Generalized Linear Mixed Models

Dale Barr
University of Glasgow

Overview

1. Introduction to generalized linear (mixed) models
2. Logistic regression
3. Worked example (Titanic data)

Discrete data

- categorical (dichotomous/polychotomous)
  - type of linguistic structure produced (X, Y, Z)
  - region viewed in a visual world study
  - number of items recalled out of N
  - accurate or inaccurate selection
  - hired or not hired
  - Likert scales
- counts (no. opportunities ill-defined)
  - no. of speech errors in a corpus
  - no. of turn shifts in a conversation
  - no. words in a utterance

Why not treat discrete data as continuous?

- Proportions range between 0 and 1
- Variance proportional to the mean (expected probability or rate)
- Spurious interactions due to scaling effects

Generalized linear models

- Allows use of regular linear regression by projecting the DV onto an appropriate scale
- Key elements of GLMs:
  - link function
  - variance function

data	approach	link	variance	function
binary	logistic regression	logit	binomial	`glm()`, `lme4::glmer()`
count	Poisson regression	log	Poisson	`glm()`, `lme4::glmer()`
ordinal	ordinal regression	logit	binomial	`ordinal::clm()`, `ordinal::clmm()`

Logistic regression

Odds and log odds

Bernoulli trial	An event that has a binary outcome, with one outcome typically referred to as ‘success’
proportion	A ratio of successes to the total number of Bernoulli trials, proportion of days of the week that are Wednesday is 1/7 or about .14
odds	A ratio of successes to non-successes, i.e., odds of a day being Wednesday are 1 to 6, natural odds= 1/6 = .17
log odds	The (natural) log of the odds (turns multiplicative effects into additive effects)

Properties of log odds ('logit')

$log\left(\frac{p}{1-p}\right)$ or $log\left(\frac{Y}{N-Y}\right)$

where p is a proportion, N is total trials and Y is observed successes

- Scale goes from (-) to (+)
- Scale is symmetric around zero
- If negative, means that Pr(success)(<.5)
- If positive, Pr(success)(>.5)

Logistic regression

$$\eta = \beta_0 + \beta_1 X$$

- link function: $\eta = log\left(\frac{p}{1-p}\right)$
- inverse link function: $p = \frac{1}{1+exp(-\eta)}$
- getting odds from logit: exp(η)
- variance function (binomial): $np(1-p)$

Logistic Regression

Parameters

beta_0 (Intercept)

-4 0 4

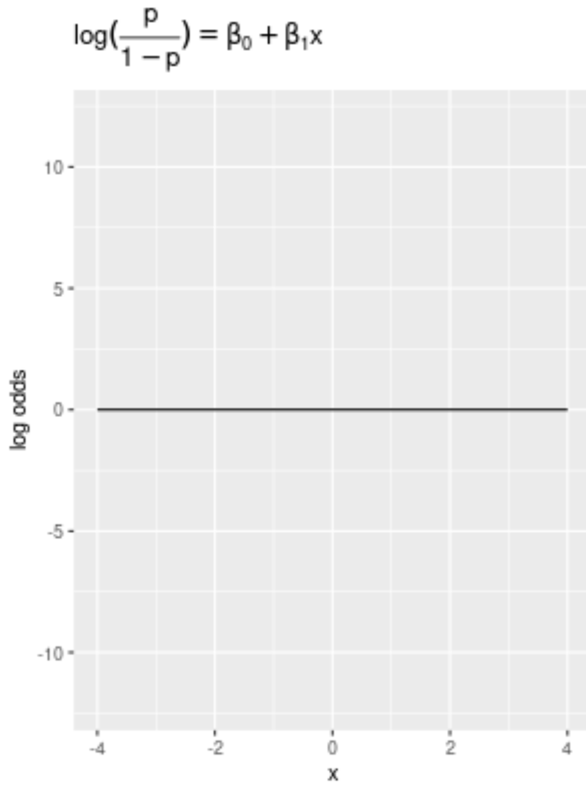
-4 -3.2 -2.4 -1.6 -0.8 0 0.8 1.6 2.4 3.2 4

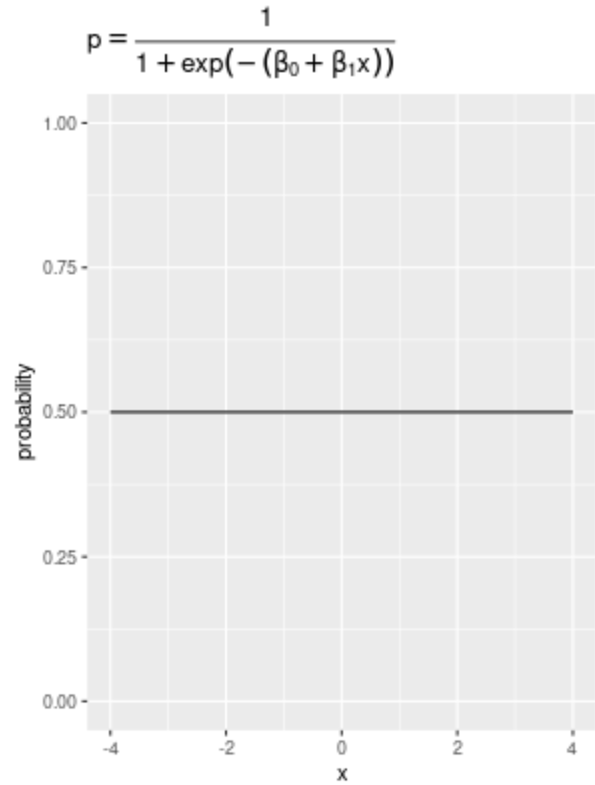
beta_1 (slope)

-4 0 4

-4 -3.2 -2.4 -1.6 -0.8 0 0.8 1.6 2.4 3.2 4

Odds Ratio (exp(beta_1)) = 1.000





Estimating logit models

- single-level data, bernoulli trials

```
mod <- glm(DV ~ IV, family = binomial(link = "logit"), ...)
```

- single-level data, binomial counts

```
mod <- glm(cbind(Y, K) ~ IV, family = binomial(link = "logit"), ...)
```

where K = N - Y

- multi-level data: same, but use `lme4::glmer()`

Worked example: Titanic data

Titanic dataset

https://www.kaggle.com/c/titanic

```
VARIABLE DESCRIPTIONS:
survival        Survival
                (0 = No; 1 = Yes)
pclass          Passenger Class
                (1st; 2nd; 3rd)
name            Name
sex             Sex
age             Age
sibsp           N Siblings/Spouses Aboard
parch           N Parents/Children Aboard
ticket          Ticket Number
fare            Passenger Fare
cabin           Cabin
embarked        Port of Embarkation
                (C = Cherbourg;
                 Q = Queenstown;
                 S = Southampton)
```

```
SPECIAL NOTES:
Pclass is a proxy for socio-economic status (SES)
 1st ~ Upper; 2nd ~ Middle; 3rd ~ Lower

Age is in Years; Fractional if Age less than One (1)
 If the Age is Estimated, it is in the form xx.5

With respect to the family relation variables (i.e. sibsp and parch)
some relations were ignored.  The following are the definitions used
for sibsp and parch.

Sibling:  Brother, Sister, Stepbrother, or Stepsister of Passenger
           Aboard Titanic
Spouse:   Husband or Wife of Passenger Aboard Titanic
           (Mistresses and Fiances Ignored)
Parent:   Mother or Father of Passenger Aboard Titanic
Child:    Son, Daughter, Stepson, or Stepdaughter of Passenger
           Aboard Titanic

Other family relatives excluded from this study include cousins,
nephews/nieces, aunts/uncles, and in-laws.  Some children travelled
only with a nanny, therefore parch=0 for them.  As well, some
travelled with very close friends or neighbors in a village, however,
the definitions do not support such relations.
```

import

```r
library("tidyverse")

dat <- readxl::read_excel("titanic4.xls")
glimpse(dat)
```

```
Rows: 1,309
Columns: 13
$ pclass    <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
$ survived  <dbl> 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1, 0, …
$ name      <chr> "Allen, Miss. Elisabeth Walton", "Allison, Master. Hudson Tr…
$ sex       <chr> "female", "male", "female", "male", "female", "male", "femal…
$ age       <dbl> 29.0000, 0.9167, 2.0000, 30.0000, 25.0000, 48.0000, 63.0000,…
$ sibsp     <dbl> 0, 1, 1, 1, 1, 0, 1, 0, 2, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, …
$ parch     <dbl> 0, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, …
$ ticket    <chr> "24160", "113781", "113781", "113781", "113781", "19952", "1…
$ fare      <dbl> 211.3375, 151.5500, 151.5500, 151.5500, 151.5500, 26.5500, 7…
$ cabin     <chr> "B5", "C22 C26", "C22 C26", "C22 C26", "C22 C26", "E12", "D7…
$ embarked  <chr> "S", "S", "S", "S", "S", "S", "S", "S", "S", "C", "C", "C", …
$ boat      <chr> "2", "11", NA, NA, NA, "3", "10", NA, "D", NA, NA, "4", "9",…
$ home.dest <chr> "St Louis, MO", "Montreal, PQ / Chesterville, ON", "Montreal…
```

survival by passenger sex

```
dat |>
  count(survived, sex)
```

```
# A tibble: 4 × 3
  survived sex        n
     <dbl> <chr>  <int>
1        0 female   127
2        0 male     682
3        1 female   339
4        1 male     161
```

```
dat |>
  group_by(sex) |>
  summarise(p = mean(survived),
            Y = sum(survived),
            N = n(), .groups="drop")
```

```
# A tibble: 2 × 4
  sex        p     Y     N
  <chr>  <dbl> <dbl> <int>
1 female 0.727   339   466
2 male   0.191   161   843
```

survival by passenger sex (model)

```r
mod <- glm(survived ~ sex, binomial(link = "logit"),  dat)
summary(mod)
```

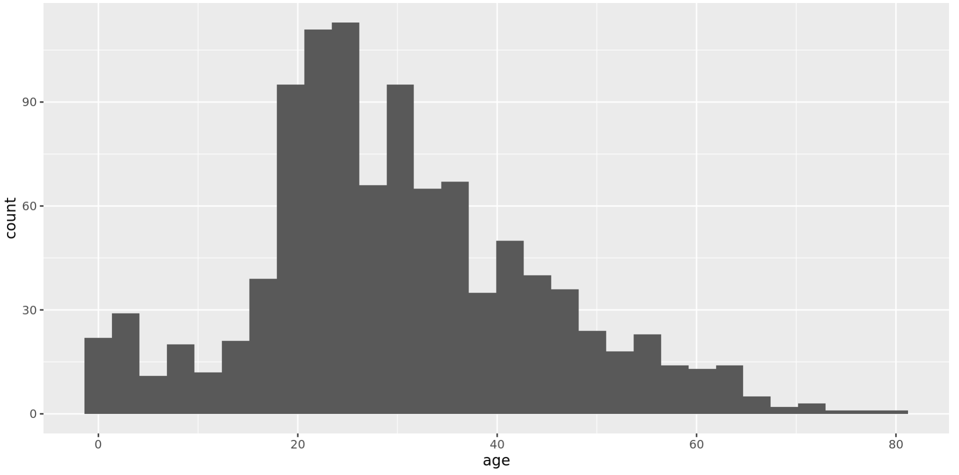
```
Call:
glm(formula = survived ~ sex, family = binomial(link = "logit"),
    data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6124  -0.6511  -0.6511   0.7977   1.8196

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.9818     0.1040   9.437   <2e-16 ***
sexmale      -2.4254     0.1360 -17.832   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1741.0  on 1308  degrees of freedom
Residual deviance: 1368.1  on 1307  degrees of freedom
AIC: 1372.1
```

age and survival

```r
## lots of NAs
dat |>
  count(f = is.na(age))
```

```
# A tibble: 2 × 2
  f         n
  <lgl> <int>
1 FALSE  1046
2 TRUE    263
```

```r
ggplot(dat, aes(age)) +
  geom_histogram()
```

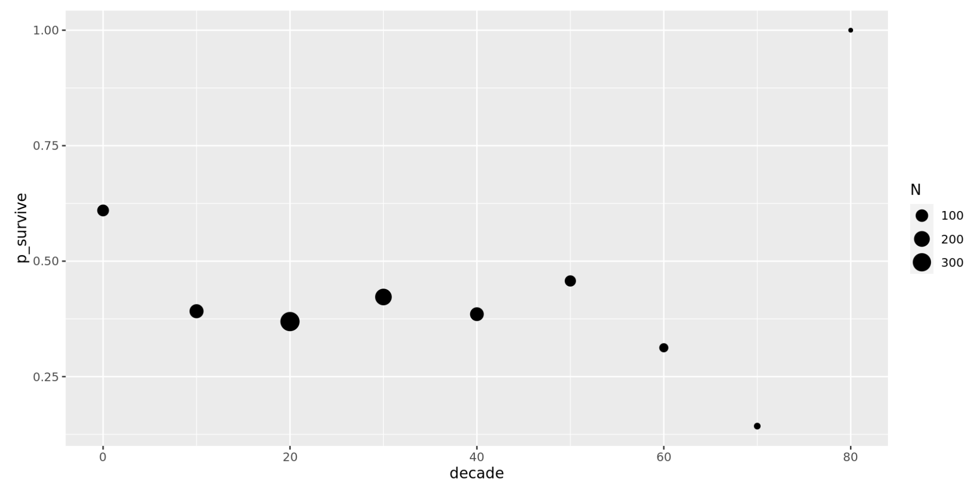


binning the data

```r
dat2 <- dat |>
  filter(!is.na(age)) |>
  mutate(decade = floor(age / 10) * 10) |>
  group_by(decade) |>
  summarise(p_survive = mean(survived),
            N = n(),
            .groups = "drop")

dat2
```

```
# A tibble: 9 × 3
  decade p_survive     N
   <dbl>     <dbl> <int>
1      0     0.610    82
2     10     0.392   143
3     20     0.369   344
4     30     0.422   232
5     40     0.385   135
6     50     0.457    70
7     60     0.312    32
8     70     0.143     7
9     80     1         1
```

```r
g <- ggplot(dat2, aes(decade, p_survive)) +
  geom_point(aes(size = N))

g
```



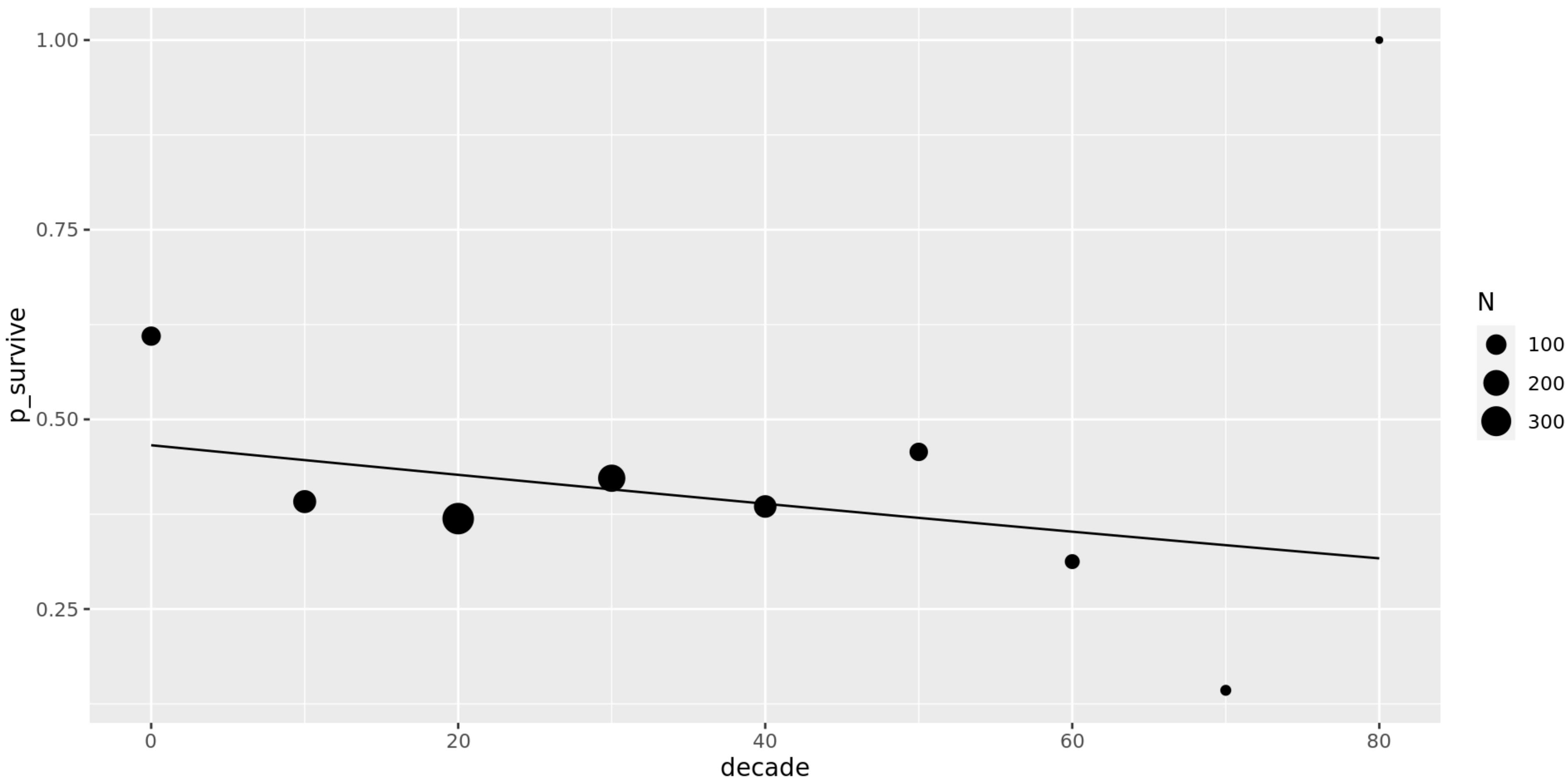
estimate

```r
mod <- glm(survived ~ age, binomial(link = "logit"), dat)
summary(mod)
```

```

Call:
glm(formula = survived ~ age, family = binomial(link = "logit"),
    data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1189  -1.0361  -0.9768   1.3187   1.5162

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.136531   0.144715  -0.943   0.3455
age         -0.007899   0.004407  -1.792   0.0731 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1414.6  on 1045  degrees of freedom
Residual deviance: 1411.4  on 1044  degrees of freedom
  (263 observations deleted due to missingness)
AIC: 1415.4
```

plot

```
newdat <- tibble(age = seq(0, 80, .2))
## see ?predict.glm
my_pred <- predict(mod, newdat, type = "response")

dat3 <- newdat |>
  mutate(p_survive = my_pred)

g + geom_line(aes(x = age, y = p_survive), data = dat3)
```



1.00
0.75
0.50
0.25
p_survive
0
20
40
60
80
decade
N
100
200
300