Lecture 1: From Linear Models to Generalized Linear Models

Dankmar Böhning

Southampton Statistical Sciences Research Institute
University of Southampton, UK

S^{3}RI, 11 - 12 December 2014

The simple regression model

Case study: BELCAP

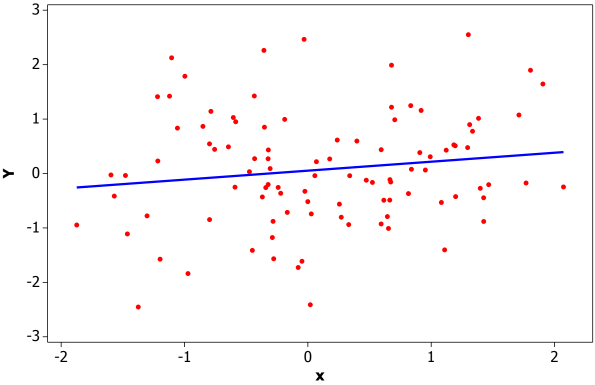
The various problems of using a simple regression model

the three elements of a GLM

The simple regression model

$$Y_i = \alpha + \beta x_i + \epsilon_i$$

- Y_i is a **response** (dependent variable, clinical endpoint, outcome) for observation i the
- x_i is a **covariate** (treatment, intervention) for observation i (might be continuous or categorical)
- α and β are unknown parameters in the model
- ϵ_i is a mean-zero normal random error: $\epsilon_i \sim N(0, \sigma^2)$


3
2
1
0
-1
-2
-3
Y
-2
-1
0
1
2
x

The simple regression model

$$Y_i = \alpha + \beta x_i + \epsilon_i$$

- testing the effect of covariate x is done by the size of the estimate $\hat{\beta}$ of β

$$t = \frac{\hat{\beta}}{s.e.(\hat{\beta})}$$

- if $|t| > 1.96$ covariate effect is **significant**

The simple regression model for several covariates

$$Y_i = \alpha + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \epsilon_i$$

- where $x_{1i}, \cdots, x_{pi}$ are the **covariates of interest**
- testing the effect of covariate x_j is done by the size of the estimate $\hat{\beta}_j$ of β_j

$$t_j = \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)}$$

- if $|t_j| > 1.96$ covariate effect is **significant**

Case study: BELCAP

- Dental epidemiological study.
- A prospective study of school-children from an urban area of Belo Horizonte, Brazil.
- The Belo Horizonte caries prevention (BELCAP) study.
- The aim of the study was to compare different methods to prevent caries.

- Children selected were all 7 years-old and from a similar socio-economic background.
- Interventions:
  - Control (3),
  - **O**ral **H**ealth **E**ducation (1),
  - **E**nrichment of the **S**chool **D**iet with rice bran (4),
  - **M**outh**W**ash (5),
  - **O**ral **HY**giene (6),
  - **ALL** four methods together (2).
- Interventions were cluster randomised to 6 different schools.

- Response, or outcome variable = DMFS index (Number of decayed, missing or filled teeth surfaces) at the end of study
- lesion of the tooth surfaces were also included in the index; graded as
  - 0 = healthy
  - 1 = light chalky spot
  - 2 = thin brown-black line
  - 3 = damage, not larger than 2mm wide
  - 4 = damage, wider than 2mm

  in the BELCAP study a lesion graded 1-4 contributed 1 to the DMFS index DMFS index was calculated at the start of the study and 2 years later (end of study). Only the 8 deciduous molars were considered.
- Potential confounders: sex (female 0 male 1), ethnicity.
- Data analysed by Böhning et al. (1999,*Journ. Royal Statist. Soc. A*).

The simple regression model for BELCAP

with $Y = DMFSe$:

$$Y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + \epsilon_i$$

or **more illustrative**

$$DMFSe_i = \alpha + \beta_1 OHE_i + \beta_2 ALL_{2i} + \beta_4 ESD_i + \beta_5 MW_i + \beta_6 OHY_i + \epsilon_i$$

- $OHE_i = \begin{cases} 1 & \text{if child } i \text{ is in intervention OHE} \\ 0 & \text{otherwise} \end{cases}$
- $ALL_i = \begin{cases} 1 & \text{if child } i \text{ is in intervention ALL} \\ 0 & \text{otherwise} \end{cases}$
- ...

analysis of BELCAP study using simple regression model

covariate	$\hat{\beta}_j$	$s.e.(\hat{\beta}_j)$	t_j	P-value
OHE	-1.795541	.5529044	-3.25	0.001
ALL	-3.826656	.5494779	-6.96	0.000
ESD	-1.711230	.5440699	-3.15	0.002
MW	-2.398767	.5231845	-4.58	0.000
OHY	-2.470469	.5540789	-4.46	0.000
α	6.779412	.3818337	17.75	0.000

The simple regression model

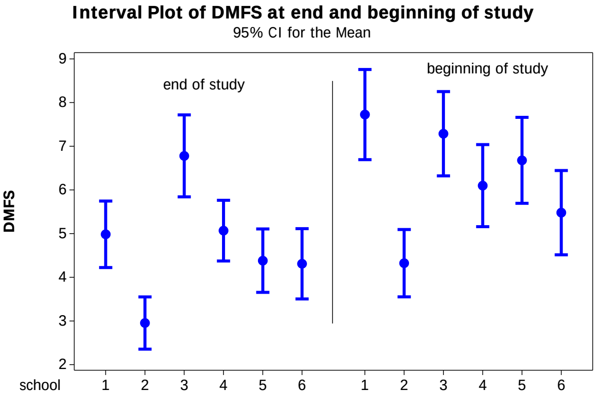
Case study: BELCAP

The various problems of using a simple regression model

the three elements of a GLM

what is problematic with this analysis: problem 1

- not all intervention schools have the same DMFS as baseline
- hence, schools with a low DMFS value at baseline will appear to have the better intervention


Interval Plot of DMFS at end and beginning of study
95% CI for the Mean
DMFS
end of study
beginning of study
9
8
7
6
5
4
3
2
school
1
2
3
4
5
6
1
2
3
4
5
6

solution: use baseline value in the model

$$DMFSe_i = \alpha + \beta_1 OHE_i + \beta_2 ALL_{2i} + \beta_4 ESD_i + \beta_5 MW_i + \beta_6 OHY_i$$

$$+\beta_7 DMFSb_i + \epsilon_i$$

- where $DMFSb_i$ is the value of the DMFS for child i at baseline

analysis of BELCAP study using simple regression model including the DMFS at baseline

covariate	$\hat{\beta}_j$	$s.e.\hat{\beta}_j$	t_j	P-value
OHE	-1.992079	.4593379	-4.34	0.000
ALL	-2.499844	.4617501	-5.41	0.000
ESD	1.179293	.4527593	-2.60	0.009
MW	-2.125991	.4347758	-4.89	0.000
OHY	-1.661519	.4621846	-3.59	0.000
DMFSb	.447653	.0237036	18.89	0.000
α	3.51747	.3611204	9.74	0.000

what is problematic with this analysis: problem 2

- DMFS is count variable, hence it cannot be negative
- there is no guarantee that the **fitted value**

$$\widehat{DMFSe}_i = \hat{\alpha} + \hat{\beta}_1 OHE_i + \hat{\beta}_2 ALL_{2i} + \hat{\beta}_4 ESD_i + \hat{\beta}_5 MW_i + \hat{\beta}_6 OHY_i + \hat{\beta}_7 DMFSb_i$$

is **nonnegative**

solution: use appropriate link function

to achieve always nonnegative values for fitted values use

$$E(DMFSe_i) = \exp[\alpha + \beta_1 OHE_i + \beta_2 ALL_{2i} + \beta_4 ESD_i + \beta_5 MW_i + \beta_6 OHY_i + \beta_7 DMFSb_i]$$

or in general

$$E(Y_i) = \exp[\alpha + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}]$$

solution: use appropriate link function

$$E(Y_i) = \exp[\alpha + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}]$$

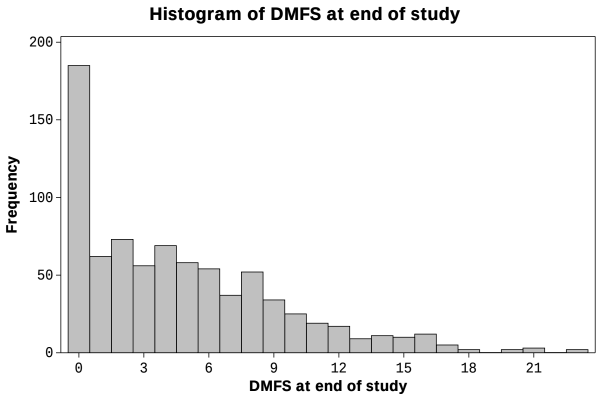
can also be written as

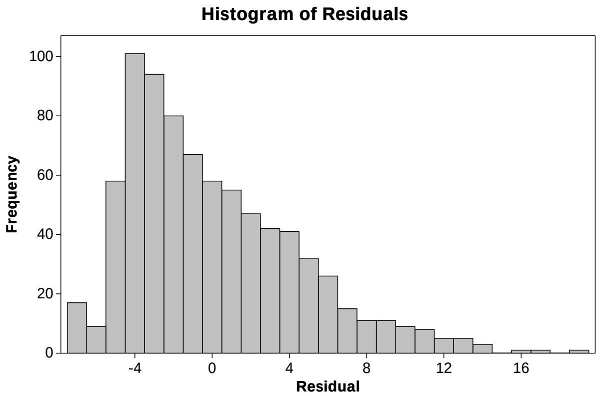
$$\log E(Y_i) = \alpha + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$

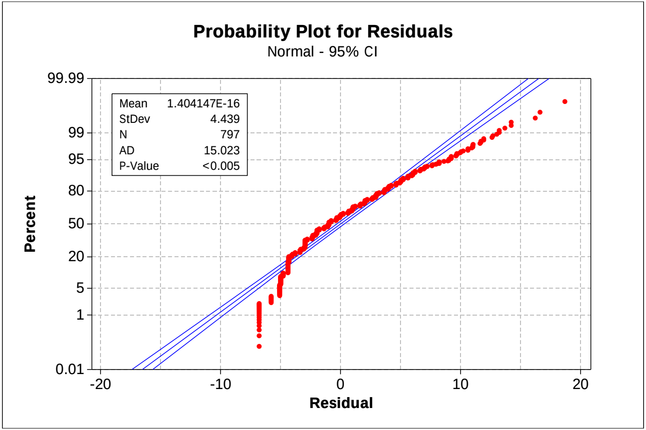
- log is called a **link function** here the log-link and the associated model is called **log-linear** model
- other valid link function would be $\sqrt{E(Y_i)}$ or similar
- the log-link is popular

what is problematic with this analysis: problem 3

- DMFS is count variable, not likely to have a **normal distribution**
- actually, only $\hat{\epsilon}_i = DMFSe_i - \widehat{DMFSe}_i$ is required to be **normal**, but also unlikely


Histogram of DMFS at end of study
Frequency
200
150
100
50
0
0
3
6
9
12
15
18
21
DMFS at end of study


Histogram of Residuals
Frequency
100
80
60
40
20
0
-4
0
4
8
12
16
Residual


Probability Plot for Residuals
Normal - 95% CI
Mean 1.404147E-16
StDev 4.439
N 797
AD 15.023
P-Value <0.005
Percent
99.99
99
95
80
50
20
5
1
0.01
-20
-10
0
10
20
Residual

The simple regression model

Case study: BELCAP

The various problems of using a simple regression model

the three elements of a GLM

elements of a generalized linear model

for study data like the BELCAP study we need to deviate from simple linear regression using **a generalized linear model** approach

1. an appropriate **linear predictor** $\alpha + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$
2. an appropriate **link function** which connects the linear predictor with the mean of the response

$$h(E(Y_i)) = \alpha + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$

3. and an appropriate **error distribution** for the response Y (other than normal)