Generalized Linear Models

Lecture 1. Introduction. Background

Statistical model

A statistical model is a class of mathematical models, with random error.

$$y = f(\beta, x) + \varepsilon$$

$y \rightarrow$ dependent variable (response, outcome), the one we want to describe (to explain, to predict)

$x \rightarrow$ explanatory variables (independent variables, factors or covariates), the ones we use to explain (to describe or to predict) the dependent variable

$\beta \rightarrow$ unknown parameters

$\varepsilon \rightarrow$ random errors, usual assumption: iid – **i**ndependent **i**dentically **d**istributed.

$f(\cdot) \rightarrow$ known function, usually linear to parameters

$= \rightarrow$ optimization (least squares, maximum likelihood)

A statistical model is a stochastic model that contains parameters (unknown constants) that need to be estimated based on assumptions about the model and the observed data.

History

- Normal distribution: linear regression (Legendre, Gauss; 19. century), analysis of variance (Fisher, 1920-1935).
- Likelihood (Fisher, 1922). Binomial distribution: dilution analysis, log-log transformation (Fisher, 1922).
  Family of exponential distributions (Fisher, 1934).
- Binomial distribution: *Probit* (Bliss, 1935), *Logit* (Berkson, 1944; Dyke, Patterson, 1952).
- Poisson distribution: count data, *Log* transformation (Birch, 1963).
- Exponential distribution, gamma distribution: survival models, *inverse* and *log* transformations (Feigl, Zelen, 1965; Zipin, Armitage, 1966; Nelder, 1966; Glasser, 1967).
- **Generalized Linear Models**: Nelder, Wedderburn (1972) family of models, estimation of parameters – MLE
- Count data models with zero-truncation and zero-inflation problems (21st century)

Before we begin with math...

The objects used in this course (try to) follow some notation rules so it's hopefully easier to follow the formulas. The rules mainly apply to letters at the end of the alphabet (x,y,z, sometimes w and few others).

- X, Y – italic capital letters usually denote random variables
- x, y – italic small letters usually denote realizations of those random variables (they may have subscripts that specify the details)
- $\boldsymbol{x}, \boldsymbol{y}$ – bold small letters usually denote column vectors of realizations, their elements are denoted by x_i and y_i, respectively
- $\mathbf{X}, \mathbf{Y}$ – bold straight capital letters usually denote matrices, their elements are denoted by x_{ij} and y_{ij}, respectively
- you should still look at the context, there are other quantities that are defined which don't always follow these rules

Linear model

$$\boldsymbol{y} = \mathbf{X}\beta + \varepsilon$$

$\boldsymbol{y} = (y_1, \ldots, y_n)^T$ – dependent variable, response

$\mathbf{X} = (\mathbf{1}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$ – design matrix $n \times p$, $\mathbf{1}$ vector of 1s

$\beta = (\beta_0, \ldots, \beta_k)^T$ – vector of unknown parameters

k – number of explanatory variables (number of unknown parameters $p = k + 1$)

$\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ – vector of random errors

Responses are regarded as random variables. Explanatory variables are usually treated as non-random measurements or observations; for example, they may be fixed by the experimental design.

Responses and explanatory variables are measured on one of the following scales: nominal, ordinal, continuous.

Methods of statistical analysis depend on the measurement scales of the response and explanatory variables.

Model fitting. Estimation

Model fitting is estimation of unknown parameters.

Methods of estimation:

- Least Squares (LS)
- Maximum Likelihood (ML)

Comments

- An important distinction between the methods of ML and LS is that the method of LS can be used without making assumptions about the distributions of the response variables. To obtain ML estimators, we need to specify the probability distribution of the response.
- In many situations, ML and LS estimations give the same estimators.
- Often, numerical methods may be needed to obtain parameter estimates that maximize the likelihood or log-likelihood function or minimize the sum of squares.

Least Squares Estimation (LSE)

Ordinary least squares (OLS) or linear least squares

Method for estimating the unknown parameters in a model, with the goal of minimizing the sum of squares of the differences between the observed responses (values of the variable being predicted) in the given dataset and those predicted by a linear function of a set of explanatory variables.

This method obtains parameter estimates that minimize the sum of squared differences

$$\sum_i (y_i - \hat{y}_i)^2 \Rightarrow min$$

Minimization of this function results in a set of normal equations, a set of simultaneous linear equations in the parameters, which are solved to find the parameter estimates.

In principle, the method can be used without making assumptions about the distributions of the response variables.

Example (LSE). Olympic winning times

To illustrate the method of least squares, consider the following 20 pairs of data points, where x is the time in years since 1900 and y is the Olympic winning time in seconds for men in the final round of the 100-meter event:

x	0	4	8	12	20	24	28	32	36	48
y	10.8	11.0	10.8	10.8	10.8	10.6	10.8	10.3	10.3	10.3
x	52	56	60	64	68	72	76	80	84	88
y	10.4	10.5	10.2	10.0	9.95	10.14	10.06	10.25	9.99	9.92

The data set covers all Olympic events held between 1900 and 1988. (Olympic games were not held in 1916, 1940, and 1944.) For this data,
$\bar{x} = 45.6, \bar{y} = 10.396$, and the least squares estimates for slope and intercept are -0.011 and 10.898, respectively.

Source: Mathematica Laboratories for Mathematical Statistics, Emphasizing Simulation and Computer Intensive Methods by Jenny A. Baglivo. ASA-SIAM Series on Statistics and Applied Probability, Volume 14, 2005

Example. Olympic winning times ...

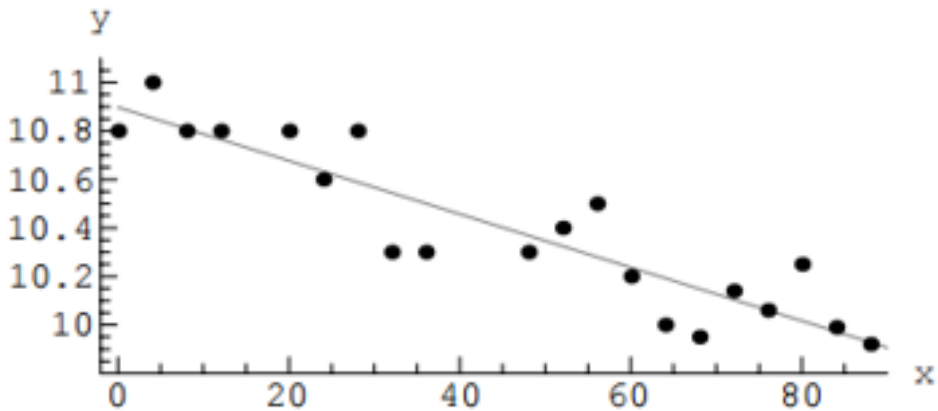


Figure 14.1. *Olympic winning time in seconds for men's 100-meter finals (vertical axis) versus year since* 1900 *(horizontal axis). The gray line is the linear least squares fit,* $y = 10.898 - 0.011x$.

The figure shows a scatter plot of the Olympic winning times data and also includes the least squares fitted line. The results suggest that the winning times have decreased at the rate of about 0.011 seconds per year during the 88 years of the study.

Maximum Likelihood Estimation

Notation

- $f(y;\theta)$ – pdf (continuous variable) or pmf (discrete variable), θ – vector of parameters
- $L(\theta, y)$ – likelihood function (function of parameters θ given data)
- $l(\theta, y) = \ln L(\theta, y)$ – log-likelihood, natural logarithm of the likelihood function

Maximum likelihood method

Method of estimating the parameters of a statistical model (based on given observations) by finding the parameter values that maximize the likelihood of getting these observations.

Because logarithm is a monotonic increasing function, the logarithm of a function achieves its maximum value at the same points as the function itself, and hence the log-likelihood can be used in place of the likelihood in maximum likelihood estimation and related techniques.

Example (MLE). Birthweight and gestational age, 1

The data (Table 2.3) shows the birthweights (in grams) and estimated gestational ages (in weeks) of 12 male and female babies born in a certain hospital. The mean ages are almost the same for both sexes but the mean birthweight for boys is higher than the mean birthweight for girls.

Boys: mean ages 38.33 weeks, mean birthweight 3024 g
Girls: mean ages 38.75 weeks, mean birthweight 2911.33 g

The question of interest is *whether the rate of increase of birth weight with gestational age is the same for boys and girls*.

Source: Dobson, A. J. (2001). An introduction to generalized linear models. Example 2.2.2

Example (MLE). Birthweight and gestational age, 2

Table 2.3 *Birthweight and gestational age for boys and girls.*

	Boys		Girls	
	Age	Birthweight	Age	Birthweight
	40	2968	40	3317
	38	2795	36	2729
	40	3163	40	2935
	35	2925	38	2754
	36	2625	42	3210
	37	2847	39	2817
	41	3292	40	3126
	40	3473	37	2539
	37	2628	36	2412
	38	3176	38	2991
	40	3421	39	2875
	38	2975	40	3231
Means	38.33	3024.00	38.75	2911.33

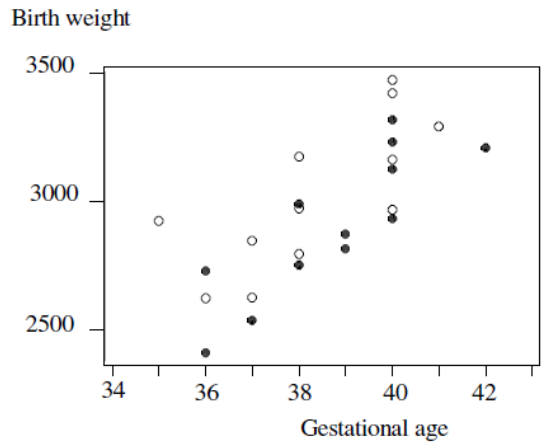


Figure 2.2 *Birthweight plotted against gestational age for boys (open circles) and girls (solid circles); data in Table 2.3.*

There is a linear trend of birth weight increasing with gestational age and the girls tend to weigh less than the boys of the same gestational age.

Example (MLE). Models. Birthweight and ...

Model 1: (the growth rate of birthweight with gestational age is different for boys and girls)

$$y_{ij} = \alpha_j + \beta_j x_{ij} + \varepsilon_{ij}$$

y_{ij} – random variable representing the birthweight of the ith baby in group j where $j = 1$ for boys and $j = 2$ for girls and $i = 1, \ldots, 12$
α_1, α_2 – intercept parameters
β_1, β_2 – parameters (depending on gender)
x_{ij} – explanatory variable , the gestational age of the ith baby, $j = 1, 2$
ε_{ij} – random error, *iid* $\varepsilon_{ij} \sim N(0, \sigma^2)$

Model 0: (the growth rates are equal)

$$y_{ij} = \alpha_j + \beta x_{ij} + \varepsilon_{ij}$$

Hypothesis: Model 1 *vs* Model 0

$$H_0 : \beta_1 = \beta_2 (= \beta); \quad H_1 : \beta_1 \neq \beta_2$$

Example (MLE). Estimation. Birthweight and ...

Model 1: Likelihood

$$L_1 = \prod_j \prod_i \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\{-\frac{1}{2\sigma^2}(y_{ij} - \alpha_j - \beta_j x_{ij})^2\}.$$

Maximum of log-likelihood

$$\widehat{l}_1 = -12 \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\widehat{S}_1$$

$$\widehat{S}_1 = \sum_{j=1}^{2}\sum_{i=1}^{12}(y_{ij} - a_j - b_j x_{ij})^2$$

Model 0: Likelihood

$$L_0 = \prod_j \prod_i \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\{-\frac{1}{2\sigma^2}(y_{ij} - \alpha_j - \beta x_{ij})^2\}.$$

Maximum of log-likelihood

$$\widehat{l}_0 = -12 \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\widehat{S}_0, \quad \widehat{S}_0 = \sum_{j=1}^{2}\sum_{i=1}^{12}(y_{ij} - a_j - b x_{ij})^2$$

Example (MLE). Results. Birthweight and ...

Table 2.5 *Analysis of data on birthweight and gestational age in Table 2.3.*

Model	Slopes	Intercepts	Minimum sum of squares
0	$b = 120.894$	$a_1 = -1610.283$	$\widehat{S}_0 = 658770.8$
		$a_2 = -1773.322$	
1	$b_1 = 111.983$	$a_1 = -1268.672$	$\widehat{S}_1 = 652424.5$
	$b_2 = 130.400$	$a_2 = -2141.667$	

If H_0 is correct, the minimum values $\hat{S}_1$ and $\hat{S}_0$ should be nearly equal.
If H_0 is correct, the improvement in fit is $\frac{1}{\sigma^2}(S_0 - S_1) \sim \chi^2_1$.
However, as σ^2 is unknown, we have to eliminate σ^2 using the ratio of $\frac{\hat{S}_0 - \hat{S}_1}{\sigma^2}$ to $\frac{\hat{S}_1}{\sigma^2}$ each divided by the relevant degrees of freedom, and get F-distribution.
$F = \frac{(\hat{S}_0 - \hat{S}_1)/(J-1)}{\hat{S}_1/(JK-2J)} = \frac{(658770.8 - 652424.5)/1}{652424.5/20} = 0.19 \quad (J = 2, K = 12)$. **Conclusion?**

Motivating example (1)

Challenger Disaster Example

In January 1986, the space shuttle Challenger exploded shortly after launch. An investigation was launched into the cause of the crash and attention focused on the rubber O-ring seals in the rocket boosters. At lower temperatures, rubber becomes more brittle and is a less effective sealant. At the time of the launch, the temperature was 31°F.

Could the failure of the O-rings have been predicted?

In the 23 previous shuttle missions for which data exists, some evidence of damage due to blow by and erosion was recorded on some O-rings. Each shuttle had two boosters, each with three O-rings. For each mission, we know the number of O-rings out of six showing some damage and the launch temperature.

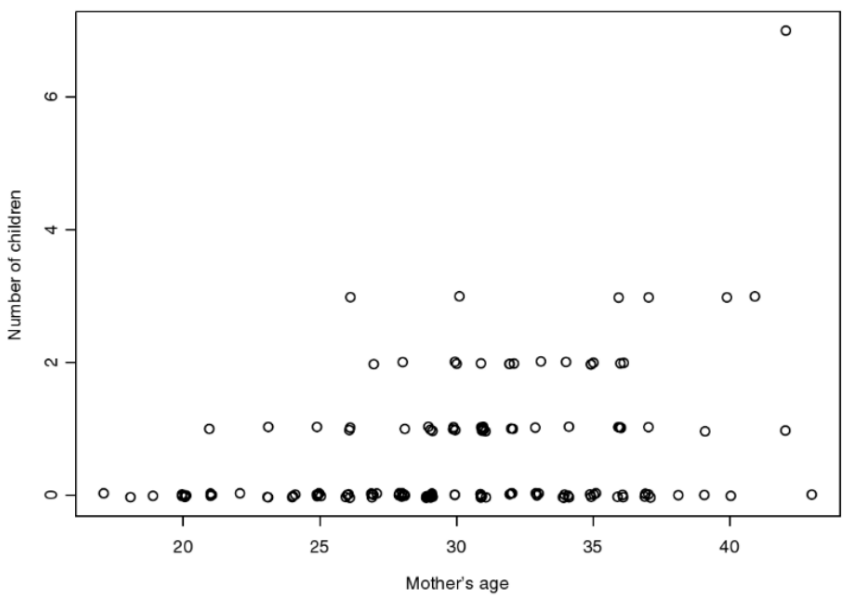
The standard linear model is clearly not directly suitable here, it is better to develop a model that is directly suited for binomial data.

Source: J.J.Faraway (2006). Extending the Linear Model with R. Generalized Linear, Mixed Effects and Nonparametric Regression Models.

Motivating example (2)

Number of children

This data set contains the number of children for each of 141 pregnant women. The age of each mother or mother-to-be is also recorded (Leader 1994). Figure plots the number of children versus mother's age. Since both variables are integers, the points fall on a grid.
As the mother's age increases there is a tendency for more children. However it is not clear whether the relationship is linear or curvilinear.



Source: P. de Jong and G. Z. Heller (2008). Generalized Linear Models for Insurance Data.

Motivating example (3)

Third party claims

Third party insurance is a compulsory insurance for vehicle owners in Australia. It insures vehicle owners against injury caused to other drivers, passengers or pedestrians, as a result of an accident.
This data set records the number of third party claims in a twelve-month period between 1984 and 1986 in each of 176 geographical areas (local government areas) in New South Wales, Australia. Areas are grouped into 13 statistical divisions. Other recorded variables are the number of accidents, the number of people killed or injured and population.

Source: P. de Jong and G. Z. Heller (2008). Generalized Linear Models for Insurance Data.

Response distribution

In classical linear models the response variable is often assumed normally distributed, but this is not the only type of responses we meet in practice.

Some examples:

- Continuous response variable, symmetrical or non-symmetrical
- Binary response variable
- Count or rate as a response variable

In fact, the normal model is rarely adequate.

We have to consider

- Discrete random variables (Bernoulli, Binomial, Poisson, Negative Binomial)
- Continuous random variables (Normal, Inverse-Normal, Log-Normal, Gamma)

Distributions related to Normal distribution

1. χ^2-distribution

Sum of squares of n independent standard normal random variables follows χ^2-distribution: if $X_i \sim N(0,1)$ and $Y = \sum_{i=1}^{n} {X_i}^2$ then $Y \sim \chi^2(n)$.
$EY = n$, $DY = 2n$, n – degrees of freedom

2. Student's t-distribution

If $X \sim N(0,1)$ and $Y_n \sim \chi^2(n)$ are independent then r.v. $T = \sqrt{n}\dfrac{X}{\sqrt{Y_n}}$ follows t-distribution (with degrees of freedom n), $T \sim t_n$

3. F-distribution

Ratio of two independent χ^2-distributed r.v.-s follows Fisher's F-distribution:

$$\frac{Y_m/m}{Y_n/n} \sim F_{m,n}$$

where $Y_m \sim \chi^2(m)$ and $Y_n \sim \chi^2(n)$