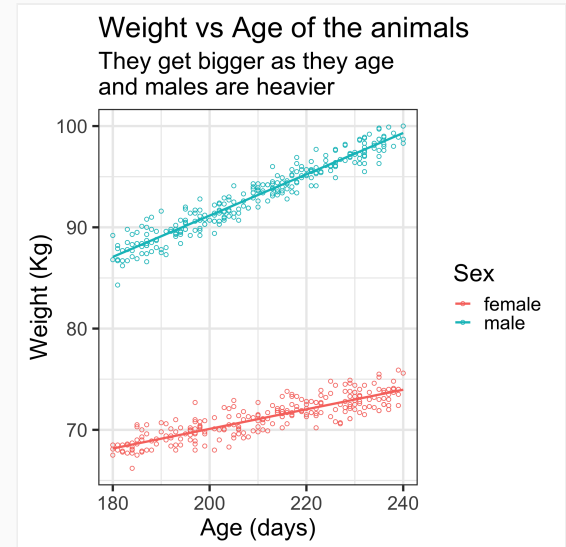
The outcome, *weight*, is related to the predictors by one or more straight lines.

For binary data we want

Our outcome to be 0 or 1

So rather than modelling the outcome.

We model the **probability** of something e.g. being diseased...



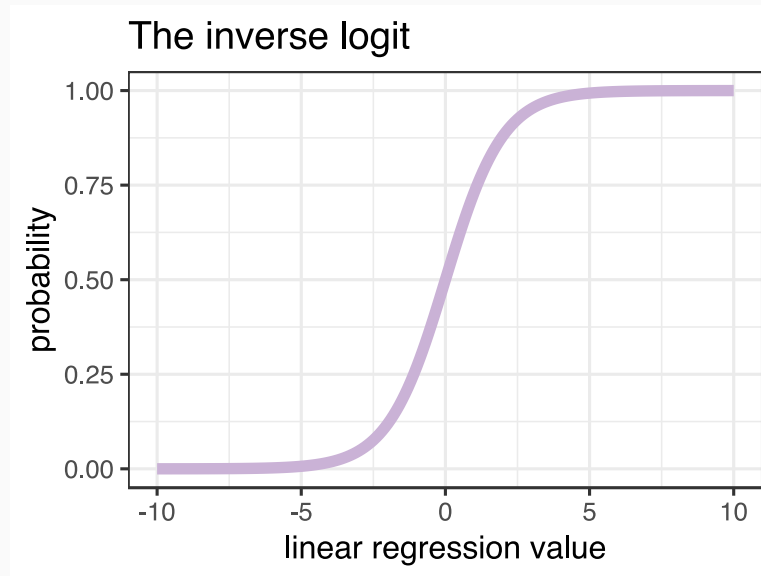
The logistic bit...

Linear regression models model numbers, any numbers!

Probabilities go from...

0 to 1

So we need to turn any number into 0 - 1



In fact the regression value is the log of the odds of the outcome.

The logistic bit 2

So we have an outcome, e.g. being diseased vs healthy, that is coded 0 or 1

And our model is

$$\log_e\left(\frac{prob}{1 - prob}\right) \sim \beta_0 + \beta_1 age + \beta_2 treatment$$

or in english

The log of the odds of an animal being diseased are modelled by a linear combination of the predictor variables

Worked example in R

R code for logistic regression

```
head(dat)
```

ID	treatment	age	region	supp	sex	weight	status	status01
A0001	control	219	A	6.730799	male	94.9	diseased	1
A0002	treated	218	A	7.950240	male	94.4	diseased	1
A0003	treated	214	D	5.617446	male	93.5	healthy	0
A0004	treated	194	D	8.489726	male	88.4	diseased	1
A0005	control	185	D	7.126925	female	69.0	diseased	1
A0006	treated	235	D	4.882058	female	73.0	healthy	0

A linear model of weight

```
mod_weight <- lm(weight ~ age + sex, data = dat)
```

A logistic regression model of disease status

```
mod_disease <- glm(status01 ~ treatment + age, family = binomial, data = dat)
```

The output

```
print(summary(mod_disease), digits = 3)
```

Call:

```
glm(formula = status01 ~ treatment + age, family = binomial,  
     data = dat)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.992	-1.375	0.702	0.796	1.016

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.80595	1.27517	-1.42	0.157
treatmenttreated	-0.54537	0.21523	-2.53	0.011 *
age	0.01524	0.00608	2.51	0.012 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 570.95 on 499 degrees of freedom
Residual deviance: 558.11 on 497 degrees of freedom
AIC: 564.1

Number of Fisher Scoring iterations: 4

The output

```
print(summary(mod_disease), digits = 3)
```

Call:

```
glm(formula = status01 ~ treatment + age, family = binomial,  
     data = dat)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.992	-1.375	0.702	0.796	1.016

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.80595	1.27517	-1.42	0.157
treatmenttreated	-0.54537	0.21523	-2.53	0.011 *
age	0.01524	0.00608	2.51	0.012 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 570.95 on 499 degrees of freedom
Residual deviance: 558.11 on 497 degrees of freedom
AIC: 564.1

Number of Fisher Scoring iterations: 4

The output

Lets get 'tidy output...

```
tidy(mod_disease) #tidy from the broom package
```

```
# A tibble: 3 x 5
```

	term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
1	(Intercept)	-1.81	1.28	-1.42	0.157
2	treatmenttreated	-0.545	0.215	-2.53	0.0113
3	age	0.0152	0.00608	2.51	0.0122

odds ratios

The estimates = log(odds ratios)

i.e.

$$\frac{\text{odds of outcome if have factor}}{\text{odds of outcome if dont have factor}}$$

So we get odds ratios by 'inverse logging them'.

We can remove the intercept.

```
tidy(mod_disease) %>%  
  filter(term ≠ "(Intercept)") %>%  
  mutate(OR = exp(estimate))
```

A tibble: 2 x 6

	term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>	OR <dbl>
1	treatmenttreated	-0.545	0.215	-2.53	0.0113	0.580
2	age	0.0152	0.00608	2.51	0.0122	1.02

A results table

```
tidy(mod_disease) %>%  
  mutate(OR = exp(estimate)) %>%  
    bind_cols(exp(confint_tidy(mod_disease)) %>%  
      as_tibble()  
    ) %>%  
  filter(term  $\neq$  "(Intercept)") %>%  
  select(term, OR, `conf.low`, `conf.high`, p.value)
```

term	OR	conf.low	conf.high	p.value
treatmenttreated	0.580	0.378	0.880	0.011
age	1.015	1.003	1.028	0.012

But what does it mean?

Interpreting the odds ratios

term	OR	2.5 %	97.5 %	p.value
treatmenttreated	0.580	0.378	0.880	0.011
age	1.015	1.003	1.028	0.012

Odds ratios **multiply**

Categorical predictors

How many times greater the odds of outcome are **if** the risk factor (etc) is present.

So for the treatment variable (which can be control or treatment) the odds of disease if treated are 0.58 **times greater** than if untreated (control).

Interpreting the odds ratios

term	OR	2.5 %	97.5 %	p.value
treatmenttreated	0.580	0.378	0.880	0.011
age	1.015	1.003	1.028	0.012

Odds ratios **multiply**

Numerical predictors

How many times greater the odds of outcome are for **each unit change** in the variable

So for the age variable the odds of disease are 1.015 **times greater** for each day older.

So for 3 days it's $1.015 \times 1.015 \times 1.015 \simeq 1.046$.

Things to watch out for

Factor levels

How does R know if you are predicting 'healthy' or 'diseased'?

Perfect predictors

E.g. all the males are diseased and all the females are healthy

Linear on logit 🤖

Disease risk might go up and then down

Model selection - a blank page

More help

Veterinary Epi Research - Ian Dahoo