

Generalized linear models (GLM)

Nora Wickelmaier

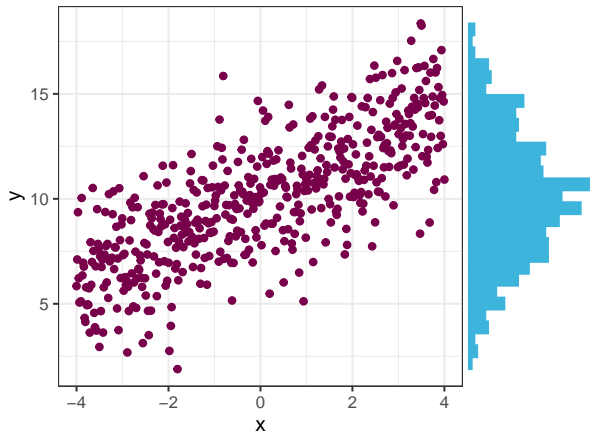
Last modified: November 5, 2024

Outline

- ① Logistic regression
- ② Poisson regression
- ③ Overdispersion

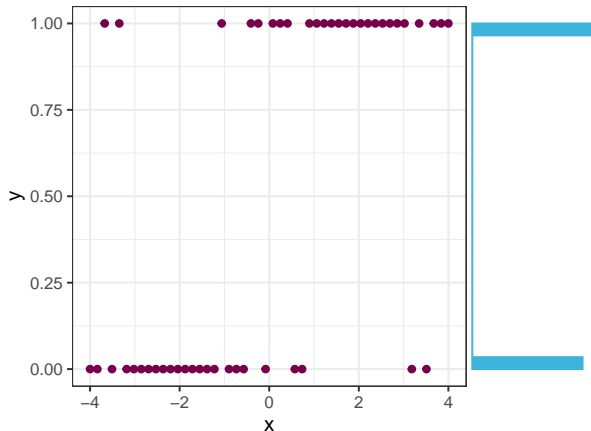
Linear regression

- Constant change in predictor leads to constant change in response variable
- Implies that response variable can vary indefinitely in both directions (or only varies by a relative small amount)






Extending linear regression for arbitrary distributions

- Generalized linear models allow for response variables that have **arbitrary distributions**
- An arbitrary function of the response variable (the link function) varies linearly with the predictors (rather than assuming that the response itself must vary linearly)





Intuition I

- Example 1: Model that predicts probability of making yes/no choice
 - How likely is it for a given person to go to the beach as a function of temperature?
 - A model might predict, for example, that a change in 10 degrees makes a person two times more or less likely to go to the beach
 - That means the odds are doubling: from 2:1 odds to 4:1 odds and so on

Temperature	10° C	20° C	30° C
			
Odds	1 : 1	2 : 1	4 : 1

Intuition II

- Example 2: Model that predicts a certain count
 - A model might predict a constant rate of increased beach attendance (e.g., an increase of 10 degrees leads to a doubling in beach attendance, and a drop of 10 degrees leads to a halving in attendance)
 - This prediction would be independent of the size of the beach

Temperature	20° C	30° C
		
Number	5	10

Generalized linear models

- A generalized linear model is defined by

$$g(E(y)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k,$$

where $g()$ is the link function that links the mean to the linear predictor

- The response y is assumed to be independent and to follow a distribution from the exponential family
- In R, a GLM is fitted by

```
glm(y ~ x1 + x2 + ... + xk, family(link), data)
```

Families

- Each response distribution admits a variety of link functions to connect the mean with the linear predictor:

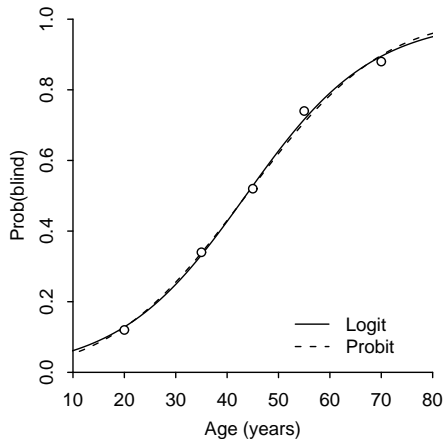
<i>## Family name</i>	<i>Link functions</i>
binomial	logit, probit, log, cloglog
gaussian	identity, log, inverse
Gamma	identity, inverse, log
inverse.gaussian	$1/\mu^2$, identity, inverse, log
poisson	log, identity, sqrt
quasi	logit, probit, cloglog, identity, inverse, log, $1/\mu^2$, sqrt

- A GLM is a specific combination of a response distribution, a link function, and a linear predictor

① Logistic regression

Binomial regression

- Logit or probit models are special cases of GLMs for binomial response variables
- Artificial example: congenital eye disease



Logit model

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 AGE$$

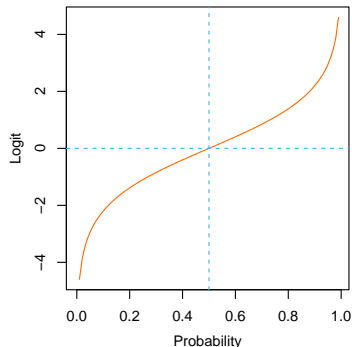
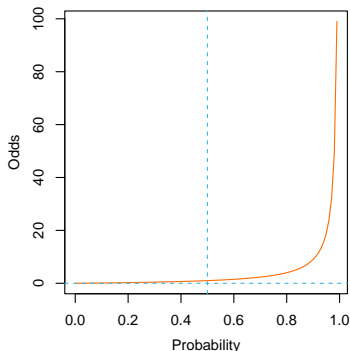
Probit model

$$\Phi^{-1}(p) = \beta_0 + \beta_1 AGE$$

Probabilities, Odds, Logit

Probabilities, Odds and Logits are measures for the tendency that an event occurs

- $Odds = \frac{p}{1-p}$
number of expected events per complementary event
- $logit = \log(Odds)$
logarithm of odds



p	0.01	0.05	0.33	0.50	0.66	0.95	0.99
Odds	1/99	1/19	0.49	1	1.94	19	99
logit p	-4.60	-2.94	-0.71	0	0.66	2.94	4.60

Logit model / logistic regression

- We want to model the probability of y with the logistic function

$$p(y) = \frac{1}{1 + e^{-z}} \quad \text{with } y \sim \text{Binom}(n, p)$$

- How do we get the logit model $\log \frac{p}{1-p} = \beta_0 + \beta_1 x$ from that?

Logit model / logistic regression

- We want to model the probability of y with the logistic function

$$p(y) = \frac{1}{1 + e^{-z}} \quad \text{with } y \sim \text{Binom}(n, p)$$

- How do we get the logit model $\log \frac{p}{1-p} = \beta_0 + \beta_1 x$ from that?

$$\begin{aligned} \log \left(\frac{p}{1-p} \right) &= \log(p) - \log(1-p) = \log \left(\frac{1}{1 + e^{-z}} \right) - \log \left(1 - \frac{1}{1 + e^{-z}} \right) \\ &= \log(1) - \log(1 + e^{-z}) - \log(e^{-z}) + \log(1 + e^{-z}) \\ &= -\log(e^{-z}) \\ &= z := \beta_0 + \beta_1 x \end{aligned}$$

$$\text{with } 1 - \frac{1}{1+e^{-z}} = \frac{e^{-z}}{1+e^{-z}}$$

Refresher: Logarithm rules

Product: $\log(xy) = \log x + \log y$

Quotient: $\log \frac{x}{y} = \log x - \log y$

Power: $\log(x^p) = p \log x$

Root: $\log \sqrt[p]{x} = \frac{\log x}{p}$

Fitting binomial regression models

```
dat <- data.frame(x = c(20, 35, 45, 55, 70),  
                  n = rep(50, 5),  
                  y = c(6, 17, 26, 37, 44))  
  
glm1 <- glm(cbind(y, n - y) ~ x, family = binomial, data = dat)  
glm2 <- glm(cbind(y, n - y) ~ x, family = binomial(probit), data = dat)  
  
# Parameter estimates  
summary(glm1)  
  
# Interpretation as odds ratio  
exp(coef(glm1))  
# --> Odds of going blind are increased by a factor  
# of 1.08 when age increases by one year
```

Goodness of fit and predictions

```
# Compare to saturated model
glms <- glm(cbind(y, n - y) ~ factor(x), family = binomial, data = dat)

# Likelihood ratio test
anova(glml, glms, test = "Chisq")

# Predictions based on new observations (see ?predict.glm)
newx <- 0:100
predict(glml, newdata = data.frame(x = newx), type = "response")
```


Exercise

- In a psychophysical experiment two LEDs are presented to a subject: a standard with 40 cd/m² and a comparison with varying intensities
- The subject is supposed to say which stimulus is brighter; each comparison is presented 40 times

x (cd/m ²)	37	38	39	40	41	42	43
y (positiv)	2	3	10	25	34	36	39

- Estimate parameters c and a of the logistic psychometric function

$$p_{pos} = \frac{1}{1 + \exp\left(-\frac{x - c}{a}\right)}$$

using `glm()` with $\text{logit}(p_{pos}) = \beta_0 + \beta_1 x$ where $a = 1/\beta_1$ and $c = -\beta_0/\beta_1$.

Exercise

- Calculate the intensity x for which $p_{pos} = 0.5$ (Point of Subjective Equality, PSE)
- Create a plot for the probability to give a positive answer depending on the intensity of the comparison
- Use `predict()` to obtain the predicted values and add the logistic psychometric function to the plot
- Use `abline()` to add parameter c to the plot
- Use a likelihood ratio test to assess how well the model fits the data
- Is there reason to assume that there is any overdispersion?

② Poisson regression

Poisson distribution

- The Poisson distribution is popular for modeling the number of times an event occurs in an interval of time or space

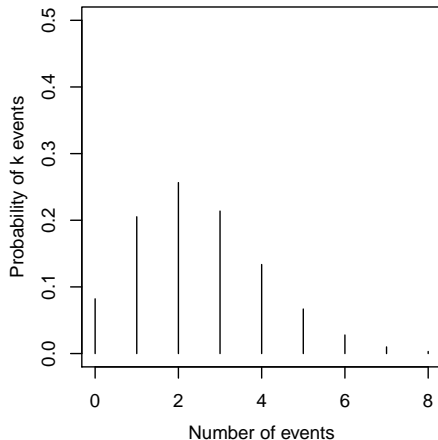
$$P(X = k) = \exp(-\lambda) \frac{\lambda^k}{k!}$$

where k is the number of events in a certain interval and λ is the average number of events per interval

- $X \sim \text{Poisson}(\lambda)$ with $E(X) = \lambda$ and $\text{Var}(X) = \lambda$

Poisson distribution

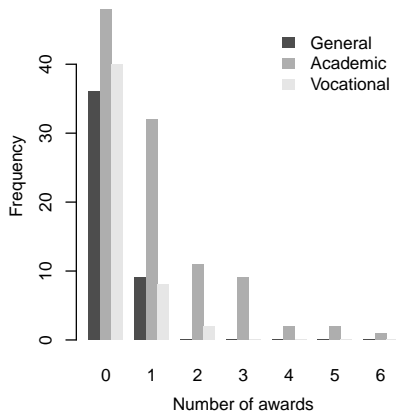
Poisson distribution with probability of events for $\lambda = 2.5$



```
x <- 0:8  
px <- dpois(x, lambda = 2.5)  
plot(px ~ x, type = "h")
```

Poisson regression

Poisson regression is used to model count variables¹



Variables

- Number of awards earned by students at one high school
- Type of program in which student was enrolled (vocational, general, or academic)
- Score on students' final exam in math

¹Simulated dataset from https://stats.idre.ucla.edu/stat/data/poisson_sim.csv (UCLA: Statistical Consulting Group, 2024)

Poisson regression

- We estimate a poisson regression using a generalized linear model with `family = poisson` and link function `log`

$$g(E(y)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\log(\mu) = \beta_0 + \beta_1 \textit{prog} + \beta_2 \textit{math}$$

with $y_i \sim \text{Poisson}(\lambda)$

Poisson regression

```
# Read data
dat <- read.csv("poisson_sim.csv")

# Define factors
dat$prog <- factor(dat$prog, levels = 1:3,
                  labels = c("General", "Academic", "Vocational"))
dat$id <- factor(dat$id)

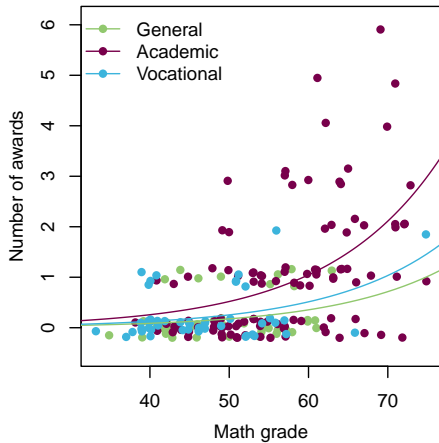
# Fit poisson regression
m1 <- glm(num_awards ~ prog + math, family = poisson, data = dat)
summary(m1)

# Evaluate goodness-of-fit
1 - pchisq(m1$deviance, df = m1$df.residual)
```


Poisson regression

- The results show that the model fits the data with $G^2(196) = 189.45$, $p = 0.6182$
- The expected number of awards when in the academic program is $\exp(1.0839) = 2.96$ times the expected number for the general program when math grade is held constant
- The expected number of awards increases by a factor of $\exp(0.0702) = 1.07$ when math grade increases by one unit (and program is held constant)

Predictions of the poisson regression



Variables

- Number of awards earned by students at one high school
- Type of program in which student was enrolled (vocational, general, or academic)
- Score on students' final exam in math

③ Overdispersion

Overdispersion

- Overdispersion is the presence of greater variability in a data set than would be expected based on a given statistical model
- It means that the underlying distributional assumptions might be violated
- The binomial and the poisson distribution are both less flexible than, e. g., the normal distribution, since they only have one free parameter and, therefore, the variance cannot be adjusted independently of the mean
- We can include a so-called overdispersion parameter φ into both models
- For the poisson regression, instead of assuming $E(y) = Var(y) = \mu$, we model $Var(y) = \varphi\mu$

Overdispersion

- In R this is done by changing the family argument to quasibinomial or quasipoisson.

```
# Fit poisson regression  
m2 <- glm(num_awards ~ prog + math, family = quasipoisson, data = dat)  
summary(m2)
```

```
# --> Results show that estimated parameters are still the same,  
# but standard errors are slightly higher
```

Some things to consider

- When using `family = "quasipoisson"` or `family = "quasibinomial"`, likelihood-ratio tests are not meaningful anymore (even though R will let you do them)
- Goodness-of-fit tests are not necessarily meaningful for *continuous* predictors for poisson and binomial regression, so use with caution (see, e. g., <http://thestatsgeek.com/2014/04/26/deviance-goodness-of-fit-test-for-poisson-regression/>)

Exercise²

- Fit a regression model to the `Affairs` data set from the `AER` package in R
- The variable `affairs` is the number of extramarital affairs in the past year and is our response variable
- Include the variables `gender`, `age`, `yearsmarried`, `children`, `religiousness`, `education` and `rating` as predictors
- `religiousness` ranges from 1 (anti) to 5 (very) and `rating` is a self rating of the marriage, ranging from 1 (very unhappy) to 5 (very happy)
- Assess the Goodness-of-fit using the deviance
- Assess overdispersion and decide if a model with an extra dispersion parameter might be indicated
- Compare the confidence intervals for the estimated parameters for both models

²Inspired by

https://rstudio-pubs-static.s3.amazonaws.com/1047952_9306ae04c1de4543812af559d777dd72.html

References

UCLA: Statistical Consulting Group. (2024). Poisson Regression – R Data Analysis Examples [accessed 2024-10-14].
<https://stats.oarc.ucla.edu/r/dae/poisson-regression/>