Chapter 3: Maximum Likelihood Estimation

MATH 60604A: Statistical Modelling

HEC Montréal
Department of decision sciences

Overview of course material

Type of variable Y	Independent Observations	Method
Continuous	Yes	**Simple linear regression (chap 2 part 1)** **Multiple linear regression (chap 2 part 2)** **Special cases: t-test and ANOVA (chap 2 part 3)** **Models for survival data (chap 6)**
Continuous	No (ex : longitudinal study)	**Regression with random effects (chap 5)**
Binary	Yes	**Logistic Regression (chap 4)**
Count	Yes	**Poisson Regression (chap 4)**

Chapter overview

Introduction

- In general, when we're interested in fitting a particular model (whether it be a simple model or a very complex model), we must estimate the model parameters. It is precisely these parameters that allow to specify the actual form of the underlying model.
  - Ex: If we're interested in fitting a linear regression model, we must estimate the regression coefficients β_j (that quantify the effects of the explanatory variables X_j and allow to specify the mean of Y) and we must estimate σ^2 (which will allow to do inference, e.g. calculate standard errors, confidence intervals, p-values)
- There exists many different estimation techniques, ex:
  - least squares estimation
  - method of moments
  - maximum likelihood estimation
  - Bayes estimator
  - etc.

Introduction to ML estimation

- The focus of this chapter is on maximum likelihood estimation, which is one of the most common estimation techniques in statistics.
- General setup: suppose we have observed data $(x_1, x_2, \ldots, x_n)$ which we believe come from some parametric model $F_\theta(x)$
  - that is, we assume that the observed data are realizations of $X \sim F_\theta(x)$, i.e. X follows a distribution F_θ which is parametrized in terms of θ
  - ex: we have observations $(x_1, \ldots, x_n)$ which we assume come from a Normal distribution $\mathcal{N}(\mu, \sigma^2)$
- Our goal is to estimate the parameter θ
  - note that θ could be one-dimensional, i.e. there is only one parameter to estimate (ex: μ), or θ could consist of a multi-dimensional vector of parameters (ex: μ, σ^2)
- Maximum likelihood estimation allows to estimate the parameters for any type of model by maximizing a specific criterion: the likelihood

Introduction to ML estimation

- The likelihood function is a function of the unknown parameter (or parameter vector) θ, which we wish to estimate, and the observed data $(x_1, x_2, \ldots, x_n)$.
  - Formally, the likelihood function is the joint probability of the observed data $(x_1, \ldots, x_n)$, for an arbitrary value of θ – it's the probability of observing what we observe!
- For a random sample, i.e. when $X_1, \ldots, X_n$ are i.i.d., the likelihood function can be written as follows:
  - for a continuous distribution with density $f_\theta(x)$

$$\boxed{L(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i)}$$

  - for a discrete distribution with probability mass function $P_\theta(x)$

$$\boxed{L(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} P_\theta(X = x_i)}$$

Introduction to ML estimation

- Thus, the likelihood function (which we can denote by $L(\theta)$ for simplicity) represents the probability of observing the sample $(x_1, \ldots, x_n)$ for a given value of θ
  - Note that $L(\theta)$ is considered a function of θ, and the observed values $(x_1, \ldots, x_n)$ are considered fixed (known values)
- The idea of maximum likelihood (ML) estimation is to estimate θ by the value which maximizes the likelihood $L(\theta)$
  - in other words, the maximum likelihood estimator (MLE) $\hat{\theta}$ is such that the probability of observing the given sample is as large as possible, i.e. $\hat{\theta}$ is the value of θ that makes the observed sample the most likely possible
- We can write the MLE as

$$\boxed{\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, L(\theta | x_1, \ldots, x_n)}$$

where Θ denotes the parameter space, that is, the set of possible values that θ can take on

Ex: introduction to ML estimation

- Suppose we want to estimate the probability that a particular event occurs, without incorporating any explanatory variables in the model.
- The event of interest either happens, or not - it's a binary outcome.
  - Ex: flipping a coin and seeing if it comes up heads or tails
- An appropriate distribution for modelling such a random event is the Bernoulli distribution; recall:
  - A random variable X follows a Bernoulli distribution with parameter p $(0 \leq p \leq 1)$ if $X \in \{0, 1\}$, i.e. X can only take on the values 0 or 1, and the probability that $X = 1$ is given by p:

$$P(X = 1) = p$$

and thus

$$P(X = 0) = 1 - p$$

Ex: introduction to ML estimation

- Typically, "1" is used to denote the occurrence of the event of interest, that is, we use 1 to denote a "success" and 0 to denote a "failure"
  - Ex: does a client buy a certain product (1) or not (0), does a study participant succeed at carrying out a specific task (1=yes, 0=no), does the coin turn up heads (1) or tails (0), etc...
- We're interested in estimating the probability of obtaining the outcome "1" (i.e. the event that $X = 1$), that is, we're interested in estimating the parameter p of the underlying Bernoulli model.

Ex: introduction to ML estimation

- Suppose that we have a random sample size of n with $X_1, X_2, \ldots, X_n$ assumed to come from a Bernoulli distribution with parameter p.
- A compact way of writing the model for observation i is:

$$P(X_i = x_i | p) = p^{x_i}(1-p)^{(1-x_i)},$$

  for $x_i = 0, 1$

- Since the observations are i.i.d., the joint probability of the observed sample is simply the product of the probabilities for each observation:

$$P(X_1 = x_1, \ldots, X_n = x_n | p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{(1-x_i)}$$

  this represents the probability of observing the sample $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$

Ex: introduction to ML estimation

- In this particular example (Bernoulli random variables), the likelihood function is then given by:

$$L(p|x_1,\ldots,x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{(1-x_i)}$$

- Note that we can rewrite this as

$$L(p) = p^{\sum_{i=1}^{n} x_i}(1-p)^{\sum_{i=1}^{n}(1-x_i)}$$

- Recall: the basic idea in ML estimation is to consider the observed values as fixed and to view $L(p)$ as a function of the parameters (here there's only one parameter: p).
- For a given value of p, $L(p)$ is the probability of observing this sample.

Ex: introduction to ML estimation

- The ML estimator for p is defined as the value of p that maximizes the likelihood function $L(p)$:

$$\hat{p} = \underset{p \in (0,1)}{\operatorname{argmax}} \ L(p)$$

- In other words, the MLE $\hat{p}$ is the value of p, within the interval $(0, 1)$, such that the probability of observing the given sample is as large as possible.

Ex: introduction to ML estimation

- Suppose that for this example we have $n = 10$ observations:

  $$x_1 = 1, x_2 = 0, x_3 = 1, x_4 = 1, x_5 = 1, x_6 = 1, x_8 = 1, x_9 = 0, x_{10} = 1$$

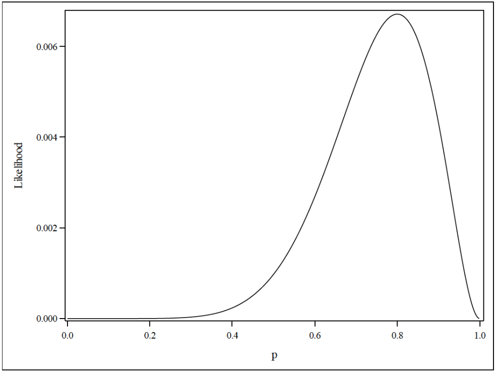
  $\rightarrow$ there are 8 “1” and 2 “0”.

- The likelihood function is therefore:

$$L(p) = p^8(1-p)^2$$

Ex: introduction to ML estimation

- Here is a plot of the likelihood function $L(p)$ as a function of p :



we see that the likelihood is maximized at the value $p = 0.8$. For this sample, the ML estimate for p is then $\hat{p} = 0.8$.

  - This seems reasonable, since p is the (theoretical) probability of having a "1" and 0.8 is the observed proportion of 1's in our sample.

Ex: introduction to ML estimation

- In optimization problems, it is usually easier to work with a sum rather than a product.
- Since the log of a product is equal to the sum of logs, that is,

$$ln(ab) = ln(a) + ln(b)$$

  we can also work with the log of the likelihood.
- We call this function the log-likelihood (LL)
- Since the logarithm is a strictly increasing function, maximizing the log of the likelihood is equivalent to maximizing the likelihood.

Ex: introduction to ML estimation

- In our example, the LL function is:

$$LL(p) = \ln\left\{\prod_{i=1}^{n} p^{x_i}(1-p)^{1-xi}\right\} = \sum_{i=1}^{n} ln\left\{p^{x_i}(1-p)^{1-xi}\right\}$$

- By using the property $\ln(a^b) = b\ln(a)$, this expression can be simplified as:

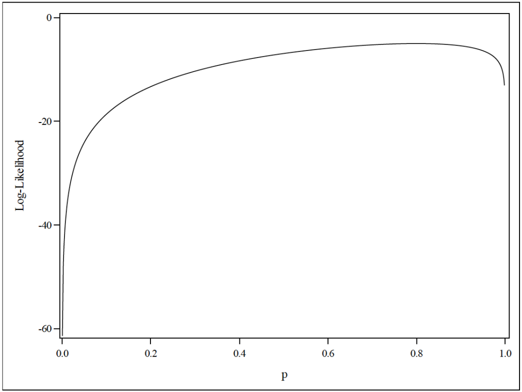
$$LL(p) = \ln(p)\sum_{i=1}^{n} x_i + \ln(1-p)\left\{n - \sum_{i=1}^{n} x_i\right\}$$

- In our numerical example, with eight 1's and two 0's, this function is then:

$$LL(p) = 8\ln(p) + 2\ln(1-p)$$

Ex: introduction to ML estimation

- Here is the plot of the log-likelihood, $LL(p)$, as a function of p:



- It's a little less clear this time, but the maximum is still achieved at the value $p = 0.8$ for this sample.

Ex: introduction to ML estimation

- Generally, the maximum of the likelihood or log-likelihood must be found numerically, using optimization algorithms
- Luckily, we don't need to worry about this, since the the algorithms used in most software are reliable and efficient, at least for the kinds of models seen in this course.
- But for more complex models, like generalized linear mixed models, the convergence of optimization algorithms can be more problematic.

Ex: introduction to ML estimation

- In certain simple cases, it is actually possible to derive an analytic formula for the ML estimator.
  - This is the case in our example with the Bernoulli distribution.
- Finding the maximum of a simple function can be done using a very basic method: finding the point where the derivative of the function equals 0.
- In our example, we see that differentiating $LL(p)$ with respect to p gives

$$\frac{\partial}{\partial p} LL(p) = \frac{1}{p} \sum_{i=1}^{n} x_i - \frac{1}{(1-p)} \left(n - \sum_{i=1}^{n} x_i \right)$$

- If we solve the equation $\frac{\partial}{\partial p} LL(p) = 0$, we get:

$$p = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Ex: introduction to ML estimation

- So the ML estimator of p is:

$$\hat{p}_{mle} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

which is equal to the proportion of 1's in the sample $\rightarrow$ the MLE $\hat{p}_{mle}$ is simply the sample mean $\bar{X}$!

  - In our numerical example, we found that $\hat{p}_{mle} = 0.8$ since 80% of the values were equal to 1.

Chapter overview

General formulation of ML estimation and its properties

- In the preceding example, there was only one parameter, but most of the time the model contains a large number of parameters.
  - ex: in a linear regression model, we wish to estimate the regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ as well as σ^2
- Let θ denote the parameter vector, that is, the vector containing all the model parameters
  - ex: $\theta = p$ in the Bernoulli example
  - ex: $\theta = (\beta_0, \beta_1, \ldots, \beta_p, \sigma^2)$ in a linear regression model
- The idea of ML estimation is always the same:
  - define the likelihood $L(\theta)$ as the joint probability of the observed sample
  - we treat the observed values $(x_1, \ldots, x_n)$ as fixed and view $L(\theta)$ as a function of the parameter vector θ
  - the MLE $\hat{\theta}$ is the value of θ that maximizes the likelihood function $L(\theta)$, or equivalently, maximizes the log-likelihood $LL(\theta)$.

General formulation of ML estimation and its properties

- Suppose we have a random sample of size n, $X_1, \ldots, X_n \overset{iid}{\sim} f_\theta$ and we observe the values $x_1, x_2, \ldots, x_n$.
- Since $X_1, \ldots, X_n$ are independent and have the same f_θ, we can express the likelihood function as

$$L(\theta) = \prod_{i=1}^{n} f_\theta(x_i)$$

and the log-likelihood function by

$$LL(\theta) = \sum_{i=1}^{n} \ln f_\theta(x_i)$$

- The ML estimator is given by

$$\hat{\theta}_{mle} = \underset{\theta \in \Theta}{\arg\max}\, L(\theta) = \underset{\theta \in \Theta}{\arg\max}\, LL(\theta)$$

- (Under regularity conditions) this amounts to solving

$$\frac{\partial}{\partial \theta} LL(\theta) = \mathbf{0}$$

Example: Poisson distribution

- Suppose we're interested in modelling the number of times an event of interest occurs in a given time period.
  - Ex: number of car accident claims in a year, number of deaths at a particular hospital in a month, etc.
- The Poisson distribution can be used to model this, recall:
  - A random variable X follows a Poisson distribution with parameter λ ($\lambda > 0$) if $X \in \{0, 1, 2, \ldots\}$, i.e. X can only take on non-negative integer values, and the probability that $X = x$ is given by:

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

    for any $x \in \{0, 1, 2, \ldots\}$

Example: Poisson distribution

- Suppose we have a random sample of size n, $X_1, \dots, X_n$ where each X_i is independent and follows a Poisson distribution with parameter λ, and that we observe the values $X_1 = x_1, \dots, X_n = x_n$

- The likelihood is

$$L(\lambda | x_1, \dots, x_n) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{x_i}}{x_i!}$$

- The log-likelihood is then

$$LL(\lambda | x_1, \dots, x_n) = -n\lambda + \ln(\lambda) \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \ln(x_i!)$$

- The ML estimator can then be found by solving

$$\frac{\partial}{\partial \lambda} LL(\lambda) = -n + \frac{1}{\lambda} \sum_{i=1}^{n} X_i = 0$$

and thus

$$\hat{\lambda}_{mle} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Example: gamma distribution

- Suppose we're interested in modeling a non-negative, continuous random variable
  - Ex: the cost of an incurred insurance claim, the time in between bus arrivals, etc.
- A gamma distribution can be used to model this.
  - A random variable X follows a gamma distribution with parameters (α, β), with both α (shape parameter) and β (scale parameter) $\in (0, \infty)$, and density

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta}$$

where $\Gamma(\cdot)$ is the gamma function, defined as

$$\Gamma(z) = \int_0^{\infty} x^{z-1} e^{-x} dx$$

note that when z is a positive integer, $\Gamma(z) = (z-1)!$.

Example: gamma distribution

- Suppose we have a random sample $X_1, \ldots, X_n$ (i.i.d.) where each X_i follows a gamma distribution with the same parameters α and β.
- The likelihood function is then

$$L(\alpha, \beta | X_1, \ldots, X_n) = \{\Gamma(\alpha)\}^{-n} \beta^{-\alpha n} \times \left(\prod_{i=1}^{n} X_i^{\alpha-1} \right) \times \exp\left(-\frac{1}{\beta} \sum_{i=1}^{n} X_i \right)$$

and the corresponding log-likelihood is

$$\ell(\alpha, \beta | X_1, \ldots, X_n) = -n \ln\{\Gamma(\alpha)\} - \alpha n \ln(\beta) + (\alpha - 1) \sum_{i=1}^{n} \ln(X_i) - \frac{1}{\beta} \sum_{i=1}^{n} X_i$$

- Taking partial derivatives with respect to the parameters gives (the score equations)

$$\frac{\partial}{\partial \alpha} \ell(\alpha, \beta | X_1, \ldots, X_n) = -n \frac{\partial \ln\{\Gamma(\alpha)\}}{\partial \alpha} - n \ln(\beta) + \sum_{i=1}^{n} \ln(X_i)$$

$$\frac{\partial}{\partial \beta} \ell(\alpha, \beta | X_1, \ldots, X_n) = -\frac{\alpha n}{\beta} + \frac{n \bar{X}}{\beta^2}$$

Example: gamma distribution

- Note that solving $\frac{\partial}{\partial\beta}\ell(\alpha,\beta|X_1,\ldots,X_n)$ leads to

$$\frac{\partial}{\partial\beta}\ell(\alpha,\beta|X_1,\ldots,X_n)=0 \Leftrightarrow \hat{\beta}=\frac{\bar{X}}{\alpha}$$

- Plugging the above into $\frac{\partial}{\partial\alpha}\ell(\alpha,\beta|X_1,\ldots,X_n)$, and noting that $\frac{\partial \ln\{\Gamma(\alpha)\}}{\partial\alpha}=\psi(\alpha)$, yields

$$\frac{\partial}{\partial\alpha}\ell(\alpha,\beta|X_1,\ldots,X_n)=-n\psi(\alpha)-n\ln(\bar{X})+n\ln(\alpha)+\sum_{i=1}^{n}\ln(X_i)$$

  And so,

$$\frac{\partial}{\partial\alpha}\ell(\alpha,\beta|X_1,\ldots,X_n)=0 \Leftrightarrow \psi(\alpha)-\ln(\alpha)=\frac{1}{n}\sum_{i=1}^{n}\ln(X_i)-\ln(\bar{X})$$

- There is no closed-form solution for the above, however, we can solve it numerically.

Example: Normal distribution

- Suppose we have a sample size of n, where $X_1, X_2, \ldots, X_n$ are independent and are assumed to come from a normal distribution with mean μ and variance σ^2,

$$X_i \sim \mathcal{N}(\mu, \sigma^2)$$

- In this model, there are two parameters, and thus the parameter vector θ is two-dimensional: $\theta = (\mu, \sigma^2)$; recall:
    - A random variable X follows a $\mathcal{N}(\mu, \sigma^2)$ if $X \in \mathbb{R}$, i.e. X can take on any value in $(-\infty, \infty)$, and has probabililty density function:

    $$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

    for any $x \in \mathbb{R}$

Example: Normal distribution.

- For a sample $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$, the likelihood function is:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \end{aligned}$$

- The log-likelihood is:

$$LL(\theta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

- This is a function of two variables, μ and σ^2. The ML estimators are obtained by finding the values of μ and σ^2 that maximize this function.

Example: Normal distribution

- This is another case where we're able to solve the problem analytically.
- The ML estimators are found by simultaneously solving:

$$\frac{\partial}{\partial \mu} LL(\theta) = 0 \quad \text{and} \quad \frac{\partial}{\partial \sigma^2} LL(\theta) = 0$$

- It can be shown that the MLEs are:

$$\hat{\mu}_{mle} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{\sigma}^2_{mle} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Example: Normal distribution

- Note that we would intuitively expect the estimator of the theoretical mean μ to simply be the sample mean.
- However, the estimate of the theoretical variance σ^2 is slightly different than the one seen in introductory statistics courses: usually the population variance is estimated by the sample variance given by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- The difference between $\hat{\sigma}^2_{mle}$ and S^2 is minimal: in one case we divide by n and in the other we divide by $(n-1)$.
  - The two versions converge to the same value as n tends towards infinity.

Example: Normal distribution

- The ML estimator for the population mean μ is unbiased, meaning that:

$$E(\hat{\mu}_{mle}) = \mu$$

- The sample variance is an unbiased estimator for the population variance σ^2

$$E(S^2) = \sigma^2$$

- Note, however, that

$$\hat{\sigma}^2_{mle} = \frac{n-1}{n} S^2$$

  and thus the ML estimator of the variance is slightly biased:

$$E(\hat{\sigma}^2_{mle}) = \frac{n-1}{n} \sigma^2$$

  But this bias tends towards 0 when n goes to ∞.

- This shows that the MLE is not necessarily unbiased, but the MLE is asymptotically unbiased, i.e., as $n \to \infty$.

Chapter overview

Properties of the ML estimator

- Suppose we have a random sample of size n, $X_1, X_2, \ldots, X_n$ from an underlying model parametrized in terms of θ (i.e. $X_1, \ldots, X_n \overset{iid}{\sim} F_\theta$). Under certain regularity conditions, the ML estimator $\hat{\theta}_{mle}$ of θ has the following properties:
  - $\hat{\theta}_{mle}$ is consistent, that is, $\hat{\theta}_{mle} \xrightarrow{P} \theta$ as $n \to \infty$
  - $\hat{\theta}_{mle}$ is asymptotically normal, specifically

    $$\sqrt{n}(\hat{\theta}_{mle} - \theta) \xrightarrow{d} \mathcal{N}(0, \Sigma_\theta)$$

    where Σ_θ is the asymptotic variance (or covariance matrix) and θ is the true (population level) value of the parameter θ

    The asymptotic variance is given by $\Sigma_\theta = \mathcal{I}(\theta)^{-1}$ where $\mathcal{I}(\theta)$ is the Fisher Information matrix and is given by $\mathcal{I}(\theta) = E_\theta \left\{ \left(\frac{\partial}{\partial \theta} \ \ln f_\theta(X) \right)^2 \right\}$
  - $\hat{\theta}_{mle}$ is asymptotically efficient, that is, it has the smallest asymptotic variance amongst all (unbiased) estimators
  - invariance property: if $\psi = g(\theta)$ for some function $g(\cdot)$, then the MLE of ψ is $\hat{\psi}_{MLE} = g(\hat{\theta}_{MLE})$

Properties of the ML estimator

- The first property is that the ML estimator converges towards the correct value as the sample size increases. Even though it's not necessarily unbiased (as we saw in the last example), this property says that it's asymptotically unbiased.
- The second property is that the ML estimator approximately follows a normal distribution when n is large. We can use this property to perform inference (i.e. calculate CIs and perform hypothesis tests).
- The third property says that the ML estimator is efficient since it has the smallest variance possible amongst a large class of estimators.
- Basically, the ML estimator has several nice properties that makes it a desirable estimation method for statistical analysis.

Example: linear regression

- Recall the linear regression model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \epsilon_i$$

  for $i = 1, \ldots, n$ where ϵ_i are independent $\mathcal{N}(0, \sigma^2)$ random variables.

- This model has $p+2$ parameters: the $p+1$ regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ and the variance parameter σ^2
- In chapter 2 we saw how to estimate the parameters using the least squares method; we will now revisit estimation for the linear regression model using the ML method.

Example: linear regression

Recall:

- The random error terms ϵ_i have mean 0 and thus

$$\mu_i = E(Y_i|X_{i1}, \ldots, X_{ip}) = \beta_0 + \beta_1 X_{i1} + \cdots \beta_p X_{ip}$$

- The random error terms ϵ_i are assumed to follow a normal distribution $\mathcal{N}(0, \sigma^2)$ and thus

$$Y_i|X_{i1}, \ldots, X_{ip} \sim \mathcal{N}(\mu_i, \sigma^2)$$

- Thus, we can write the log-likelihood for the parameter vector $\theta = (\beta_0, \beta_1, \ldots, \beta_p, \sigma^2)$ as

$$\begin{aligned} LL(\theta) &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i-1}^{n}(Y_i - \mu_i)^2 \\ &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i-1}^{n}(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_p X_{ip})^2 \end{aligned}$$

Example: linear regression

- It is clear that maximizing this function with respect to $\beta_0, \dots, \beta_p$ means maximizing

$$-\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{i1} - \ldots - \beta_p X_{ip})^2$$

which is equivalent to minimizing

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{i1} - \ldots - \beta_p X_{ip})^2$$

- But this is exactly what we minimized in the least-squares method; that is, we minimized the sum squared errors!
- Consequently, the estimator of the β parameters from the least-squares method can be seen as ML estimates under the assumption of normality.
- Therefore, the estimators that we used in the last chapter have all the nice properties of ML estimators.

Example: linear regression

- We can verify that the ML estimator for the variance σ^2 is:

$$\hat{\sigma}^2_{mle} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \ldots - \beta_p X_{ip})^2$$

- However, we saw in the last chapter that the most commonly used estimate of σ^2 is:

$$\hat{\sigma}^2 = \frac{SS_E}{n - \text{nuber of parameters in the regression part}}$$

which is equal to:

$$\hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \ldots - \hat{\beta}_p X_{ip})^2$$

- This estimator is unbiased for σ^2. The estimator $\hat{\sigma}^2_{mle}$ is slightly biased, though it's asymptotically unbiased.

Example: linear regression

- The ML method usually gives a slight bias for estimating variance parameters. This bias becomes negligible when the sample size increases, since the ML converges towards the true value. However, we can still correct this.
- The estimation method REML (residual maximum likelihood or restricted maximum likelihood) is a variation on ML that tries to correct this bias.
  - The REML method is based on a slightly different formulation of the *LL* for estimating σ^2, which acknowledges estimation of β.
- This is the method we'll use for mixed linear models for longitudinal/correlated data (and this is often the default method used by various statistical softwares).

Chapter overview

Likelihood and inference

- The theory of ML estimation can be used for inference in different ways
- Wald test:
  - For a null hypothesis of the form $H_0 : \theta = \theta_0$, the test statistic

$$W = \frac{\hat{\theta} - \theta_0}{\hat{se}(\hat{\theta})}$$

    can be derived based on the MLE $\hat{\theta}$ and using the asymptotic properties of the MLE (i.e. that $\hat{\theta}_{mle} \approx \mathcal{N}(\theta, \sigma_\theta / n)$). This general form for a test is referred to as a *Wald test*. For large samples, W is approximately $\mathcal{N}(0,1)$ under H_0.
- Score test:
  - From the log-likelihood function LL, define the following

$$u(\theta) = \frac{\partial \, LL(\theta)}{\partial \theta}, \qquad I(\theta) = -E\left\{ \frac{\partial^2 \, LL(\theta)}{\partial \theta^2} \right\}$$

    The *score statistic*, given by $S = u(\theta_0)/\sqrt{I(\theta_0)}$ can be shown to approximately follow a $\mathcal{N}(0,1)$ distribution under $H_0 : \theta = \theta_0$. This can be used for inference.

Likelihood and inference

- Likelihood ratio test (LRT)
  - For a null hypothesis of the form $H_0 : \theta = \theta_0$, define $LL(\theta_0)$ to be the value of the log-likelihood evaluated at θ_0 and $LL(\hat{\theta}_{mle})$ be the log-likelihood evaluated at the MLE $\hat{\theta}_{mle}$. Define the test statistic D as the difference
  $$D = -2\left\{LL(\theta_0) - LL(\hat{\theta}_{mle})\right\}$$
  For large n, it can be shown that D is approximately χ^2 under H_0 with degrees of freedom equal to the difference in the dimensions of the parameter spaces Θ (unrestricted, H_1) and Θ_0 (under H_0).
  - We'll discuss the LRT again in a few slides...

Variants of likelihood

Sometimes, a variation of the likelihood function can be used for inference

- Suppose a model is parametrized in terms of θ and ϕ, where θ is of interest and ϕ is a *nuisance* parameter.

- Conditional likelihood
  - Suppose that
  $$f(y|\theta,\phi) \propto f(t_1|t_2,\theta)f(t_2|\theta,\phi)$$
  where t_1 and t_2 are statistics (i.e. functions of y). Inference for θ can be based on the *conditional likelihood*
  $$L_c(\theta) = f(t_1|t_2,\theta)$$

- Marginal likelihood
  - Suppose that
  $$f(y|\theta,\phi) \propto f(s_1,s_2,a|\theta,\phi) = f(a)f(s_1|a,\theta)f(s_2|s_1,a,\theta,\phi)$$
  where s_1, s_2, a are statistics. Inference for θ can be based on the *marginal likelihood*
  $$L_m(\theta) = f(s_1|a,\theta)$$

Variants of likelihood

- Profile likelihood
  - The profile likelihood for θ is

$$L_p(\theta) = \max_{\phi} L(\theta, \phi)$$

if $\tilde{\theta}$ is the maximum of $L_p(\theta)$, then $\tilde{\theta} = \hat{\theta}_{mle}$

- Quasi-likelihood
  - Consider a quasi-likelihood function, which has a similar form to a known likelihood (from a known distribution), but slightly different (usually to allow for overdispersion).
- Restricted maximum likelihood (REML)
  - Recall: in the linear regression model, the estimator for the variance parameter σ^2 is SS_E/n and is in fact biased. The idea of REML is to consider an alternative form of the likelihood for estimating σ^2, which acknowledges estimation of β. We will revisit this concept in the context of linear mixed models.

Likelihood-based tools for model comparison

- We often want to compare how well different models fit our data.
  - We need tools to choose the "best" model.
- There are 3 important quantities involving the ML method that provide a measure of model fit:
  - $-2LL(\hat{\theta}_{mle})$
  - AIC
  - BIC

$-2LL(\hat{\theta}_{mle})$

- This is simply -2 times the value of the log-likelihood evaluated at the MLE $\hat{\theta}_{mle}$, i.e. at the maximum of the function LL.
- This can be viewed as a measure of the quality of the model fit. The smaller the value $-2LL(\hat{\theta}_{mle})$, the better the fit.
  - This follows since when the likelihood is large, the better the fit.
- However, this value doesn't account for model complexity.
  - We can make this value as small as we want by building more and more complex models (with more parameters).
  - This can lead to "over-fitting", meaning the model is not a good representation of the underlying relationship.

$-2LL(\hat{\theta}_{mle})$

- Consider the case of linear regression; we can show that

$$-2LL(\hat{\theta}_{mle}) = n\ln(SS_E/n) + n\left(\ln(2\pi) + 1\right)$$

where

$$SS_E = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

is the sum of the squared errors.

- Thus, in the case of linear regression, $-2LL(\hat{\theta}_{mle})$ is essentially equivalent to measuring the quality of the model by the sum of the squared errors (the second term in the expression is constant).
- The problem is that SS_E will always decrease (or at worst stay the same) when we add variables to the model.
- It's thus possible to decrease the value of $-2LL(\hat{\theta}_{mle})$ by adding useless complexity to the model.

AIC and BIC

- The AIC and BIC are quantities that measure how well the model fits the data, while penalizing for the number of model parameters.
- The idea is to find a balance between the model fit and parsimony (i.e. the amount of parameters in the model), therefore guarding against over-fitting.
- They're defined as:
  - $AIC = -2LL(\hat{\theta}_{mle}) + 2(\text{number of parameters in the model})$
  - $BIC = -2LL(\hat{\theta}_{mle}) + \ln(n)(\text{number of parameters in the model})$
- They're very simple to use. The smaller the AIC (or BIC), the better the model fit.
- The BIC penalizes the number of parameters more than the AIC, and will therefore choose a model with fewer parameters.

Likelihood ratio test

- We just saw that the quantity $-2LL(\hat{\theta}_{mle})$ is a measure of model fit and can be used to define model selection criteria (AIC and BIC).
- But these critera do not constitute formal hypothesis tests on the parameters.
- We can also use the quantity $-2LL(\hat{\theta}_{mle})$ to construct tests that compare models.
- *To simplify our notation, we will shorten this to $-2LL$ from now on. This refers to : -2 times the log-likelihood evaluated at the MLE $\hat{\theta}_{mle}$.*

Likelihood ratio test

- To perform a likelihood ratio test (LRT), we need to consider two nested models:
  - a "full" model (this is the "complete", or more complex model)
  - a "reduced" model (this is a subset of the full model, for example, assuming some of the parameters in the full model are equal to 0).
- To formally compare the two models, we look at the difference of the $-2LL$ values from the two models.
- More precisely, the procedure involves fitting two models:
  - The 1st model is the "complete" model. We call this $-2LL(complete)$, i.e. the value of $-2LL$ for this model.
  - The 2nd model is the"reduced" model. We call this $-2LL(reduced)$, i.e. the value of $-2LL$ for this model.
  - For example, the complete model is a regression model with 4 predictor variables and the reduced model includes only the first 2 predictor variables.

Likelihood ratio test

- In the LRT, the underlying null hypothesis is that the complete and reduced models are not different in terms of goodness of fit.
- The test statistic is defined as:

$$\boxed{D = [-2LL(reduced)] - [(-2LL(complete)]}$$

- If H_0 is true, then the difference D between the $-2LL$ values approximately follows a Chi-squared distribution with degrees of freedom equal to the difference in the number of parameters between the two models.
  - We can calculate the p-value for this test using the Chi-squared distribution.
- This is a very general testing procedure that comes from ML estimation.

Chapter overview

What you should know

- Understand what the likelihood function represents
- Understand the principle of maximum likelihood estimation (how to find MLEs, and the properties of these estimators)
- Model comparison criteria: AIC/BIC
- Model comparison tool: LRT test (Likelihood Ratio Test)