

# Likelihoods

History of the course:

## 0.1 Outline

# 1 Likelihood Construction and Estimation

Why do Statisticians love likelihood-based estimation?

1.

2.

3.

4.

Downsides?

1.

2.

## 1.1 Introduction

**Definition:** Suppose random variables  $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$  has joint density or probability mass function  $f_{\mathbf{Y}}(\mathbf{y}, \boldsymbol{\theta})$  where  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_b)$ . Then the *likelihood function* is

$$L(\boldsymbol{\theta}|\mathbf{Y}) = f_{\mathbf{Y}}(\mathbf{Y}, \boldsymbol{\theta}).$$

**Key concept:** In all situations, the likelihood is the joint density of the observed data to be analyzed.

### 1.1.1 Notation

Given  $\mathbf{y}$ , note that  $L(\boldsymbol{\theta}|\mathbf{y}) : \mathbb{R}^b \rightarrow \mathbb{R}$ .

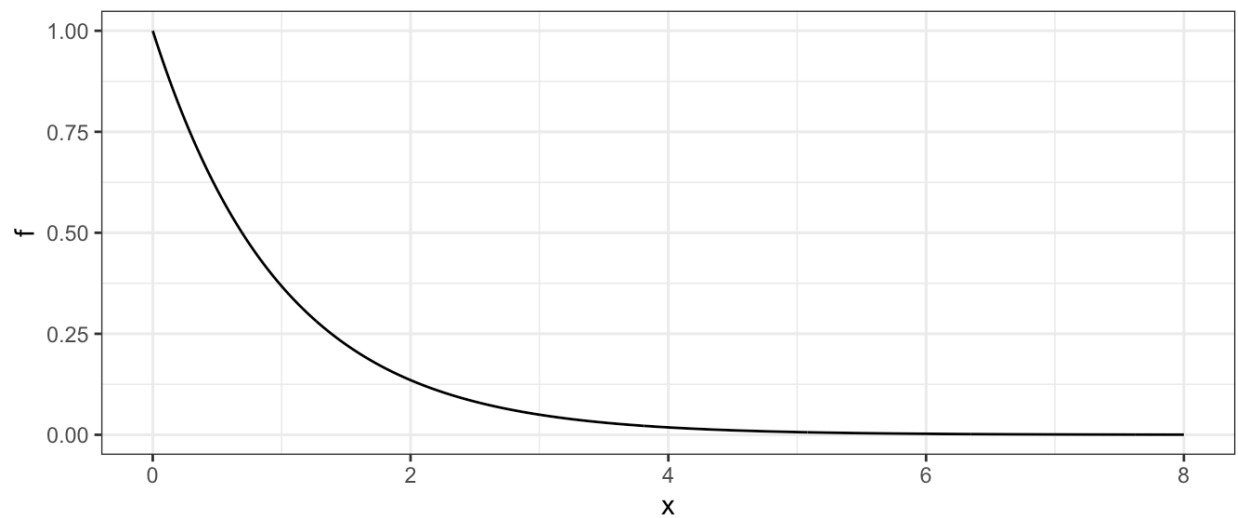
Generally, we optimize  $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}|\mathbf{y})$ .

How?

**Example:** Suppose we have  $Y_1, \dots, Y_n \stackrel{iid}{\sim} \text{Exp}(\lambda)$ . The likelihood function is defined as

```
# likelihood simulation
n <- 10
lambda <- 1

# plot of exponential(lambda) density
data.frame(x = seq(0, 8, .01)) |>
  mutate(f = dexp(x, rate = lambda)) |>
  ggplot() +
  geom_line(aes(x, f))
```



```

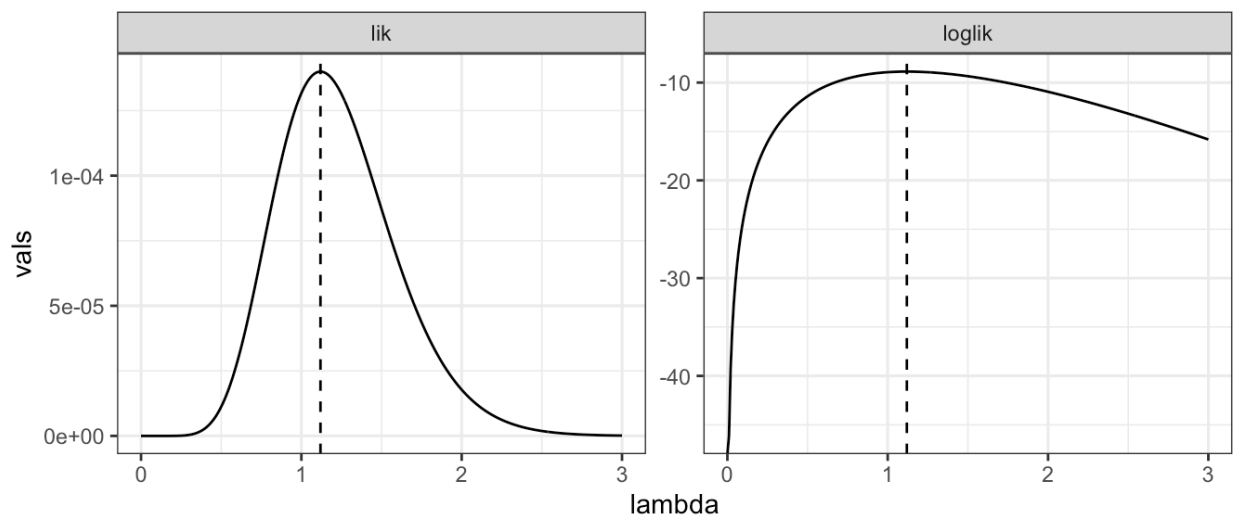
# define likelihood
loglik <- function(lambda, data)
{
  lik <- prod(dexp(data, rate = lambda))
  loglik <- sum(dexp(data, rate = lambda, log = T))

  out <- data.frame(lik = lik, loglik = loglik)
  return(out)
}

# simulate data
data <- rexp(n = n, rate = lambda)

# plot likelihood and loglikelihood
data.frame(lambda = seq(0, 3, by = .01)) |>
  rowwise() |>
  mutate(loglik = loglik(lambda, data)) |>
  unnest(cols = c(loglik)) |>
  pivot_longer(-lambda, names_to = "func", values_to =
"vals") |>
  ggplot() +
  geom_vline(aes(xintercept = 1 / mean(data)), lty = 2) + #
max likelihood estimate is 1/mean
  geom_line(aes(lambda, vals)) +
  facet_wrap(~func, scales = "free")

```



The likelihood function is random!

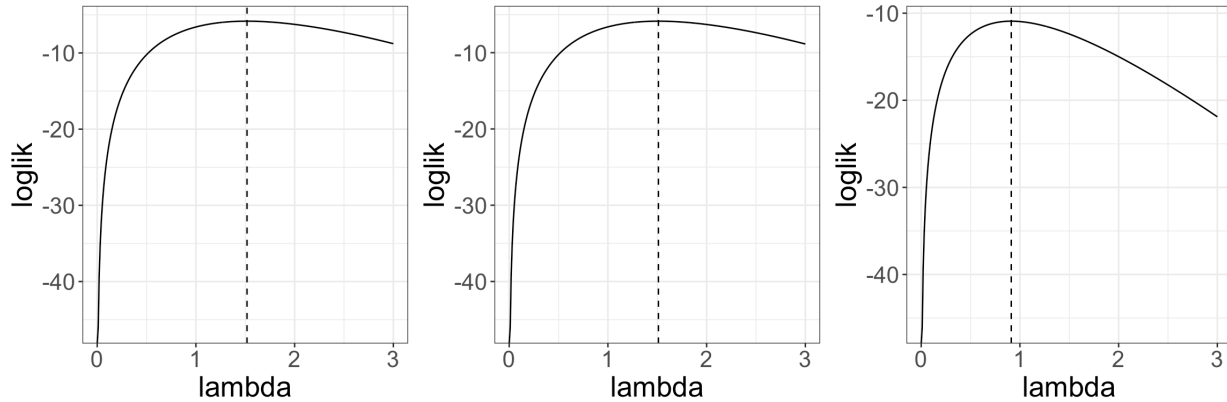
```

for(i in seq_len(3)) {
  # simulate data
  data <- rexp(n = n, rate = lambda)

  # plot likelihood and loglikelihood
  data.frame(lambda = seq(0, 3, by = .01)) |>
    rowwise() |>
    mutate(loglik = loglik(lambda, data)) |>
    unnest(cols = c(loglik)) |>
    ggplot() +
      geom_vline(aes(xintercept = 1 / mean(data)), lty = 2) +
      # max likelihood estimate is 1/mean
      geom_line(aes(lambda, loglik)) +
      theme(text = element_text(size = 20)) -> p ## make
      legible in notes

  print(p)
}

```



**Your Turn:** What is the effect of sample size on the log-likelihood function? Make a plot showing the log-likelihood function that results from  $n = 10$  vs.  $n = 100$  with corresponding MLE.

## 1.2 Construction

The use of the likelihood function in parameter estimation is easiest to understand in the case of discrete iid random variables.

### 1.2.1 Discrete IID Random Variables

Suppose each of the  $n$  random variables in the sample  $Y_1, \dots, Y_n$  have probability mass function  $f(y; \theta) = P_\theta(Y_1 = y)$ ,  $y = y_1, y_2, \dots$ . The likelihood is then defined as:

$$L(\theta|\mathbf{Y}) = \text{joint density of observed random variables}$$

In other words,



**Example (Fetal Lamb Movements):** Data on counts of movements in five-second intervals of one fetal lamb ( $n = 240$  intervals:)

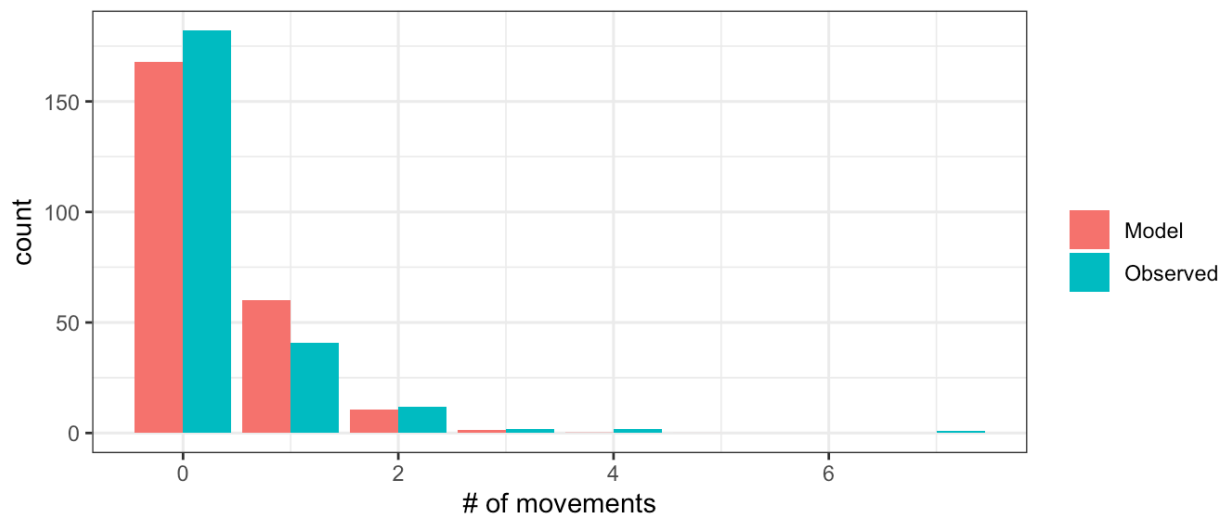
No. of Movements	0	1	2	3	4	5	6	7
Count	182	41	12	2	2	0	0	1

Assume a Poisson model:  $P(Y = y) = f_Y(y; \lambda) = \frac{\exp(-\lambda)\lambda^y}{y!}$ . Then the likelihood is

Equating the derivative of the loglikelihood with respect to  $\lambda$  to zero and solving results in the MLE

$$\hat{\lambda}_{\text{MLE}} =$$

This is the best we can do with this model. But is it good?



### 1.2.2 Multinomial Likelihoods

The multinomial distribution is a generalization of the binomial distribution where instead of 2 outcomes (success or failure), there are now  $k \geq 2$  outcomes.

The probability mass function is

For  $N_1, \dots, N_k$ ,  $N_i =$  the number of balls in  $i^{\text{th}}$  urn,

The maximum likelihood estimator of  $p_i$ :

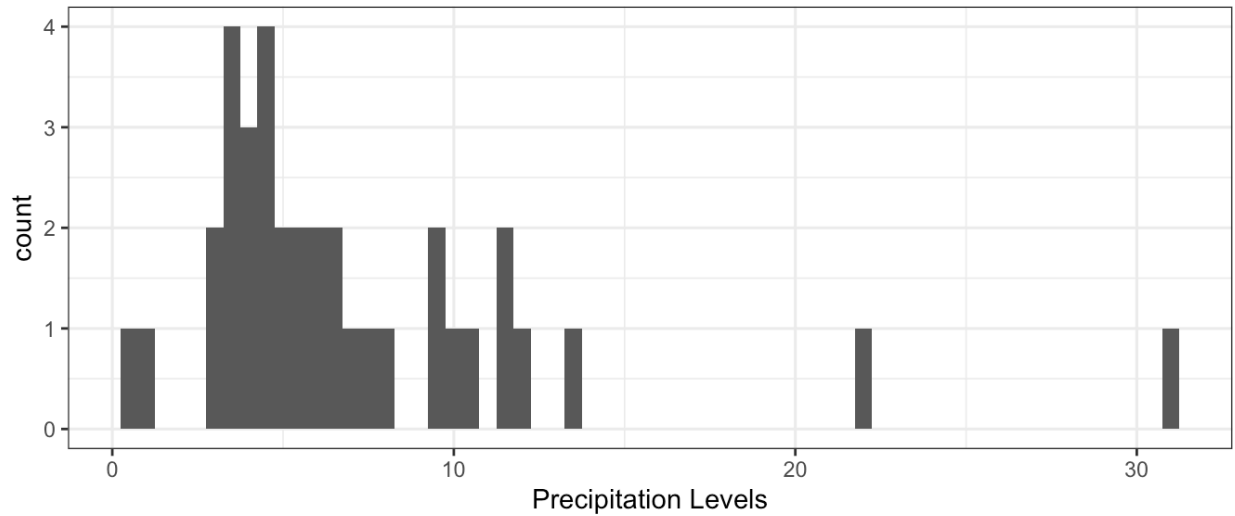
More interesting multinomial likelihoods arise when the  $p_i$  are modeled as a function of a lesser number of parameters  $\theta_1, \dots, \theta_m$ ,  $m < k - 1$ .

**Example (Capture-Recapture):** To estimate fish survival during a specific length of time (e.g., one month), a common approach is to use a removal design.

### 1.2.3 Continuous IID Random Variables

Recall: the likelihood is the joint density of data to be analyzed.

**Example (Hurricane Data):** For 36 hurricanes that had moved far inland on the East Coast of the US in 1900-1969, maximum 24-hour precipitation levels during the time they were over mountains.



We model the precipitation levels with a gamma distribution, which has density

$$f(y; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} \exp(-y/\beta), \quad y > 0, \alpha, \beta > 0.$$

This leads to the likelihood

Of course, this cannot be interpreted as a probability because

To get a probability, need to go from a density to a measure.

But it may be useful to think of the value of the likelihood as being proportional to a probability.

More formally, begin with the definition of a derivative

$$g'(x) = \lim_{h \rightarrow 0^+} \frac{g(x+h) - g(x-h)}{2h}.$$

Let  $F$  be the cumulative distribution function of a continuous random variable  $Y$ , then (if the derivative exists)

$$f(y) = \lim_{h \rightarrow 0^+} \frac{F(y+h) - F(y-h)}{2h} =$$

If we substitute this definition of a density into the definition of the likelihood

Compare this to the iid discrete case:

**Example (Hurricane Data, Cont'd):** Recall with a gamma model, the likelihood for this example is

$$L(\boldsymbol{\theta}|\mathbf{Y}) = \{\Gamma(\alpha)\}^{-n} \beta^{-n\alpha} \left\{ \prod Y_i \right\}^{\alpha-1} \exp\left(-\sum y_i/\beta\right),$$

and log-likelihood

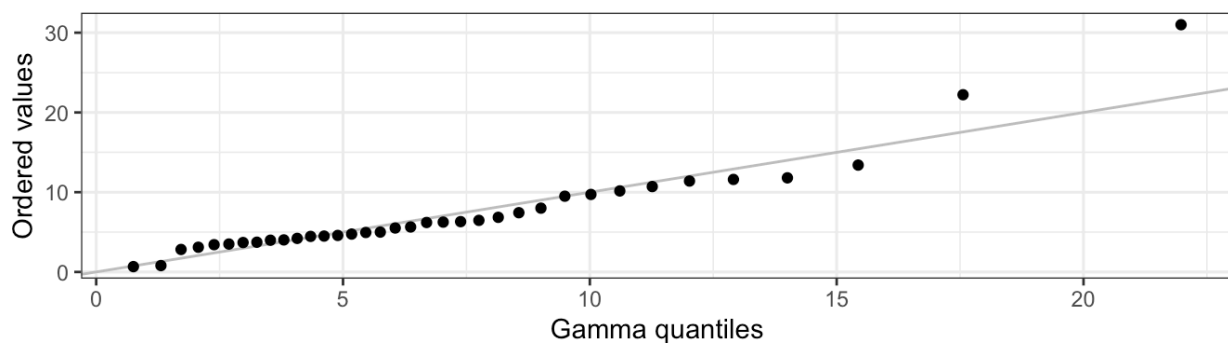
$$\ell(\boldsymbol{\theta}) =$$

```
## loglikelihood function
neg_gamma_loglik <- function(theta, data) {
  -sum(log(dgamma(data, theta[1], scale = theta[2])))
}

## maximize
mle <- nlm(neg_gamma_loglik, c(1.59, 4.458), data =
hurr_rain)
mle$estimate
```

```
## [1] 2.187214 3.331862
```

```
## Gamma QQ plot
data.frame(theoretical = qgamma(ppoints(hurr_rain),
mle$estimate[1], scale = mle$estimate[2]),
          actual = sort(hurr_rain)) |>
  ggplot() +
  geom_abline(aes(intercept = 0, slope = 1), colour =
"grey") +
  geom_point(aes(theoretical, actual)) +
  xlab("Gamma quantiles") + ylab("Ordered values")
```



### 1.2.4 Mixtures of Discrete and Continuous RVs

Some data  $Y$  often have a number of zeros and the amounts greater than zero are best modeled by a continuous distribution.

Ex:

In other words, they have positive probability of taking a value of exactly zero, but continuous distribution otherwise.

A sensible model would assume  $Y_i$  are iid with cdf

$$F_Y(y; p, \boldsymbol{\theta}) = \begin{cases} 0 & y = 0 \\ p & y = 0 \\ p + (1 - p)F_T(y; \boldsymbol{\theta}) & y > 0 \end{cases}$$

where  $0 < p \leq 1$  is  $P(Y = 0)$  and  $F_T(y; \boldsymbol{\theta})$  is a distribution function for a continuous positive random variable.

Another way to write this:

How to go from here to get a likelihood?



One approach: let  $n_0$  be the number of zeroes in the data and  $m = n - n_0$  be the number of non-zero  $Y_i$ . This leads to an intuitive way to construct the likelihood for iid  $Y_1, \dots, Y_n$  distributed according to the above distribution:

$$L(\boldsymbol{\theta}|\mathbf{Y}) = \lim_{h \rightarrow 0^+} \left( \frac{1}{2h} \right)^m \prod_{i=1}^m \{F_Y(Y_i + h; p, \boldsymbol{\theta}) - F_Y(Y_i - h; p, \boldsymbol{\theta})\}$$

Feels a little arbitrary in how we are defining different weights on our likelihood for discrete and continuous parts.

Turns out, it doesn't matter! (Need some STAT 630/720 to see why.)

**Definition (Absolute Continuity)** On  $(\mathbb{X}, \mathcal{M})$ , a finitely additive set function  $\phi$  is *absolutely continuous* with respect to a measure  $\mu$  if  $\phi(A) = 0$  for each  $A \in \mathcal{M}$  with  $\mu(A) = 0$ . We also say  $\phi$  is *dominated* by  $\mu$  and write  $\phi \ll \mu$ . If  $\nu$  and  $\mu$  are measures such that  $\nu \ll \mu$  and  $\mu \ll \nu$  then  $\mu$  and  $\nu$  are *equivalent*.

**Theorem (Lebesgue-Randon-Nikodym)** Assume that  $\phi$  is a  $\sigma$ -finite countably additive set function and  $\mu$  is a  $\sigma$ -finite measure. There exist unique  $\sigma$ -finite countably additive set functions  $\phi_s$  and  $\phi_{ac}$  such that  $\phi = \phi_{ac} + \phi_s$ ,  $\phi_{ac} \ll \mu$ ,  $\phi_s$  and  $\mu$  are mutually singular and there exists a measurable extended real valued function  $f$  such that

$$\phi_{ac}(A) = \int_A f d\mu, \quad \text{for all } A \in \mathcal{M}.$$

If  $g$  is another such function, then  $f = g$  a.e. wrt  $\mu$ . If  $\phi \ll \mu$  then  $\phi(A) = \int_A f d\mu$  for all  $A \in \mathcal{M}$ .

**Definition (Radon-Nikodym Derivative)**  $\phi = \phi_{ac} + \phi_s$  is called the *Lebesgue decomposition*. If  $\phi \ll \mu$ , then the density function  $f$  is called the *Radon-Nikodym derivative* of  $\phi$  wrt  $\mu$ .

So what?



### 1.2.5 Proportional Likelihoods

Likelihoods are equivalent for point estimation as long as they are proportional and the constant of proportionality does not depend on unknown parameters.

Why?

Consider if  $Y_i, i = 1, \dots, n$  are iid continuous with density  $f_Y(y; \boldsymbol{\theta})$  and  $X_i = g(Y_i)$  where  $g$  is increasing and continuously differentiable. Because  $g$  is one-to-one, we can construct  $Y_i$  from  $X_i$  and vice versa.

More formally, the density of  $X_i$  is  $f_X(x; \boldsymbol{\theta}) = f_Y(h(x); \boldsymbol{\theta})h'(x)$ , where  $h = g^{-1}$ , and

$$L(\boldsymbol{\theta}|\mathbf{X}) =$$

**Example (Likelihood Principle):** Consider data from two different sampling plans:

1. A binomial experiment with  $n = 12$ . Let  $Y_i = 1$  if  $i^{\text{th}}$  trial is a success and 0 otherwise.

$$L_1(p|\mathbf{Y}) = \binom{12}{S} p^S (1-p)^{12-S}, \text{ where } S = \sum_{i=1}^n Y_i$$

2. A negative binomial experiment, i.e. run the experiment until three zeroes are obtained.

$$L_2(p|\mathbf{Y}) = \binom{S+2}{S} p^S (1-p)^3.$$

The ratio of these likelihoods is

$$\frac{L_1(p|\mathbf{Y})}{L_2(p|\mathbf{Y})} =$$

Suppose  $S = 9$ . Is all inference equivalent for these likelihoods? Debatable.

**The likelihood principle** states all the information about  $\theta$  from an experiment is contained in the actual observation  $\mathbf{y}$ . Two likelihood functions for  $\theta$  (from the same or different experiments) contain the same information about  $\theta$  if they are proportional.

### 1.2.6 Empirical Distribution Function as MLE

Recall the empirical cdf:

Suppose  $y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(n)}$  are the order statistics of an iid sample from an unknown distribution function  $F_Y$ . Our goal is to estimate  $F_Y$ .

$$\hat{F}_Y(y) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(y \geq y_{(i)})$$

Is this a “good” estimator of  $F_Y$ ?

Yes, because it's MLE.

Suppose  $Y_1, \dots, Y_n$  are iid with distribution function  $F(y)$ . Here  $F(y)$  is the unknown parameter.

An approximate likelihood for  $F$  is

$$L_h(F|\mathbf{Y}) = \prod_{i=1}^n \{F(Y_i + h) - F(Y_i - h)\}$$

### 1.2.7 Censored Data

Censored data occur when the value is only partially known. This is different from *truncation*, in which the data does not include any values below (or above) a certain limit.

For example, we might sample only households that have an income above a limit,  $L_0$ . If all incomes have distribution  $F(x; \theta)$ , then for  $y > L_0$ ,

$$P(Y_1 \leq y | Y_1 > L_0) =$$

The likelihood is then

#### 1.2.7.1 Type I Censoring

Suppose a random variable  $X$  is normally distributed with mean  $\mu$  and variance  $\sigma^2$ , but whenever  $X \leq 0$ , all we observe is that it is less than or equal to 0. If the sample is set to 0 in the censored cases, then define

$$Y = \begin{cases} 0 & \text{if } X \leq 0 \\ X & \text{if } X > 0. \end{cases}$$

The distribution function of  $Y$  is



Suppose we have a sample  $Y_1, \dots, Y_n$  and let  $n_0$  be the number of sample values that are 0. Then  $m = n - n_0$  and

We might have censoring on the left at  $L_0$  and censoring on the right at  $R_0$ , but observe all values of  $X$  between  $L_0$  and  $R_0$ . Suppose  $X$  has density  $f(x; \boldsymbol{\theta})$  and distribution function  $F(x; \boldsymbol{\theta})$  and

$$Y_i = \begin{cases} L_0 & \text{if } X_i \leq L_0 \\ X_i & \text{if } L_0 < X_i < R_0 \\ R_0 & \text{if } X_i \geq R_0 \end{cases}$$

If we let  $n_L$  and  $n_R$  be the number of  $X_i$  values  $\leq L_0$  and  $\geq R_0$  then the likelihood of the observed data  $Y_1, \dots, Y_n$  is

We could also let each  $X_i$  be subject to its own censoring values  $L_i$  and  $R_i$ . For the special case of right censoring, define  $Y_i = \min(X_i, R_i)$ . In addition, define  $\delta_i = \mathbb{I}(X_i \leq R_i)$ . Then the likelihood can be written as

**Example (Equipment failure times):** Pieces of equipment are regularly checked for failure (but started at different times). By a fixed date (when the study ended), three of the items had not failed and therefore were censored.

y	2	72	51	50	33	27	14	24	4	21
delta	1	0	1	0	1	1	1	1	1	0

Suppose failure times follow an exponential distribution  $F(x; \sigma) = 1 - \exp(-x/\sigma)$ ,  $x \geq 0$ . Then

$$L(\sigma|\mathbf{Y}) =$$

### 1.2.7.2 Random Censoring

So far we have considered censoring times to be fixed. This is not required.

This leads to random censoring times, e.g.  $R_i$ , where we assume that the censoring times are independent of  $X_1, \dots, X_n$  and iid with distribution function  $G(t)$  and density  $g(t)$ .

Let's consider the contributions to the likelihood:

which results in

$$L(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\delta}) =$$

## 1.3 Likelihoods for Regression Models

We will start with linear regression and then talk about more general models.

### 1.3.1 Linear Model

Consider the familiar linear model

$$Y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \dots, n,$$

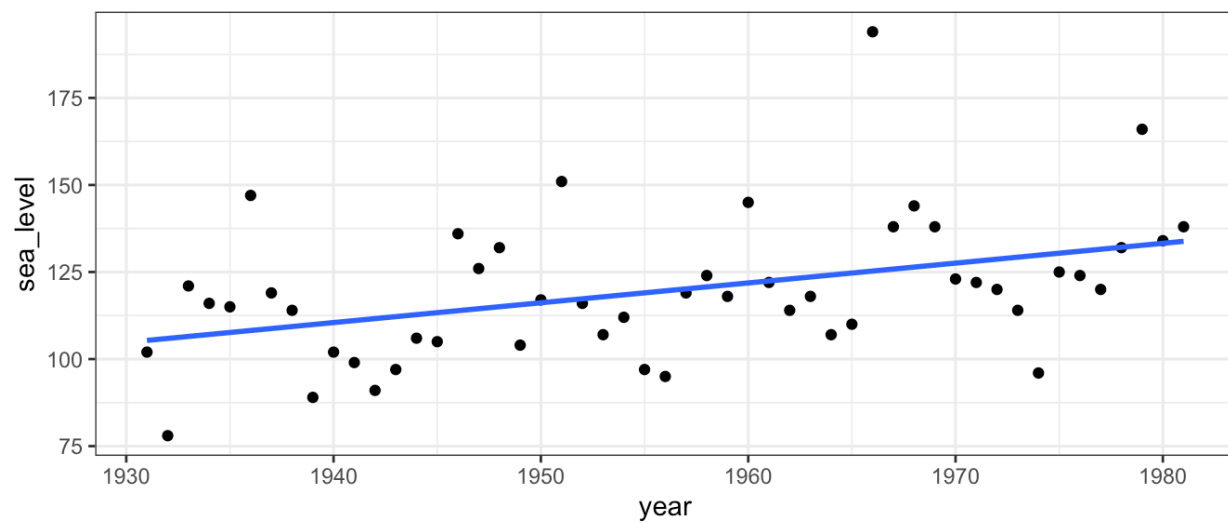
where  $\mathbf{x}_1, \dots, \mathbf{x}_n$  are known nonrandom vectors.

For likelihood-based estimation,

$$L(\boldsymbol{\beta}, \sigma | \{Y_i, \mathbf{x}_i\}_{i=1}^n) =$$

What do you do when  $\epsilon_i$  are not Gaussian?

**Example (Venice sea levels):** The annual maximum sea levels in Venice for 1931–1981 are :



### 1.3.2 Additive Errors Nonlinear Model

### 1.3.3 Generalized Linear Models

Imagine an experiment where individual mosquitos are given some dosage of pesticide. The response is whether the mosquito lives or dies. The data might look something like:

**Goal:** Model the relationship between the predictor and response.

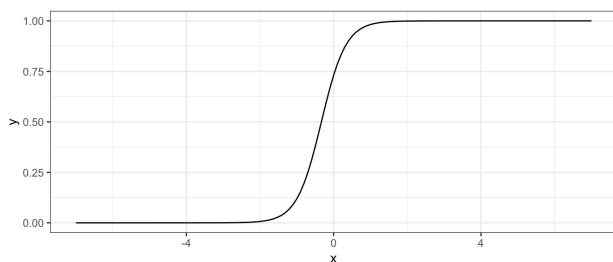
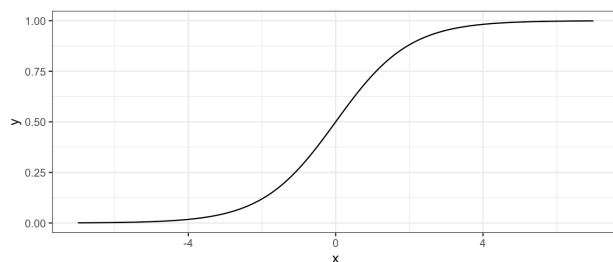
**Question:** What would a curve of best fit look like?

**Refined Goal:**

Let's build a sensible model.

**Step 1:** Find a function that behaves the way we want.

```
# understanding the logistic function  
# first, theta just equals x  
x <- seq(-7, 7, .1)  
theta <- x  
y <- exp(theta)/(1 + exp(theta))  
ggplot() + geom_line(aes(x, y))  
  
# now, let theta be a linear function of x  
theta <- 1 + 3*x  
y <- exp(theta)/(1 + exp(theta))  
ggplot() + geom_line(aes(x, y))
```



**Step 2:** Build a stochastic mechanism to relate to a binary response.

**Step 3:** Put Step 1 and Step 2 together.

Fitting our model: Does OLS make sense?



Consider the likelihood contribution.

$$L_i(\mathbf{p}_i | Y_i) =$$

So the log-likelihood contribution is

$$\ell_i(\mathbf{p}_i) =$$

Recall, we said  $p_i = \frac{\exp(\theta_i)}{1 + \exp(\theta_i)}$  was sensible.

Which gives us,

$$\ell_i(\theta_i) =$$

So the log-likelihood is

$$\ell(\theta_1, \dots, \theta_n) =$$

To optimize?

```
## data on credit default
data("Default", package = "ISLR")
head(Default)
```

```
## default student balance income
## 1 No No 729.5265 44361.625
## 2 No Yes 817.1804 12106.135
## 3 No No 1073.5492 31767.139
## 4 No No 529.2506 35704.494
## 5 No No 785.6559 38463.496
## 6 No Yes 919.5885 7491.559
```

```
## fit model with ML
m0 <- glm(default ~ balance, data = Default, family =
binomial)
tidy(m0) |> kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	-10.6513306	0.3611574	-29.49221	0
balance	0.0054989	0.0002204	24.95309	0

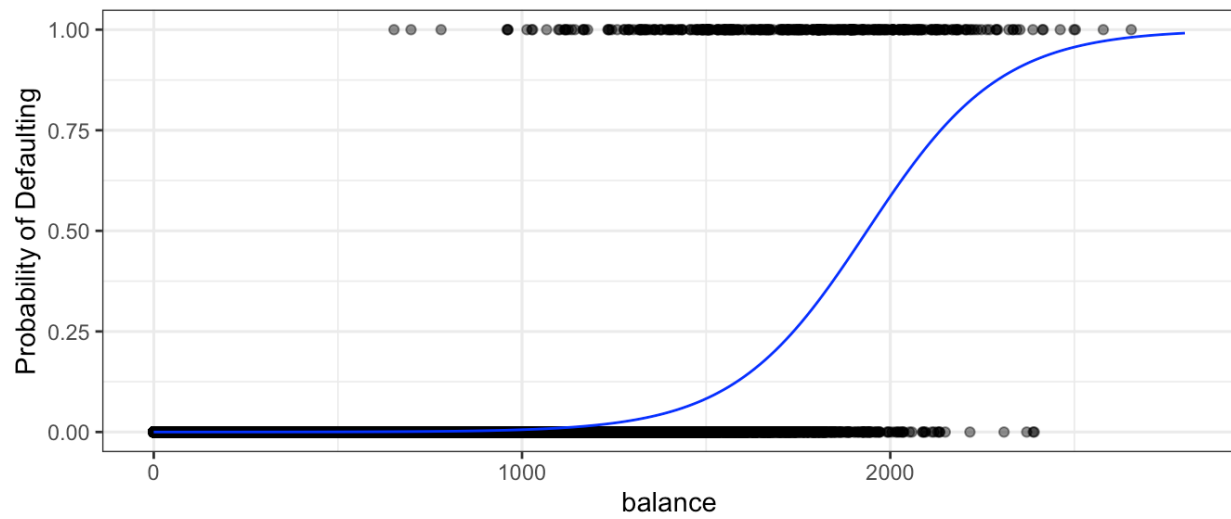
```
glance(m0) |> kable()
```

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	nobs
2920.65	9999	-798.2258	1600.452	1614.872	1596.452	9998	10000

```
## plot the curve
x_new <- seq(0, 2800, length.out = 200)
theta <- m0$coefficients[1] + m0$coefficients[2]*x_new
p_hat <- exp(theta)/(1 + exp(theta))

ggplot() +
  geom_point(aes(balance, as.numeric(default) - 1), alpha =
0.5, data = Default) +
```

```
geom_line(aes(x_new, p_hat), colour = "blue") +  
ylab("Probability of Defaulting")
```



In general, a GLM is three pieces:

1. The random component

2. The systemic component

3. A linear predictor

Remarks:

**Example (Poisson regression):**

Consider a general family of distributions:

$$\log f(y_i; \theta_i, \phi) = \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi).$$

**Example (Normal model):**

We can learn something about this distribution by considering its mean and variance. Because we don't have an explicit form of the density, we rely on two facts:

$$1. \mathbf{E} \left[ \frac{\partial \log f(Y_i; \theta_i, \phi)}{\partial \theta_i} \right] = 0.$$

$$2. \mathbf{E} \left[ \frac{\partial^2 \log f(Y_i; \theta_i, \phi)}{\partial \theta_i^2} \right] + \mathbf{E} \left[ \left( \frac{\partial \log f(Y_i; \theta_i, \phi)}{\partial \theta_i} \right)^2 \right] = 0.$$

$$\text{For } \log f(y_i; \theta_i, \phi) = \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi),$$

**Example (Bernoulli model):**

$$f(y_i; p_i) = p_i^{y_i} (1 - p_i)^{1-y_i}$$

Finally, back to modelling. Our **goal** is to build a relationship between the mean of  $Y_i$  and covariates  $\boldsymbol{x}_i$ .

**Example (Bernoulli model, cont'd):**



## 1.4 Marginal and Conditional Likelihoods

Consider a model which has  $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ , where  $\boldsymbol{\theta}_1$  are the parameters of interest and  $\boldsymbol{\theta}_2$  are nuisance parameters.

One way to improve estimation for  $\boldsymbol{\theta}_1$  is to find a one-to-one transformation of the data  $\mathbf{Y}$  to  $(\mathbf{V}, \mathbf{W})$  such that either

The key feature is that one component of each contains only the parameter of interest.

**Example (Neyman-Scott problem):** Let  $Y_{ij}, i = 1, \dots, n, j = 1, 2$  be independent normal random variables with possible different means  $\mu_i$  but the same variance  $\sigma^2$ .

Our goal is to estimate  $\sigma^2$ . Should we be able to?

Following the usual arguments,

$$\hat{\mu}_{i,\text{MLE}} = \frac{Y_{i1} + Y_{i2}}{2}$$
$$\hat{\sigma}_{\text{MLE}}^2 = \frac{1}{2n} \sum_{i=1}^n \sum_{j=1}^2 (Y_{ij} - \hat{\mu}_{i,\text{MLE}})^2$$

$$E[\hat{\sigma}_{\text{MLE}}^2] =$$

A reworking of the data seems more promising. Let,

$$V_i = \frac{Y_{i1} - Y_{i2}}{\sqrt{2}} \quad \text{and} \quad W_i = \frac{Y_{i1} + Y_{i2}}{\sqrt{2}}$$

For conditional likelihoods, we can often exploit the existence of sufficient statistics for the nuisance parameters under the assumption that the parameter of interest is known.

**Example (Exponential Families):** The structure of exponential families is such that it is often possible to exploit their properties to eliminate nuisance parameters. Let  $Y$  have a density of the form

$$f(\mathbf{y}; \boldsymbol{\eta}) = h(\mathbf{y}) \exp \left\{ \sum_{i=1}^s \eta_i T_i(\mathbf{y}) - A(\boldsymbol{\eta}) \right\},$$

then

Thus, exponential families often provide an automatic procedure for finding  $\mathbf{W}$  and  $\mathbf{V}$ .

**Example (Logistic Regression):** For binary  $Y_i$ , the standard logistics regression model is

$$P(Y_i = 1) = p_i(\mathbf{x}_i, \boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})}$$

and the likelihood is

$$L(\boldsymbol{\beta} | \mathbf{Y}, \mathbf{X}) =$$

## 1.5 The Maximum Likelihood Estimator and the Information Matrix

We have now talked about how to construct likelihoods in a variety of settings, now we can use those constructions to formalize how we make inferences about model parameters.

Recall the score function

$$S(\mathbf{Y}, \boldsymbol{\theta}) =$$

Generally, the maximum likelihood estimator  $\hat{\boldsymbol{\theta}}_{\text{MLE}}$  is the value of  $\boldsymbol{\theta}$  where the maximum (over the parameter space  $\Theta$ ) of  $L(\boldsymbol{\theta}|\mathbf{Y})$  is attained.

Under the assumption that the log-likelihood is continuously differentiable, then

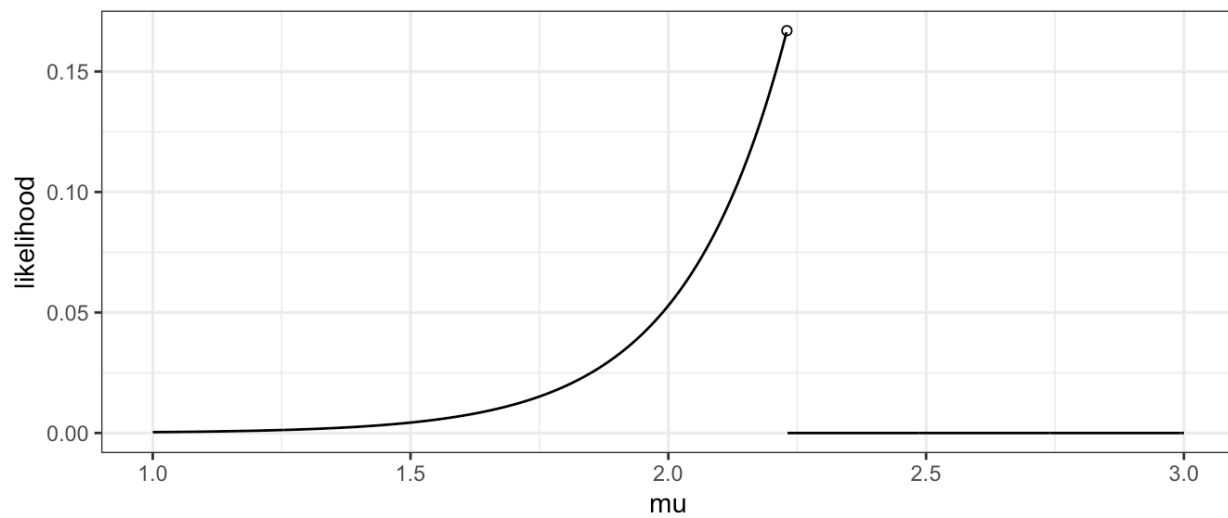
But not always (?!).

**Example (Exponential threshold model):** Suppose that  $Y_1, \dots, Y_n$  are iid from the exponential distribution with a threshold parameter  $\mu$ ,

$$f(y; \mu) = \begin{cases} \exp\{-(y - \mu)\} & \mu < y < \infty \\ 0 & \text{otherwise,} \end{cases}$$

for  $-\infty < \mu < \infty$ .

Consider the artificial data set  $\mathbf{y} = [2.47, 2.35, 2.23, 3.53, 2.36]$ .



### 1.5.1 The Fisher Information Matrix

The Fisher information matrix  $I(\boldsymbol{\theta})$  is defined as the  $b \times b$  matrix where

$$I_{ij}(\boldsymbol{\theta}) =$$

In matrix form,

$$I(\boldsymbol{\theta}) =$$



Fisher information facts:

1. The Fisher information matrix is the variance of the score contribution.

2. If regularity conditions are met,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}) \xrightarrow{d} \text{N}_b(0, I(\boldsymbol{\theta})^{-1}).$$

3. If  $b = 1$ , then any unbiased estimator must have variance greater than or equal to  $\{nI(\boldsymbol{\theta})\}^{-1}$

4. The information matrix is related to the curvature of the log-likelihood contribution.

### 1.5.2 Observed Information

The information matrix is not random, but it is also not observable from the data.

Let  $Y_1, \dots, Y_n$  be iid with density  $f_Y(y_i; \boldsymbol{\theta})$ . The log likelihood is defined as

taking two derivatives and dividing by  $n$  results in

**Definition:** The matrix  $n\bar{I}(Y; \hat{\boldsymbol{\theta}}_{\text{MLE}})$  is called the sample information matrix, or the *observed information matrix*.

Why use  $I(\boldsymbol{\theta}) = \mathbf{E} \left[ -\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \log f(Y_1; \boldsymbol{\theta}) \right]$  as the basis for an estimator, rather than  $I(\boldsymbol{\theta}) = \mathbf{E} \left[ \left\{ \frac{\partial}{\partial \boldsymbol{\theta}^\top} \log f(Y_1; \boldsymbol{\theta}) \right\} \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \log f(Y_1; \boldsymbol{\theta}) \right\} \right]$ ?

Now let's prove the asymptotic normality of the MLE (in the scalar case).