UNIVERSITY OF MANNHEIM
School of Social Sciences

Quantitative Methods in Political Science
The Likelihood Theory of Statistical Inference

Thomas Gschwend | Oliver Rittmann | Viktoriia Semenova | David M. Grundmanns

Week 10 - 10 November 2021

The Course: Overview

Roadmap

- Understand and model stochastic processes
- Understand statistical inference
- Implement it mathematically and learn how to estimate it
  - OLS
  - Maximum Likelihood
- Implement it using software
  - R
  - Basic programming skills

Beyond the normal distribution

- So far, we dealt with continuous dependent variables.
- Many observed variables in political science are fundamentally discrete:
  - vote choice
  - voter preferences
  - international event counts
  - political party identification
- To estimate those models, we need a technology different from ordinary least squares.

In this lecture ...

- we introduce a first, simple, example of a non-continuous model and
- introduce a general estimation technique, maximum likelihood, which we will use throughout the rest of the course.

Motivating example: pass or fail an exam

- Imagine an exam which you can either pass or fail.
- Let Y be the outcome with two possible states:
  - pass, $y = 1$, with probability p_1
  - fail: $y = 0$, with probability p_0
  - obviously, $p_0 + p_1 = 1$
- In practical applications, p is unknown. Thus, we wanna estimate it.
  - take a sample of size N to estimate it
  - assume a probability distribution for p
- Say we want to estimate p_1, the probability of passing the exam, via a sample of $N = 10$.
- We observe: `y=(1,1,0,1,0,1,0,1,0,1)`, so 6 students passed the exam.
- How likely is that? Distribution?

Binomial distribution of exam results

- The variable Y ("y out of N students passed") follows a binomial distribution.
- Its probability function has just two parameters:
  - the probability of passing the exam, p
  - the number of observations of the sample, N.

$$
\begin{aligned}
P(Y = y|N, p) &= \binom{N}{y} p^y (1-p)^{N-y} \\
&= \frac{N!}{(N-y)!y!} p^y (1-p)^{N-y}
\end{aligned}
$$

- To estimate p 'by hand' we let it vary from 0 to 1 and calculate the corresponding value of $P(Y = y|N = 10, p)$

Resulting table of probabilities $P(Y = y|N = 10, p)$

	y										
p	0	1	2	3	4	5	6	7	8	9	10
0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.1	0.35	0.39	0.19	0.06	0.01	0.00	0.00	0.00	0.00	0.00	0.00
0.2	0.11	0.27	0.30	0.20	0.09	0.03	0.01	0.00	0.00	0.00	0.00
0.3	0.03	0.12	0.23	0.27	0.20	0.10	0.04	0.01	0.00	0.00	0.00
0.4	0.01	0.04	0.12	0.21	0.25	0.20	0.11	0.04	0.01	0.00	0.00
0.5	0.00	0.01	0.04	0.12	0.21	0.25	0.21	0.12	0.04	0.01	0.00
0.6	0.00	0.00	0.01	0.04	0.11	0.20	0.25	0.21	0.12	0.04	0.01
0.7	0.00	0.00	0.00	0.01	0.04	0.10	0.20	0.27	0.23	0.12	0.03
0.8	0.00	0.00	0.00	0.00	0.01	0.03	0.09	0.20	0.30	0.27	0.11
0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.06	0.19	0.39	0.35
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

If the true p is 0.5, how many (y) most likely passed?

	y										
p	0	1	2	3	4	5	6	7	8	9	10
0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.1	0.35	0.39	0.19	0.06	0.01	0.00	0.00	0.00	0.00	0.00	0.00
0.2	0.11	0.27	0.30	0.20	0.09	0.03	0.01	0.00	0.00	0.00	0.00
0.3	0.03	0.12	0.23	0.27	0.20	0.10	0.04	0.01	0.00	0.00	0.00
0.4	0.01	0.04	0.12	0.21	0.25	0.20	0.11	0.04	0.01	0.00	0.00
0.5	0.00	0.01	0.04	0.12	0.21	0.25	0.21	0.12	0.04	0.01	0.00
0.6	0.00	0.00	0.01	0.04	0.11	0.20	0.25	0.21	0.12	0.04	0.01
0.7	0.00	0.00	0.00	0.01	0.04	0.10	0.20	0.27	0.23	0.12	0.03
0.8	0.00	0.00	0.00	0.00	0.01	0.03	0.09	0.20	0.30	0.27	0.11
0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.06	0.19	0.39	0.35
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

If the true p is 0.5, how many (y) most likely passed?

						y					
p	0	1	2	3	4	5	6	7	8	9	10
0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.1	0.35	0.39	0.19	0.06	0.01	0.00	0.00	0.00	0.00	0.00	0.00
0.2	0.11	0.27	0.30	0.20	0.09	0.03	0.01	0.00	0.00	0.00	0.00
0.3	0.03	0.12	0.23	0.27	0.20	0.10	0.04	0.01	0.00	0.00	0.00
0.4	0.01	0.04	0.12	0.21	0.25	0.20	0.11	0.04	0.01	0.00	0.00
0.5	0.00	0.01	0.04	0.12	0.21	0.25	0.21	0.12	0.04	0.01	0.00
0.6	0.00	0.00	0.01	0.04	0.11	0.20	0.25	0.21	0.12	0.04	0.01
0.7	0.00	0.00	0.00	0.01	0.04	0.10	0.20	0.27	0.23	0.12	0.03
0.8	0.00	0.00	0.00	0.00	0.01	0.03	0.09	0.20	0.30	0.27	0.11
0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.06	0.19	0.39	0.35
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

If the true p is 0.5, how many (y) most likely passed?

						y					
p	0	1	2	3	4	5	6	7	8	9	10
0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.1	0.35	0.39	0.19	0.06	0.01	0.00	0.00	0.00	0.00	0.00	0.00
0.2	0.11	0.27	0.30	0.20	0.09	0.03	0.01	0.00	0.00	0.00	0.00
0.3	0.03	0.12	0.23	0.27	0.20	0.10	0.04	0.01	0.00	0.00	0.00
0.4	0.01	0.04	0.12	0.21	0.25	0.20	0.11	0.04	0.01	0.00	0.00
0.5	0.00	0.01	0.04	0.12	0.21	0.25	0.21	0.12	0.04	0.01	0.00
0.6	0.00	0.00	0.01	0.04	0.11	0.20	0.25	0.21	0.12	0.04	0.01
0.7	0.00	0.00	0.00	0.01	0.04	0.10	0.20	0.27	0.23	0.12	0.03
0.8	0.00	0.00	0.00	0.00	0.01	0.03	0.09	0.20	0.30	0.27	0.11
0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.06	0.19	0.39	0.35
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

… but do we really care about that?

						y					
p	0	1	2	3	4	5	6	7	8	9	10
0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.1	0.35	0.39	0.19	0.06	0.01	0.00	0.00	0.00	0.00	0.00	0.00
0.2	0.11	0.27	0.30	0.20	0.09	0.03	0.01	0.00	0.00	0.00	0.00
0.3	0.03	0.12	0.23	0.27	0.20	0.10	0.04	0.01	0.00	0.00	0.00
0.4	0.01	0.04	0.12	0.21	0.25	0.20	0.11	0.04	0.01	0.00	0.00
0.5	0.00	0.01	0.04	0.12	0.21	0.25	0.21	0.12	0.04	0.01	0.00
0.6	0.00	0.00	0.01	0.04	0.11	0.20	0.25	0.21	0.12	0.04	0.01
0.7	0.00	0.00	0.00	0.01	0.04	0.10	0.20	0.27	0.23	0.12	0.03
0.8	0.00	0.00	0.00	0.00	0.01	0.03	0.09	0.20	0.30	0.27	0.11
0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.06	0.19	0.39	0.35
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

We observed $y = 6$. Which p most likely generated that?

						y					
p	0	1	2	3	4	5	6	7	8	9	10
0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.1	0.35	0.39	0.19	0.06	0.01	0.00	0.00	0.00	0.00	0.00	0.00
0.2	0.11	0.27	0.30	0.20	0.09	0.03	0.01	0.00	0.00	0.00	0.00
0.3	0.03	0.12	0.23	0.27	0.20	0.10	0.04	0.01	0.00	0.00	0.00
0.4	0.01	0.04	0.12	0.21	0.25	0.20	0.11	0.04	0.01	0.00	0.00
0.5	0.00	0.01	0.04	0.12	0.21	0.25	0.21	0.12	0.04	0.01	0.00
0.6	0.00	0.00	0.01	0.04	0.11	0.20	0.25	0.21	0.12	0.04	0.01
0.7	0.00	0.00	0.00	0.01	0.04	0.10	0.20	0.27	0.23	0.12	0.03
0.8	0.00	0.00	0.00	0.00	0.01	0.03	0.09	0.20	0.30	0.27	0.11
0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.06	0.19	0.39	0.35
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

We observed $y = 6$. Which p most likely generated that?

						y					
p	0	1	2	3	4	5	6	7	8	9	10
0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.1	0.35	0.39	0.19	0.06	0.01	0.00	0.00	0.00	0.00	0.00	0.00
0.2	0.11	0.27	0.30	0.20	0.09	0.03	0.01	0.00	0.00	0.00	0.00
0.3	0.03	0.12	0.23	0.27	0.20	0.10	0.04	0.01	0.00	0.00	0.00
0.4	0.01	0.04	0.12	0.21	0.25	0.20	0.11	0.04	0.01	0.00	0.00
0.5	0.00	0.01	0.04	0.12	0.21	0.25	0.21	0.12	0.04	0.01	0.00
0.6	0.00	0.00	0.01	0.04	0.11	0.20	0.25	0.21	0.12	0.04	0.01
0.7	0.00	0.00	0.00	0.01	0.04	0.10	0.20	0.27	0.23	0.12	0.03
0.8	0.00	0.00	0.00	0.00	0.01	0.03	0.09	0.20	0.30	0.27	0.11
0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.06	0.19	0.39	0.35
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Maximum Likelihood Principle

The value of p for which the observed data are most likely (i.e. have highest probability of being observed) is called the *maximum likelihood estimate* ($\hat{p}_{ML}$).

- How do we get there?
- The solution turns out to be the likelihood, $L(p|y, N)$, defined as function consisting of values *proportional* to the traditional probability (density) distribution for different (hypothetical) values of p
- We can describe this likelihood as a function:

$$L(p|y, N) = \frac{N!}{(N-y)!y!} p^y (1-p)^{N-y},$$

- L represents the likelihood that the observed data ($N = 10$ with $y = 6$ passed exams) is generated by the assumed DGP (here: binomial) for every p.
- The maximum of this likelihood function L is our Maximum Likelihood Estimate (MLE).

Log-Likelihood

- In practice, the log of the likelihood is usually used – taking the logarithm changes products to sums, making calculations easier. Therefore

$$L(p|y,N) = \frac{N!}{(N-y)!y!}p^y(1-p)^{N-y},$$

becomes

$$\begin{aligned} logL(p|y,N) &= log\left[\frac{N!}{(N-y)!y!}\right] + y\,logp + (N-y)\,log(1-p) \\ &= y\,logp + (N-y)\,log(1-p) \end{aligned}$$

- Thus for $y = 6$ and $N = 10$ we get:

$$log\,L(p|y=6,N=10) = 6 \cdot log(p) + 4 \cdot log(1-p)$$

Log-Likelihood Values of $log\, L(p|y = 6, N = 10)$

This yields the following table:

	y										
p	0	1	2	3	4	5	6	7	8	9	10
0	–	–	–	–	–	–	–	–	–	–	–
0.1	-1.05	-0.95	-1.64	-2.86	-4.50	-6.51	-8.89	-11.65	-14.82	-18.53	-23.03
0.2	-2.23	-1.32	-1.20	-1.60	-2.43	-3.63	-5.20	-7.15	-9.52	-12.41	-16.09
0.3	-3.57	-2.11	-1.45	-1.32	-1.61	-2.27	-3.30	-4.71	-6.54	-8.89	-12.04
0.4	-5.11	-3.21	-2.11	-1.54	-1.38	-1.61	-2.19	-3.16	-4.55	-6.45	-9.16
0.5	-6.93	-4.63	-3.12	-2.14	-1.58	-1.40	-1.58	-2.14	-3.12	-4.63	-6.93
0.6	-9.16	-6.45	-4.55	-3.16	-2.19	-1.61	-1.38	-1.54	-2.11	-3.21	-5.11
0.7	-12.04	-8.89	-6.54	-4.71	-3.30	-2.27	-1.61	-1.32	-1.45	-2.11	-3.57
0.8	-16.09	-12.41	-9.52	-7.15	-5.20	-3.63	-2.43	-1.60	-1.20	-1.32	-2.23
0.9	-23.03	-18.53	-14.82	-11.65	-8.89	-6.51	-4.50	-2.86	-1.64	-0.95	-1.05
1	–	–	–	–	–	–	–	–	–	–	–

which, again, leads to $\hat{p}_{ML} = .6$

Likelihood and (Log-)Likelihood

- Both, the likelihood and log likelihood function gave the same estimate for p

$$\hat{p}_{ML} = max(L(p|y, N)) = max(logL(p|y, N)) = 0.6$$

- More general, for any set of parameters θ

$$\hat{\theta}_{ML} = max(L(\theta)) = max(logL(\theta)),$$

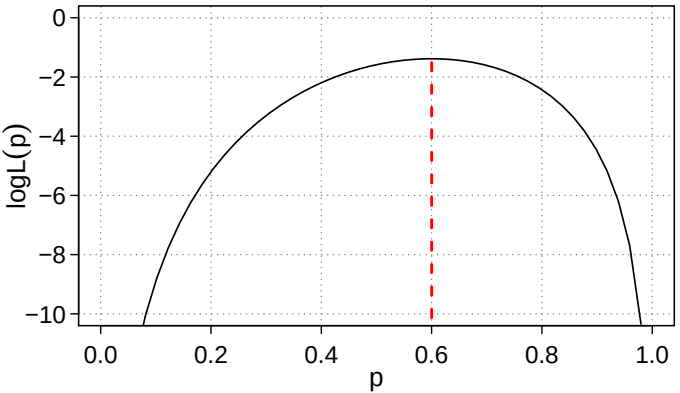
  i.e. both functions are maximum likelihood estimators.

- While it is easy to find the ML for simple problems by plugging in values and looking over the resulting table, when using more complex models we search for the maximum directly.
- We have to find the maximum of a function, which we know how to do. Modern statistics packages and computer algebra programs contain search routines for finding maxima.

Maximizing the (Log-)Likelihood

- Example: plot binomial $log\, L(p|y = 6, N = 10) = 6 \cdot log(p) + 4 \cdot log(1 - p)$
- We find the maximum via

$$\frac{dlogL(p)}{dp} = 0$$



The Log-Likelihood Function

- We can do this analytically for tractable likelihood functions. Here we have $log\, L(p|y = 6, N = 10) = 6 \cdot log(p) + 4 \cdot log(1 - p)$
- Maximizing this log-likelihood function gives:

$$\frac{\partial\, L(p \mid N = 10, y = 6)}{\partial\, p} = \frac{6}{p} - \frac{4}{1-p}$$

- The corresponding first-order condition is given as:

$$\begin{aligned} \frac{6}{p} - \frac{4}{1-p} &= 0 \\ 6(1-p) &= 4p \\ \hat{p}_{ML} &= 0.6 \end{aligned}$$

- This corresponds to the value that we read off the table.
- Choosing $p = 0.6$ makes it most likely to observe 6 'pass and 4 'fail' from 10 repeated Bernoulli trials. It maximizes the likelihood.

Maximum Likelihood Estimation (MLE)

- Maximum likelihood estimation searches the parameter space and chooses model parameters θ of the assumed DGP that most likely lead to observe the particular data y we have.
- It is a powerful tool as it is highly flexible and can be used to estimate, e.g., non-linear models.
- It is used to estimate logit and probit models, choice models (ordered probit, multinomial logit), or count models (Poisson regression, (zero-inflated) negative binomial models). Which we will discuss in the following weeks.
- MLE can also be used to estimate the linear model and retrieves the same coefficients as OLS does.

Maximum likelihood estimation of a linear regression

- Since maximum likelihood is a general technique, we can easily estimate a standard linear regression model via maximum likelihood instead of using ordinary least squares.
- Let's go back to one of our earlier examples: the relationship between the two party vote share and economic growth in US presidential elections.
- Remember the basic regression equation:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon$$

- The parameters we are going to estimate are $\theta = (\beta_0, \beta_1, \sigma^2)$
- First, we specify a suitable probability model of the data generating process:

$Y_i \sim N(\mu_i, \sigma^2)$ Stochastic component
$\mu_i = \beta_0 + \beta_1 x_i$ Systematic component

Specify the probability model

- We assumed that Y_i is distributed normal, $Y_i \sim N(y_i|\mu_i, \sigma^2)$, hence for ith observation y_i we get

$$Pr(Y_i = y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(Y_i - \mu_i)^2}{2\sigma^2}\right)$$

- Since we assume a simple random sample, we have N realizations (of identically and independently distributed random variables). Thus, the probability model is:

$$\begin{aligned} Pr(Y_1, \cdots, Y_N) &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(Y_i - \mu_i)^2}{2\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-N/2}\ exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \mu_i)^2\right] \\ &= (2\pi\sigma^2)^{-N/2}\ exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i)^2\right] \end{aligned}$$

Set-up the Likelihood Function

- The likelihood that the observed data is generated by the assumed DGP is

$$L(\beta_0, \beta_1, \sigma^2) = (2\pi\sigma^2)^{-N/2} \, exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i)^2\right]$$

- This likelihood function describes the *data generating process* completely – i.e. it contains all the unknown parameters β_0, β_1 and the variance term σ^2.
- Finally (although this is optional) we take the log of the likelihood function, since it is easier to maximize:

$$\begin{aligned} logL(\beta_0, \beta_1, \sigma^2) &= -\frac{N}{2} log(2\pi\sigma^2) \; - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i)^2 \\ &= -\frac{N}{2} log(\sigma^2) \; - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i)^2 \end{aligned}$$

Obtaining standard errors

- To get standard errors of our model, we use the negative second partial derivatives of the log-likelihood function with respect to the ML estimates (the "Hessian" matrix)
- Taking the expectation yields the ("Fisher") information matrix which indicates the precision of your parameter estimates: $I(\hat{\theta}_{ML})$.
- The inverse of the information matrix is an estimate of the variance-covariance matrix of the ML parameter estimates
- Finally, you get standard errors by taking the square root of the variances (which are on the diagonal).

$$se(\hat{\theta}_{ML}) = \sqrt{\widehat{Var}(\hat{\theta}_{ML})} = 1/\sqrt{I(\hat{\theta}_{ML})}$$

- Remember that second derivatives tell the rate of the curvature of the function. Therefore: steep LL curve $\rightarrow$ high rate of curvature (quite sure that we have reached "true" maximum) $\rightarrow$ large second derivative $\rightarrow$ inversion makes SE small

Summary: Maximum Likelihood Estimation

Easy steps to obtain MLE

1. Formulate a suitable probability model of the data-generating process including assumptions of how Y (and therefore also how ϵ) is distributed (i.e., stochastic component) and a parameterization of stuff that gets estimated (i.e., systematic component).
2. Write down the likelihood function based on your parameterization
3. Maximize it! Take the first partial derivative of the log-likelihood function with respect to the parameters (also called *score function*) and set it equal to 0.
4. Solve to find the parameter estimates
5. Take the negative second partial derivatives of the log-likelihood function at the maximum likelihood estimates to get the Fisher information. Take the square root of the inverse of this quantity to find the standard errors of your parameters.
6. .... or let R do the job in step 3 and 5 for you!