Multiple regression and generalized linear models (GLM)

Nora Wickelmaier

November 7, 2022

Outline

1. Multiple linear regression

2. Generalized linear model
   - Logistic regression
   - Poisson regression

Exercise

- Create two vectors x and y with 100 observations each and $X \sim N(1,1)$ and $Y \sim N(2,1)$.
- Create a data frame with variables `id`, `group` and `score`. x and y are your score values.
- Conduct a t test assuming that X and Y are independent having the same variances.
- Then use the function `aov()` to compute an analysis of variance for these data.
- Use then function `lm()` for a linear regression with predictor `group` and dependent variable `score`.
- Compare your results.

Extending simple linear regression

Additional predictors	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \varepsilon$
Nonlinear models	$\log y = \beta_0 + \beta_1 \log x + \varepsilon$
Nonadditive models	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$
Generalized linear models	$g(E(y)) = \beta_0 + \beta_1 x$
Mixed-effects models	$y = \beta_0 + \beta_1 x_1 + \beta_2 \mathit{time} + \upsilon_0 + \upsilon_1 \mathit{time} + \varepsilon$

. . .

❶ Multiple linear regression

Multiple linear regression

- Empirical observations consist of tuples for each observation unit

  $$(y_i, x_{i1}, \ldots, x_{ip}) \ \text{ with } \ i = 1, \ldots, n$$

  and we get the stochastical model

  $$\begin{aligned} & y_i = \beta_0 + \beta_1 \cdot x_{i1} + \ldots + \beta_p \cdot x_{ip} + \varepsilon_i \\ & \varepsilon_i \sim N(0, \sigma^2) \text{ i.i.d.} \end{aligned}$$

  which transfers to

  $$y_i \sim N(\mu_i, \sigma^2) \ \text{ with } \ \mu_i = \beta_0 + \beta_1 \cdot x_{i1} + \ldots + \beta_p \cdot x_{ip}$$

- The criterion variable y is always a metric variable, whereas the predictor variables $x_1, \ldots, x_p$ can be either metric or categorical variables, or both

Example: Multiple linear regression

- We are fitting the following model

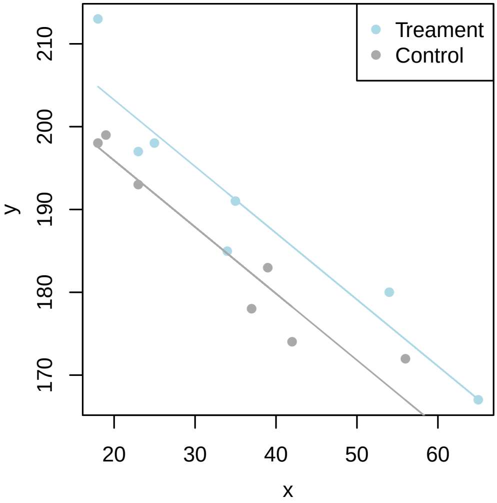
$$y_{ij} = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot z_j + \varepsilon_{ij}$$

  with $i = 1 \dots N$ and $j = 1, 2$ for two groups

- This means that we have one dummy variable for z which takes the values 0 and 1
- Hence, we get the two models

$$\begin{aligned} y_{i1} &= \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot 0 + \varepsilon_{ij} = \beta_0 + \beta_1 \cdot x_i + \varepsilon_{ij} \\ y_{i2} &= \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot 1 + \varepsilon_{ij} = (\beta_0 + \beta_2) + \beta_1 \cdot x_i + \varepsilon_{ij} \end{aligned}$$

Example: Multiple linear regression



Example: Multiple linear regression

```r
dat <- data.frame(
  x = c(18,23,25,35,65,54,34,56,72,19,23,42,18,39,37),
  y = c(213,197,198,191,167,180,185,172,153,199,193,
        174,198,183,178),
  z = rep(c("treatment", "control"), c(7, 8))
)

aggregate(y ~ z, dat, mean)
```

Example: Multiple linear regression

- We can now use the parameters to calculate adjusted means for the two groups
- The observed means are $\bar{x}_{treat} = 190.14$ and $\bar{x}_{contr} = 181.25$
- The adjusted means correspond to

$$\begin{aligned} \bar{x}_{contr} &= \beta_0 \\ \bar{x}_{treat} &= \beta_0 + \beta_2 \end{aligned}$$

  These are the means for a value of $x = 0$ which should have a meaningful interpretation

- Hence, it might be indicated to center x

Example: Multiple linear regression

```r
dat$xc <- dat$x - mean(dat$x)

lm2 <- lm(y ~ xc + z, dat)
summary(lm2)

# adjusted means
coef(lm2)[1]
coef(lm2)[1] + coef(lm2)[3]
```

Exercise

- The data set `cars` contains speed and stopping distances of 50 cars
- Estimate the regression model

$$dist_i = \beta_0 + \beta_1 speed_i + \varepsilon_i$$

- How much variance of the stopping distances is explained by speed?
- Look at the residuals of the model. Are there any systematic deviances?
- Now estimate the model

$$dist_i = \beta_0 + \beta_1 speed_i + \beta_2 speed_i^2 + \varepsilon_i$$

  Hint: Use `I(speed^2)` in the model formula in `R`
- Which model fits the data better?

2 Generalized linear model

Logistic regression

Poisson regression

Intuition I

- In linear regerssion, a constant change in a predictor leads to a constant change in the response variable
  → implies that response variable can vary indefinitely in both directions (or only varies by a relative small amount)
- Generalized linear models allow for response variables that have **arbitrary distributions** (rather than simply normal distributions), and for an arbitrary function of the response variable (the link function) to vary linearly with the predictors (rather than assuming that the response itself must vary linearly)

Intuition II

- Example 1: Model that predicts probability of making yes/no choice
  - A model that predicts the likelihood of a given person going to the beach as a function of temperature
  - A reasonable 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 th odds are doubling: from 2:1 odds to 4:1 odds and so on
- Example 2: Model that predicts a certain count
  - A realistic model would 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

https://en.wikipedia.org/wiki/Generalized_linear_model

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

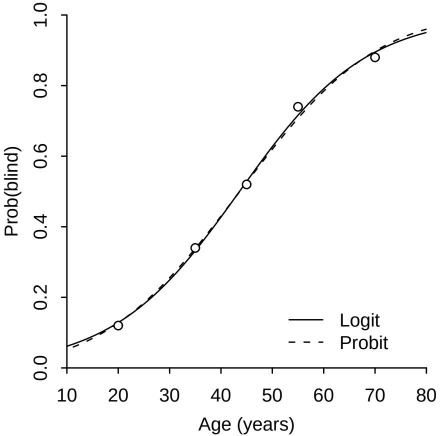
- Each response distribution admits a variety of link functions to connect the mean with the linear predictor:

```
## Family name          Link functions
   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

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

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

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))

glml <- glm(cbind(y, n - y) ~ x, binomial, dat)
glmp <- glm(cbind(y, n - y) ~ x, binomial(probit), dat)

# Parameter estimates
summary(glml)

# Interpretation as odds ratio
exp(coef(glml))
# --> 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), binomial, dat)

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

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

Exercise

- In a psychophysical experiment two LEDs are presented to a subject: a standard with 40 cd/m^2 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

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

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

Poisson distribution

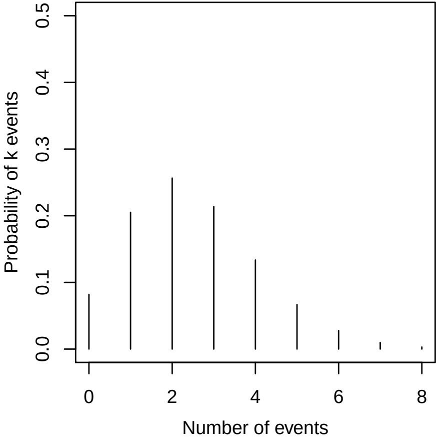
- The Poisson distribution is popular for modelling the number of times an event occurs in an interval of time or space

$$P(k \text{ events in an interval}) = \exp(-\lambda)\frac{\lambda^k}{k!}$$

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

Poisson distribution

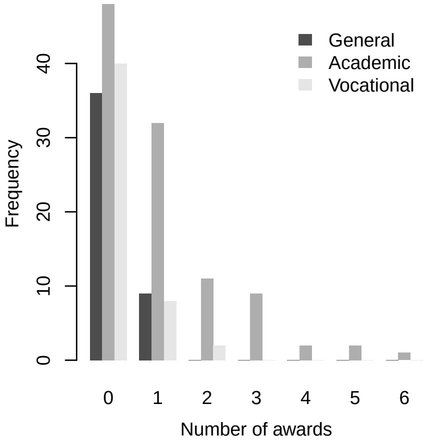
Poisson distribution with probability of events for $\lambda = 2.5$



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

Poisson regression

Poisson regression is used to model count variables[1].



- 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

[1] Simulated dataset from https://stats.idre.ucla.edu/stat/data/poisson_sim.csv

Poisson regression

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

$$
\begin{aligned}
g(E(y)) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 \\
\log(\mu) &= \beta_0 + \beta_1 prog + \beta_2 math
\end{aligned}
$$

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

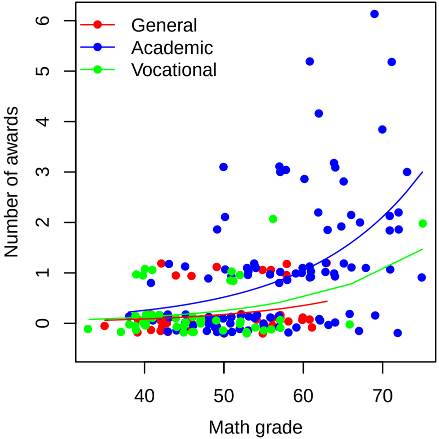
Poisson regression

```
# 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$deviance, m1$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



- 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

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

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