Introduction to Model Building - Fitting Continuous Models

All Models are Wrong, But Some are Useful

All models are wrong, but some are useful.—George Box[1]

[Mathematical models] should be made as simple as possible, but not simpler.—Albert Einstein[2]

Beware of geeks bearing formulas.—Warren Buffett, 2009, on the role of "quants" in the 2008 financial collapse.

[1] George Box (1919-2013) is a famous statistician, with several statistical procedures named after him.

[2] The reader is undoubtedly aware of Einstein's (1879-1955) famous theories of relativity, but may not know his connections to probability theory. His work on **Brownian motion**, which describes the path of a molecule as it is bombarded by others, is probabilistic in nature, and later developed into a major branch of probability theory. Einstein was also a pioneer in quantum mechanics, which is probabilistic as well. At one point, he doubted the validity of quantum theory, and made his famous remark, "God does not play dice with the universe."

Modeling in Probability and Statistics

The above quote by Box says it all. Consider for example the family of normal distributions. In real life, random variables are bounded—no person's height is negative or greater than 500 inches—and are inherently discrete, due to the finite precision of our measuring instruments. Thus, technically, no random variable in practice can have an exact normal distribution. Yet the assumption of normality pervades statistics and has been enormously successful, provided one understands its approximate nature.

The Essence of Modeling: So, the field of probability and statistics is fundamentally about *modeling*. The field is extremely useful, provided the user understands the modeling issues well. For this reason, the book contains this separate chapter on modeling issues.

Introduction

One often models one's data using a parametric family, as in Chapters 5 and 6. This chapter introduces this approach, involving core ideas of statistics, closely related to each other:

- Why might we want to fit a parametric model to our sample data?
- How do we fit such a model, i.e., how do we estimate the population parameters from our sample data?
- What constitutes a good fit?

Focus on Parametric Density Models

Our focus here will be on fitting parametric density models, thus on continuous random variables. However, the main methods introduced, the **Method of Moments** and **Maximum Likelihood Estimation**, do apply to discrete random variables as well.

Why Fit a Parametric Model?

Denote our data by $X_1, \ldots, X_n$. It is often useful to fit a parametric density model to the data. One might ask, though, why bother with a model?

- Isn't, say a histogram (see below) enough to describe the data?
- There are a couple of answers to this:

Reasons for Why Histogram are not Enough

- In our first example below, we will fit the gamma distribution. The gamma is a two-parameter family, and it's a lot easier to summarize the data with just two numbers, rather than the 20 bin heights in the histogram.
- In many applications, we are working with large systems consisting of dozens of variables. In order to limit the complexity of our model, it is desirable to have simple models of each component. For example, in models of queuing systems, if things like service times and job interarrival times can be well modeled by an exponential distribution, the analysis may simplify tremendously, and quantities such as mean job waiting times can be easily derived.

Model-Free Estimation of a Density

Model-Free Estimation of a Density

Before we start with parametric models, let's see how we can estimate a density function without them. This will introduce central issues that will arise again in regression models and machine learning, Chapter 15.

How can we estimate a population density from our sample data?

It turns out that the common histogram, so familiar from your instructors' summaries of the "distribution" of exam scores, is actually a density estimator!

Although densities themselves are not probabilities, they do tell us which regions will occur often or rarely. That is exactly what a histogram tells us.

A Closer Look

Let $X_1, X_2, \ldots, X_n$ denote our data, a random sample from a population with density f_X. Say bin i in a histogram covers the interval $(c, c+w)$. Let N_i denote the number of data points falling into the bin. This quantity has a binomial distribution with n trials and success probability

$$p = P(c < X < c+w) = \text{area under } f_X \text{ from } c \text{ to } c+w.$$

If w is small, then this implies $p \approx w \cdot f_X(c)$.

Let $\hat{p}$ and $\hat{f}_X(c)$ be estimators of p and $f_X(c)$, respectively. But since p is the probability of an observation falling into this bin, we can estimate it by $\hat{p} = \frac{N_i}{n}$ and $\hat{p} = w \cdot \hat{f}_X(c)$.

So, we have an estimate of f_X: $\hat{f}_X(c) = \frac{\hat{p}}{w} = \frac{N_i}{wn}$.

So, other than a constant factor $\frac{1}{wn}$, our histogram, which plots the N_i, is an estimate of the density f_X.

Example: BMI Data

Consider the Pima Indians diabetes study from Ch7 of the book. One of the columns is Body Mass Index (BMI). Let's plot a histogram:

```
pima <- read.csv('diabetes.csv', header = FALSE )
bmi <- pima [,6]
bmi <- bmi[bmi > 0]
hist(bmi, breaks =19, freq = FALSE )
```

The plot is shown below:

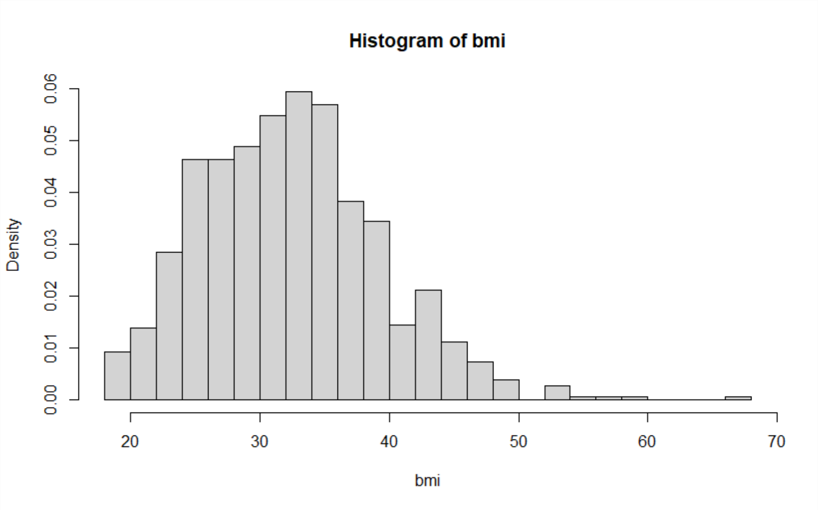

Figure 1: BMI, 20 bins

The Number of Bins

Why is there an issue with the number of bins?

- If we use too many bins, the graph will be quite choppy. Next figure shows a histogram for the BMI data with 100 bins. Presumably, the true population density is pretty smooth, so the choppiness is a problem.
- On the other hand, if we use too few bins, each bin will be very wide, so we won't get a very detailed estimate of the underlying density. In the extreme, with just one bin, the graph becomes completely uninformative.

It's instructive to think of the issue of choosing the number of bins in terms of variance and bias, the famous bias-variance tradeoff. This is a fundamental issue in statistics. We'll discuss it here in the context of density estimation (more on this in Chapter 15).

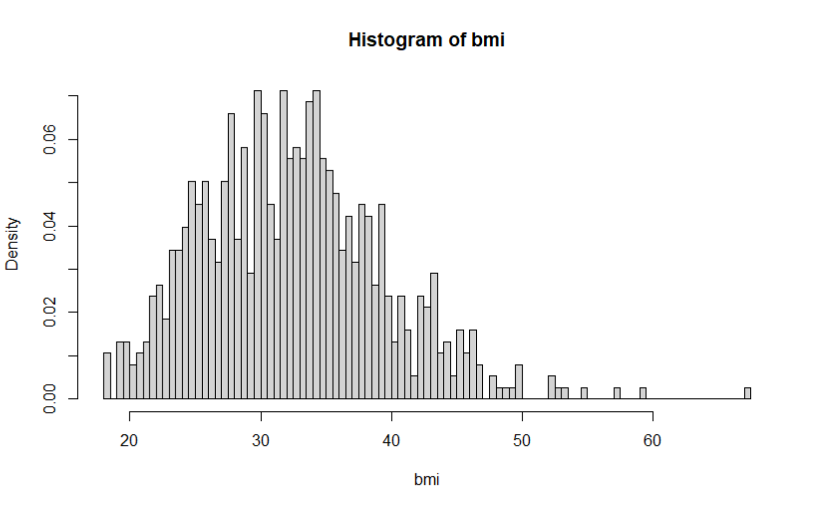

Figure 2: BMI, 100 bins

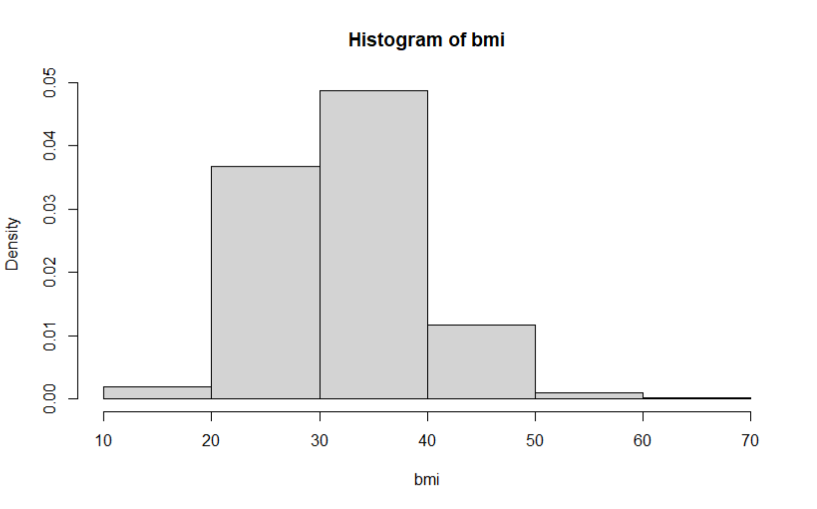

Figure 3: BMI, 3 bins

The Mean Squared Error (MSE)

Suppose we wish to estimate some population quantity θ, using an estimator $\hat{\theta}$ computed from our sample data. Then we hope to keep the mean squared error,

$$MSE = E[(\hat{\theta} - \theta)^2],$$

as small as possible. Let's expand that quantity. Write

$$\hat{\theta} - \theta = (\hat{\theta} - E[\hat{\theta}]) + (E[\hat{\theta}] - \theta) = \text{deviation} + \text{bias}.$$

So, we need to find

$E[(\text{deviation}+\text{bias})^2] = E(\text{deviation}^2)+E(\text{bias}^2)+2E(\text{deviation}\times\text{bias})$.

But bias and hence bias^2 are both constant; since bias $= E[\hat{\theta}] - \theta$ and $E[\hat{\theta}]$ and θ are both constants. So, $E(\text{bias}^2) = \text{bias}^2$ and $2E(\text{deviation} \times \text{bias}) = 2 \times \text{bias} \times E(\text{deviation})$

The Mean Squared Error (MSE) - cont'd

Also, $E(\text{deviation} = E[\hat{\theta} - E[\hat{\theta}]] = 0$ since $E[\hat{\theta}]$ is constant and $E[\hat{\theta} - E[\hat{\theta}]] = E[\hat{\theta}] - E[E[\hat{\theta}]] = E[\hat{\theta}] - E[\hat{\theta}] = 0$ So, $2E(\text{deviation} \times \text{bias}) = 2 \times \text{bias} \times E(\text{deviation}) = 0$.

Finally, $E(\text{deviation}^2) = E[(\hat{\theta} - E[\hat{\theta}])^2] = Var(\hat{\theta})$ by definition of variance.

Thus, we get the famous formula:

$$MSE = \text{Variance} + \text{Bias}^2.$$

The Bias-Variance Tradeoff - e.g. for the Two Variance Estimators

It is called a *tradeoff* because those two terms are often at odds with each other. For instance, in estimating population variance σ^2:

- The classic estimator s^2 has zero bias, whereas bias of s_m^2 is nonzero. So, the classic estimator is better in that its second term in the MSE formula is smaller.
- On the other hand, since $\frac{1}{(n-1)} > \frac{1}{n}$, the classic estimator has a larger variance, by a factor of $(n/(n-1))$ larger. Thus, s_m^2 has a smaller first term in the MSE formula.

The overall "winner" will depend on n and the size of variances of s^2 and s_m^2. Calculating the latter would be too much of a digression here, but the point is that there IS a tradeoff.

The Bias-Variance Tradeoff in the Histogram Case

Let's look at the bin width issue in the context of variance and bias.

- If the bins are too narrow, then for a given bin size, there will be a lot of variation in height of that bin from one sample to another. In other words, the variance of the height will be large.
- On the other hand, making the bins too wide produces a bias problem. Suppose, for instance, the true density $f_X(t)$ is increasing in t. Then within a bin, our estimate $\hat{f_X}(t)$ will tend to be too low near the left end of the bin and too high on the right end. If the number of bins is small, then the bin widths will be large, and bias may be a serious issue.

A General Issue: Choosing the Degree of Smoothing

Recall the quote in the Preface of this book, from the ancient Chinese philosopher Confucius:

> *[In spite of] innumerable twists and turns, the Yellow River flows east.*

Confucius' point was basically that one should, as we might put it today, "Look at the big picture," focusing on the general eastward trend of the river, rather than the local kinks. We should visually "smooth" our image of the river.

In a histogram, the fewer the number of bins, the more smoothing is done. So, choosing the number of bins can be described as choosing the amount of smoothing. This is a central issue in statistics and machine learning, and will play a big role in Chapter 15 as well as here.

Automatic Selection of the Number of Bins

There are various methods for automatic selection of the number of bins. They are too complex to discuss here, but the R package `histogram` offers several such methods. Here is the package in action on the BMI data:

```
hist(bmi, freq=FALSE)
```

The default in R is that all bin widths are equal. The 'freq' argument set to 'FALSE' indicates that the histogram will display the density of the data rather than counts. This means each bar height will represent the density of observations within each bin, showing the distribution shape of the 'bmi' variable.

The plot is shown below.

Note that 11 bins were chosen. The graph looks reasonable here, but the reader should generally be a bit wary of automatic

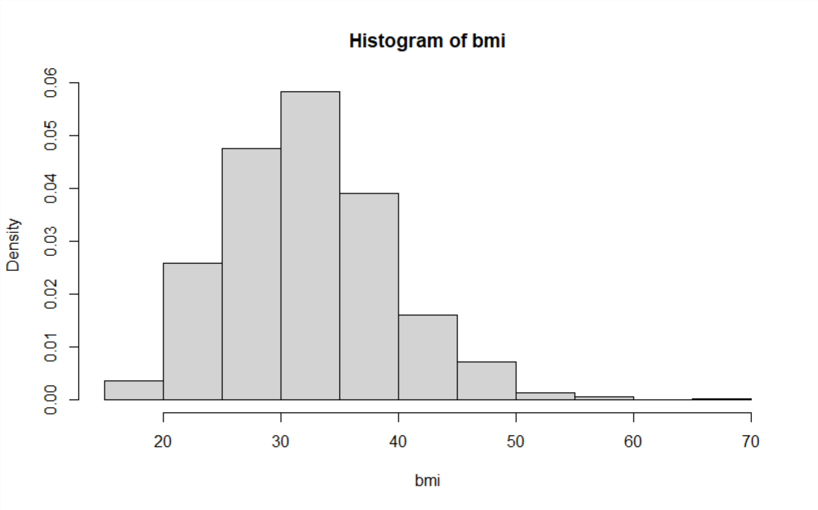

Figure 4: BMI, with R defaults

Advanced Methods for Model-Free Density Estimation

Even with a good choice for the number of bins, histograms are still rather choppy in appearance. Kernel methods aim to remedy this.

To see how they work, consider again a bin $[c - \delta, c + \delta]$ in a histogram. Say we are interested in the value of the density at a particular point t_0 in the interval. Since the histogram has constant height within the interval, that means that all data points X_i in the interval are treated as equally relevant to estimating $f_X(t_0)$.

By contrast, kernel methods put more weight on points closer to t_0. Even points outside the interval may be given some weight.

The mathematics gets a bit complex (see the Mathematical Complements section at the end of the Chapter 8) and we just show how to use this method in base R, via the `density()` function.

Using the `density()` Function in R

As with many R functions, `density()` has many optional arguments. We'll stick the defaults here, but the bandwidth, `bw`, controls the degree of smoothing, as the bin width does for histograms.

The call then is simply:

```
plot(density(bmi))
```

Note that the output of `density()` is just the estimated density values, and must be run through `plot()` to be displayed. By doing things this way, it is easy to plot more than one density estimate on the same graph.

The graph is shown below.

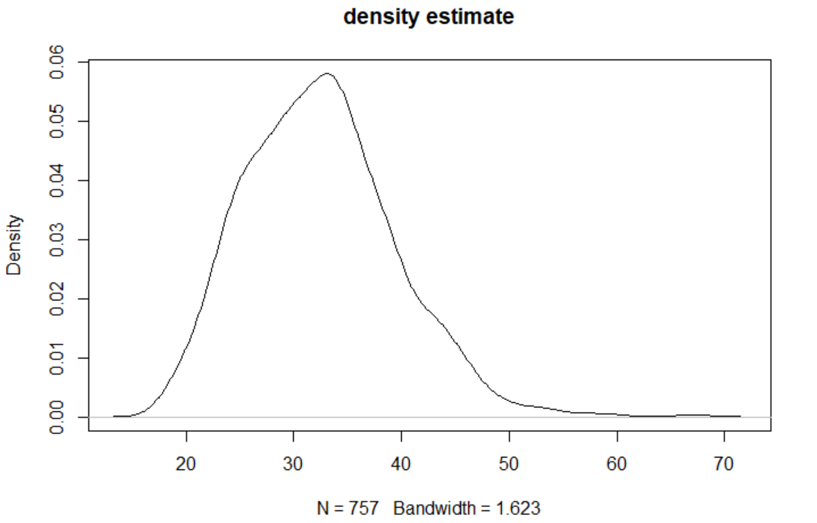

Figure 5: Kernel density estimate, BMI data

Parameter Estimation - Parametric Estimation of a Density

Parameter Estimation

To fit a parametric model such as the Gamma distribution to our data, the question then arises as to how to estimate the parameters.

Earlier, we often referred to certain estimators as being “natural”. For instance, estimating a population mean with a sample mean is intuitive. However, in many cases, it’s less clear what a “natural” estimate might be. This section introduces general methods for estimation.

Two common methods for estimating the parameters of a density are the Method of Moments (MM) and Maximum Likelihood Estimation (MLE) (two other methods are Bayesian Estimation, and EM Algorithm, not covered).

We’ll introduce these two via examples.

Method of Moments

Method of Moments (MM) gets its name from the fact that quantities like mean and variance are called moments. $E(X^k)$ is the kth moment of X, with $E[(X - EX)^k]$ being termed the kth central moment. If we have an m-parameter family, we "match" m moments.

- Oldest method of point estimation dating back to Karl Pearson in the late 1800's.
- Idea is simple, and usually, the resulting estimators need to be improved.

Method of Moments

In general, let $X_1, X_2, \ldots, X_n$ be iid from pmf or pdf $f(x|\theta_1, \ldots, \theta_k)$, we have

	sample moments	population moments
1^{st} moment	$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$	$\mu_1 = \mu_1(\theta_1, \ldots, \theta_k) = EX$
2^{nd} moment	$m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$	$\mu_2 = \mu_2(\theta_1, \ldots, \theta_k) = EX^2$
$\vdots$	$\vdots$	$\vdots$
kth moment	$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$	$\mu_k = \mu_k(\theta_1, \ldots, \theta_k) = EX^k$

To get the MoM estimators, "equate" the first k sample moments to the corresponding k population moments and solve for $(\theta_1, \ldots, \theta_k)$ in terms of $(m_1, \ldots, m_k)$. Usually, it may also help using $s_m^2 = E[(X - EX)^2] = Var(X)$.

Example: BMI Data

Let's see how well the model fits, at least visually, with an example. Recall that for a gamma-distributed $X \sim Gamma(\alpha, \beta)$, where α is the shape and β is the rate parameter,

$$E(X) = \frac{\alpha}{\beta} \quad \text{and} \quad Var(X) = \frac{\alpha}{\beta^2}$$

In MM, we simply replace population values by sample estimates (i.e. moments) in the above equations,

$$\bar{X} = \frac{\alpha}{\beta} \quad \text{and} \quad s_m^2 = \frac{\alpha}{\beta^2}$$

and solving for α and β, which yields:

$$\hat{\beta} = \frac{\bar{X}}{s_m^2} \quad \text{and} \quad \hat{\alpha} = \frac{\bar{X}^2}{s_m^2}$$

Note that we put hats "ˆ" to indicate that these are estimators of the corresponding unknown parameters.

Visual Fit of the Model

Let's see how well the model fits, at least visually:

- $\bar{x}$ is the sample mean of BMI.
- s_m^2 is the sample variance of BMI.
- $\hat{\alpha}$ is the estimate of the shape parameter.
- $\hat{\beta}$ is the estimate of the rate parameter.

The plot is shown below.

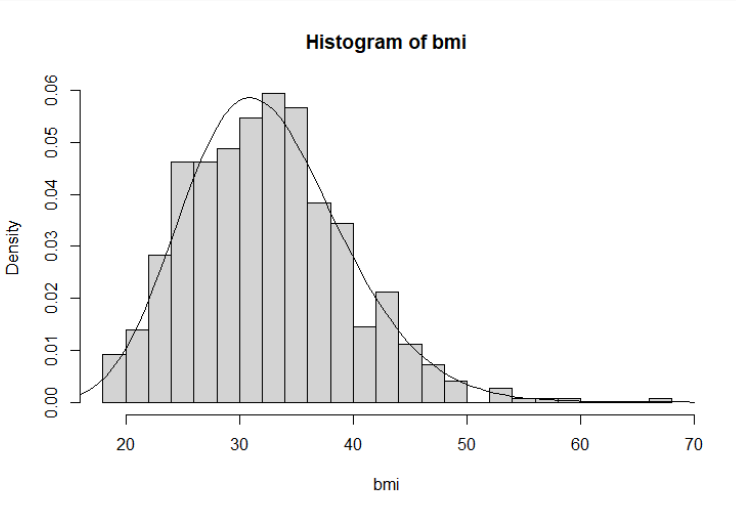

Figure 6: BMI, histogram and MoM Gamma fit

Visually, the fit looks fairly good, but keep in mind possible sources of discrepancy between the fitted model and the histogram.

Normal MoM Estimator

Suppose $X_1, X_2, \ldots, X_n$ are iid from a $N(\mu, \sigma^2)$ distribution. Find the MM estimators of μ and σ^2.

Solution:

$$\mu = \bar{X}, \quad \sigma^2 = s_m^2$$

We can get:

$$\widehat{\mu} = \bar{X}, \quad \widehat{\sigma}^2 = s_m^2 = \frac{n-1}{n} s^2$$

Binomial MoM Estimator

Suppose $X_1, X_2, \ldots, X_n$ are iid from a Bin(κ, p) distribution where both κ and p are unknown. Find the MM estimators of κ and p.

Solution: Recall that for $X \sim \text{Bin}(\kappa, p)$,

$$EX = \kappa\, p \quad \text{and} \quad Var(X) = \kappa\, p\,(1-p).$$

Then from $\overline{X} = \kappa p$ and $s_m^2 = \kappa\, p\,(1-p)$, we have

$$\widehat{\kappa} = \frac{\overline{X}^2}{\overline{X} - s_m^2} \quad \text{and} \quad \widehat{p} = \frac{\overline{X}}{\widehat{\kappa}}$$

Method of Maximum Likelihood

Preferred Method:

- The Maximum Likelihood Estimation (MLE) is widely used due to its strong theoretical properties.

General Procedure for MLE
Maximizing Likelihood:

- Given a sample $X_1, \ldots, X_n$ and parameters $\theta_1, \ldots, \theta_k$, we find the values that maximize the likelihood function.
- For differentiable problems, we maximize the log likelihood by setting its derivatives with respect to θ_j to 0 and solving for the estimators.

MLE - More Formally

Let $X_1, X_2, \ldots, X_n$ be iid from pdf or pmf $f\left(x|\theta_1, \ldots, \theta_k\right)$. The likelihood function is defined as

$$L(\boldsymbol{\theta}|\mathbf{x}) = L(\theta_1, \ldots, \theta_k | x_1, \ldots, x_n) = \\ f(x_1, \ldots, x_n | \theta_1, \ldots, \theta_k) = \prod_{i=1}^{n} f\left(x_i | \theta_1, \ldots, \theta_k\right).$$

Definition: For each sample point $\mathbf{x}$, let $\widehat{\theta}(\mathbf{x})$ be a parameter value at which $L(\theta|\mathbf{x})$ attains its maximum as a function of θ, with $\mathbf{x}$ held fixed. A *maximum likelihood estimator* (MLE) of the parameter θ based on a sample $\mathbf{X}$ is $\widehat{\theta}(\mathbf{X})$.

Log Likelihood Function

In most cases, especially when differentiation can be used, it is easier to work with the (natural) logarithm of likelihood, $\log L(\theta|\mathbf{x})$ (also known as the *log likelihood function*).

Why is this okay?

- The logarithm is a strictly increasing function, so the log likelihood attains its maximum at the same point as the likelihood itself.
- Logarithms transform products into sums, which are easier to differentiate.

Example (Bernoulli MLE)

Let $X_1, X_2, \ldots, X_n$ be iid Ber(p). Find the MLE of p for $0 \leq p \leq 1$.

Solution:

Given that $X_1, X_2, \ldots, X_n$ are iid Ber(p), the likelihood function for p is given by

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

Taking the logarithm $\log L(p) = \log\left(p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}\right) = (\sum_{i=1}^{n} x_i)\log(p) + (n - \sum_{i=1}^{n} x_i)\log(1-p)$.

Taking the derivative of $\log L(p)$ with respect to p and setting it to zero:

$$\frac{d}{dp}\log L(p) = \left(\sum_{i=1}^{n} x_i\right)\frac{1}{p} - \left(n - \sum_{i=1}^{n} x_i\right)\frac{1}{1-p} = 0.$$

Example (Bernoulli MLE) cont'd

Solving for p, we get

$$\widehat{p} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x},$$

which is the sample mean.

Therefore, the MLE of p is $\widehat{p} = \overline{X}$. $\square$

In finding MLEs, maximization takes place only over the range of the parameter space. If $L(\theta|\mathbf{x})$ cannot be maximized analytically, it may be possible to maximize it numerically.

Example (Normal$(\theta, 1)$)

Let $X_1, X_2, \ldots, X_n$ be iid $N(\theta, 1)$, find the MLE of θ.

Solution:

Given that $X_1, X_2, \ldots, X_n$ are iid $N(\theta, 1)$, the likelihood function for θ is given by

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \theta)^2\right) =$$

$$(2\pi)^{-n/2} \exp\left(-\sum_{i=1}^{n} \frac{1}{2}(x_i - \theta)^2\right).$$

Taking the logarithm of the likelihood function yields

$$\log L(\theta) = -\sum_{i=1}^{n} \left(\frac{1}{2}(x_i - \theta)^2\right) - \frac{n}{2}\log(2\pi).$$

Example (Normal$(\theta, 1)$) - cont'd

To find the MLE of θ, we need to maximize the log-likelihood function with respect to θ. We do this by taking the derivative of $\log L(\theta)$ with respect to θ and setting it to zero:

$$\frac{d}{d\theta} \log L(\theta) = \sum_{i=1}^{n} (x_i - \theta) = 0.$$

Solving for θ, we get the MLE as $\widehat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{X}$.

Maximization Over Parameter Space

Example (Binomial MLE, Number of Trials Unknown) Let $X_1, X_2, \ldots, X_n$ be iid $\mathrm{Bin}(k, p)$. Find the MLE of k where p is known and k is unknown.

Solution: The likelihood function is:

$$L(k|p, \mathbf{x}) = \prod\nolimits_{i=1}^{n} \binom{k}{x_i} p^{x_i}(1-p)^{n-x_i}.$$

Then consider the ratio: $L(k|p, \mathbf{x})/L(k-1|p, \mathbf{x})$.

Invariance Property of MLEs

Theorem (Invariance Property of MLEs) If $\widehat{\theta}$ is the MLE of θ, then $g(\widehat{\theta})$ is the MLE of $g(\theta)$ for any function $g(\theta)$.

Example:

- Let $X_1, X_2, \ldots, X_n$ be iid $N(\theta, 1)$, what is the MLE of θ^2?
- Let $X_1, X_2, \ldots, X_n$ be iid Ber(p), find the MLE of the variance and standard deviation of X_i.

Solution:

(Normal MLEs - μ and σ^2 Unknown)

Let $X_1, X_2, \ldots, X_n$ be iid from a $N(\mu, \sigma^2)$ where both μ and σ^2 are unknown. Then show that the MLE of μ is $\widehat{\mu} = \overline{X}$ and the MLE of σ^2 is $\widehat{\sigma}^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = s_m^2$.

Solution:

Given $X_1, X_2, \ldots, X_n$ iid from a $N(\mu, \sigma^2)$, the likelihood function is

$$L(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

$$= (2\pi)^{-n/2}\sigma^{-n} \exp\left(-\sum_{i=1}^{n}\frac{(x_i - \mu)^2}{2\sigma^2}\right).$$

Taking the natural logarithm of the likelihood function,

$$\log L(\mu, \sigma) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \sum_{i=1}^{n}\frac{(x_i - \mu)^2}{2\sigma^2}.$$

(Normal MLEs - μ and σ^2 Unknown) - cont'd

To maximize the log-likelihood, we take partial derivatives with respect to μ and σ and set them to zero:

For μ:

$$\frac{\partial}{\partial\mu}\log L(\mu,\sigma)=\sum_{i=1}^{n}\frac{x_i-\mu}{\sigma^2}=0,$$

which yields the MLE $\quad\widehat{\mu}=\overline{X}$.

For σ:

$$\frac{\partial}{\partial\sigma}\log L(\mu,\sigma)=-\frac{n}{\sigma}+\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^3}=0,$$

which yields the MLE $\quad\widehat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2=s_m^2$.

Estimation of Parameters of a Gamma Distribution

Problem Statement:

- Given a random sample $X_1, X_2, \ldots, X_n$ from a Gamma(c, λ) distribution with parameters c (shape) and λ (rate). Estimate the parameters.

Method of Moments
Two Parameters:

- We have two parameters to estimate: c (shape) and λ (rate).
- Use the first two moments of X, which can conveniently be the expected value EX and variance $Var(X)$.

Moment Equations:

- $EX = \frac{c}{\lambda}$ and $Var(X) = \frac{c}{\lambda^2}$.
- From the sample, we estimate $\widehat{\lambda} = \frac{\overline{X}}{s_m^2}$ and $\widehat{c} = \frac{\overline{X}^2}{s_m^2}$.

MLEs for Gamma Distribution Parameters

Given a random sample $X_1, X_2, \ldots, X_n$ from a Gamma distribution with shape parameter c and rate parameter λ, the likelihood function is:

$$L(c, \lambda) = \prod_{i=1}^{n} \frac{\lambda^c X_i^{c-1} e^{-\lambda X_i}}{\Gamma(c)} = \frac{\lambda^{nc} (\prod_{i=1}^{n} X_i^{c-1}) e^{-\lambda \sum_{i=1}^{n} X_i}}{(\Gamma(c))^n},$$

where $\Gamma(c)$ is the gamma function.

The log-likelihood function is given by:

$$\log L(c, \lambda) = nc \log \lambda + (c-1) \sum_{i=1}^{n} \log X_i - \lambda \sum_{i=1}^{n} X_i - n(\log \Gamma(c))$$

MLEs for Gamma Distribution Parameters

To find the MLE of c and λ, we need to maximize the log-likelihood function with respect to both c and λ. This is done by solving the following system of equations obtained by setting the partial derivatives of $\log L$ with respect to c and λ to zero:

$$\frac{\partial}{\partial c} \log L(c, \lambda) = 0, \quad \frac{\partial}{\partial \lambda} \log L(c, \lambda) = 0.$$

The solutions to these equations give the MLEs for c and λ. However, these equations do not generally have closed-form solutions (why?) and numerical methods are usually required to find the MLEs.

MLEs for Gamma Distribution Parameters

- The likelihood is the product of the densities for the X_i.
- We generally maximize the log likelihood for convenience.

Maximizing the Log Likelihood:

- The log likelihood for the gamma distribution involves the gamma function and its derivatives, which complicates the solution.
- Numerical methods are required to find the MLEs of c and λ.

R's mle() Function

Finding MLEs Numerically:

- R provides the `mle()` function (in `stats4` package) for finding MLEs numerically.
- The user defines a function for the negative log likelihood, and `mle()` handles the rest.
- Initial parameter guesses are required, and standard errors of estimates can be obtained.

MLE Example in R

Using mle() to Estimate Parameters:

```
ll <- function(c, lambda) {
   -sum(dgamma(x, shape=c, rate=lambda, log=TRUE))
}
summary(mle(minuslogl=ll, start=list(c=1.5, lambda=2)))
```

Output:

```
Coefficients:
       Estimate Std. Error
c      2.147605 0.08958823
lambda 1.095344 0.05144393
-2 log L: 3072.181
```

Assessing “Goodness of Fit” of a Model

Introduction to Goodness of Fit

Our example above concerned how to estimate the parameters of a gamma distribution, given a sample from the distribution. But before estimating parameters, we need to decide if the gamma model, or any model, is appropriate. How do we assess the "Goodness of Fit" of a model?

Less Formal Methods of Model Assessment

Less formal methods of assessing the fit of a model include visualizing the data through histograms or empirical cumulative distribution functions (ecdfs). These can be plotted against a fitted model to assess suitability.

Plotting ecdfs:
To visualize the fit of a model, one can plot the ecdf against a fitted model. Here's an example using R code for BMI values:

```r
#plotting ecdf's for bmi
ebmi <- ecdf (bmi)
plot (ebmi ,cex =.1, xlim=c(15,60))
curve(pgamma(x, ch, lh), 15, 60, col=2,add = TRUE)
```

Sample ecdfs

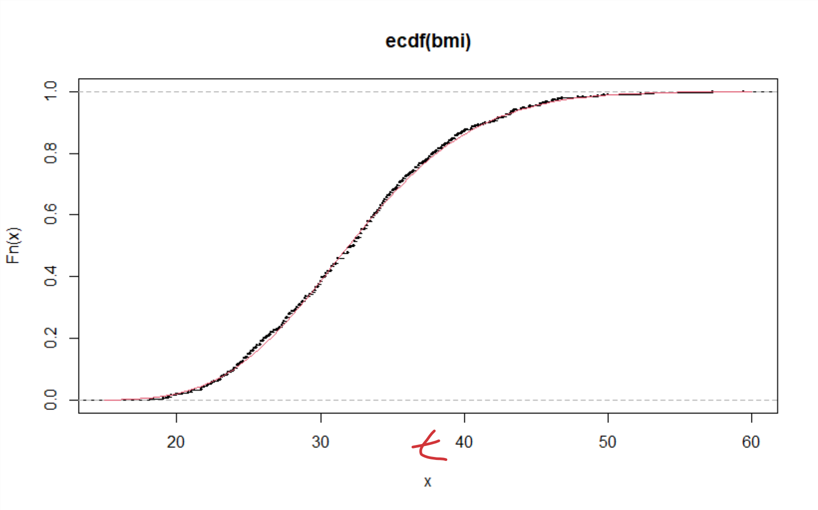

Figure 7: Ecdf and fitted cdf

Interpretation of the Plot

The plot compares the ecdf of the data (in black) to the cdf of a fitted Gamma distribution (in red). Deviations between the curves indicate areas where the Gamma model does not perfectly represent the data distribution.

Limitations of Density Estimation

While we can also compare data to models using density estimation, choosing parameters such as bin width or bandwidth is problematic. There is no foolproof method for selecting these parameters, despite the existence of theoretical guidelines aimed at minimizing integrated mean squared error.

More Formal Assessment of Goodness of Fit

In our examples above, we can do a visual assessment of how well our model fits the data, but it would be nice to have a quantitative measure of goodness of fit.

The classic assessment tool is the Chi-Squared Goodness of Fit Test, which is one of the oldest statistical methods (1900!) and still in wide use. However, a more useful measure is the Kolmogorov-Smirnov (KS) statistic.

Kolmogorov-Smirnov Statistic:

- The Kolmogorov-Smirnov (KS) statistic is a useful measure for the discrepancy between a fitted model and the true population distribution.
- It quantifies the maximum difference between the empirical cumulative distribution function (CDF) of the sample data and the CDF of the theoretical model.

Application: Fitting a Gamma Model

- Example: Fitting a Gamma distribution to BMI data.
- The KS statistic can help assess how well the beta model represents the actual data.

CDFs and the Gamma Model

- Cumulative Distribution Functions (CDFs) play a key role in KS statistic calculation.
- The `pgamma()` function in R provides the CDF for the Gamma distribution.

$$\frac{\#\text{ of obs} \le t}{n}$$

Empirical CDF of the Data

- The empirical CDF, $\hat{F}_X(t) = \frac{M(t)}{n}$, where $M(t)$ is the count of observations less than or equal to t. $F_X(t) = P(X \le t)$
- In R, the `ecdf()` function calculates the empirical CDF.

Kolmogorov-Smirnov (KS) Statistic

To make matters concrete, say we are fitting a Gamma model, with the BMI data. The KS statistic is based on cumulative distribution functions (cdfs) and measures the maximum discrepancy between the empirical cdf and the fitted cdf.

(The values of *ch* and *lh* had been previously computed.)

Now, to quantify the fit, we can calculate the K-S statistic, which measures the maximum discrepancy:

```
> ks_result <- ks.test(bmi, "pgamma", ch, lh)
> ks_result$statistic
         D
0.02375358
```

Since cdf values range in [0, 1], a maximum discrepancy of 0.024 is considered pretty good.

Accounting for Sampling Variation

To account for sampling variation, we can use a K-S confidence band.

The K-S statistic measures the fit in terms of the maximum discrepancy, which is subject to sampling variation. A K-S confidence band provides a way to quantify the confidence in the goodness of fit assessment.

(The details of constructing a K-S confidence band can be found in reference [20].)

In summary, the Kolmogorov-Smirnov (KS) statistic is a valuable tool for quantitatively assessing the goodness of fit between a fitted model and the true population distribution.