University of Edinburgh

School of Mathematics

Statistical Methodology

Serveh Sharifi

Semester 1, 2019–2020

The concepts of likelihood and regression are central to statistical inference and statistical modelling. This course develops likelihood methods for single parameter and multiparameter problems—including likelihood-based confidence regions and likelihood ratio—and offers an overview on theory and applications of mean regression models.

Course outline

1. Introduction, examples of likelihood and likelihood principle.
2. Likelihood-based inference.
3. Maximum likelihood estimation.
4. Regression.

Textbook

Wood, S. N., *Core Statistics*, Cambridge University Press, 2015.

Supplementary Resources

1. Azzalini, A., *Statistical Inference Based on the Likelihood*, Chapman & Hall, 1996.
   Good coverage of the lecture material.

2. Held, L. & Bové D. S., *Applied Statistical Inference: Likelihood and Bayes*, Springer, 2014.
   Provides an integrated overview on statistical inference.

3. Christensen, R. et al., *Bayesian Ideas and Data Analysis, An Introduction for Scientists and Statisticians*, Chapman & Hall, 2011.
   For further reading on Bayesian methods.

4. Crawley, M. J. *The R Book*, Wiley, 2012.
   Comprehensive guide to R. Introduces the language, and then covers a wide range of statistical techniques.

Contents

[1] From Christensen, R. et al., Bayesian Ideas and Data Analysis, An Introduction for Scientists and Statisticians, Chapman and Hall, 2011.

1 Introduction

1.1 Likelihood function

Consider a random sample $X_1, \ldots, X_n$ from a distribution with probability density function (p.d.f.) (or probability function, p.f., in the discrete case) $f(x;\theta)$, where θ is an unknown parameter, i.e. the random variables $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) in $f(x;\theta)$. For observations $x_1, \ldots, x_n$, we define the **likelihood function** by

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

$L(\theta)$ may be thought of as a function of variable θ with the observations $x_1, \ldots, x_n$ fixed. It represents the relative plausibilities of different θ values in the light of the data. In some problems the above definition of likelihood needs to be extended (to correctly account for the probabilities of events associated with the observed data), but for most situations the definition is adequate.

Example:

- **Binomial distribution**: Observe 3 heads in 5 throws of a (possibly biased) coin. Model by a binomial $B(5,\theta)$ distribution, where $\theta = \Pr(\text{head})$. Since $f(x;\theta) = \binom{5}{x}\theta^x(1-\theta)^{5-x}$, the likelihood is given by

$$L(\theta) = f(3;\theta) = \binom{5}{3}\theta^3(1-\theta)^2.$$

  We can use the likelihood to assess the relative plausibilities of different values of θ in the interval [0,1], given the event (3 heads) has occurred.

  We can then choose θ to ensure that what we observed is most plausible. This gives the idea of **maximum likelihood estimation**. A θ corresponding to the highest point on the likelihood is called a **maximum likelihood estimate** (MLE), and we denote such a value of θ by $\widehat{\theta}$.

  Differentiating the likelihood with respect to θ gives

$$\frac{dL(\theta)}{d\theta} = \binom{5}{3}3\theta^2(1-\theta)^2 - 2\binom{5}{3}\theta^3(1-\theta).$$

  Setting this equal to zero and solving for θ gives the MLE of θ. This leads to solving the equation

$$3(1-\theta) - 2\theta = 0.$$

  Thus, the MLE of θ is

$$\widehat{\theta} = \frac{3}{5}.$$

To plot this likelihood function in `R` (cf Figure 1) use the following code:

```r
## R code for producing Fig. 1
L <- function(theta)
  choose(5, 3) * theta^3 * (1 - theta)^2
curve(L, xlab = expression(theta), ylab = expression(L(theta)))
```

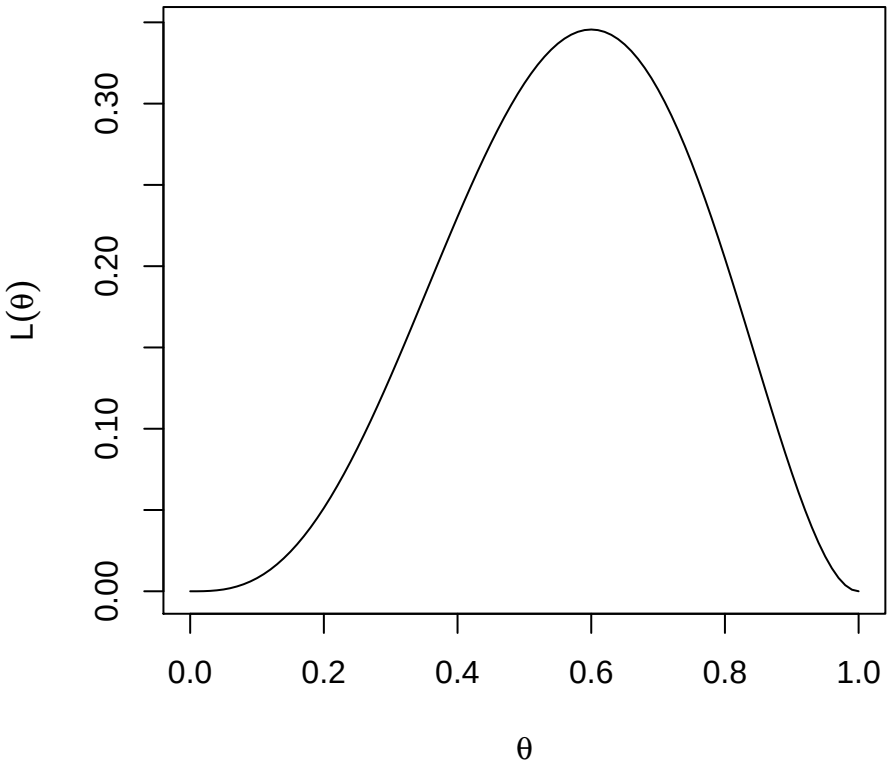

Figure 1: Likelihood function. Binomial: 3 heads in 5 trails.

In the `R` console, "?`command`" can be used to learn more about a command. (Example, try `?curve`.) To install `R` see

`https://cran.r-project.org`

From both theoretical and computational points of view it is usually more convenient to work with the (natural) logarithm of the likelihood function rather than the likelihood function. Note that the MLE can be determined by maximising the log-likelihood function. We will also see in later sections that it is the log-likelihood (and its derivatives) on which we base our inferences.

1.2 Examples

- **Poisson distribution**: Consider modelling the numbers of aphids on n leaves, and suppose that a Poisson distribution with mean parameter λ provides an appropriate

model. That is, if X is a random variable denoting the number of aphids on a leaf, then we have

$$f(x;\lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}, \quad x = 0, 1, 2, \ldots$$

Here λ is the parameter of interest, and for independent observations $x_1, \ldots, x_n$ we can determine the likelihood by taking the product of the $\lambda^{x_i} \exp(-\lambda)/x_i!$ over $i = 1, \ldots, n$. This leads to

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} \exp(-\lambda)}{x_i!} = \frac{\lambda^{\sum_{i=1}^n x_i} \exp(-n\lambda)}{\prod_{i=1}^n x_i!}, \quad \lambda > 0.$$

Taking logs (to base e) gives the log-likelihood as

$$l(\lambda) = \left(\sum_{i=1}^{n} x_i \right) \log \lambda - n\lambda + \text{const}, \quad \lambda > 0.$$

Then setting the derivative equal to zero determines the MLE

$$\frac{dl(\lambda)}{d\lambda} = \frac{\sum_{i=1}^n x_i}{\lambda} - n = 0,$$

which gives the MLE as $\widehat{\lambda} = 1/n \sum_{i=1}^n x_i = \bar{x}$, the sample mean.

Suppose we have the following sample of $x_1 = 7, x_2 = 9, x_3 = 14$ aphids on three leaves, then the MLE would be

$$\widehat{\lambda} = \frac{30}{3} = 10.$$

To plot this log-likelihood function in R (cf Figure 2) use the following code:

```
## R code for producing Fig. 2
x <- c(7, 9, 14)
n <- length(x)
const <- -sum(log(factorial(x)))
l <- function(lambda)
    sum(x) * log(lambda) - n * lambda + const
curve(l, from = 6, to = 15, xlab = expression(lambda),
      ylab = expression(l(lambda)))
```

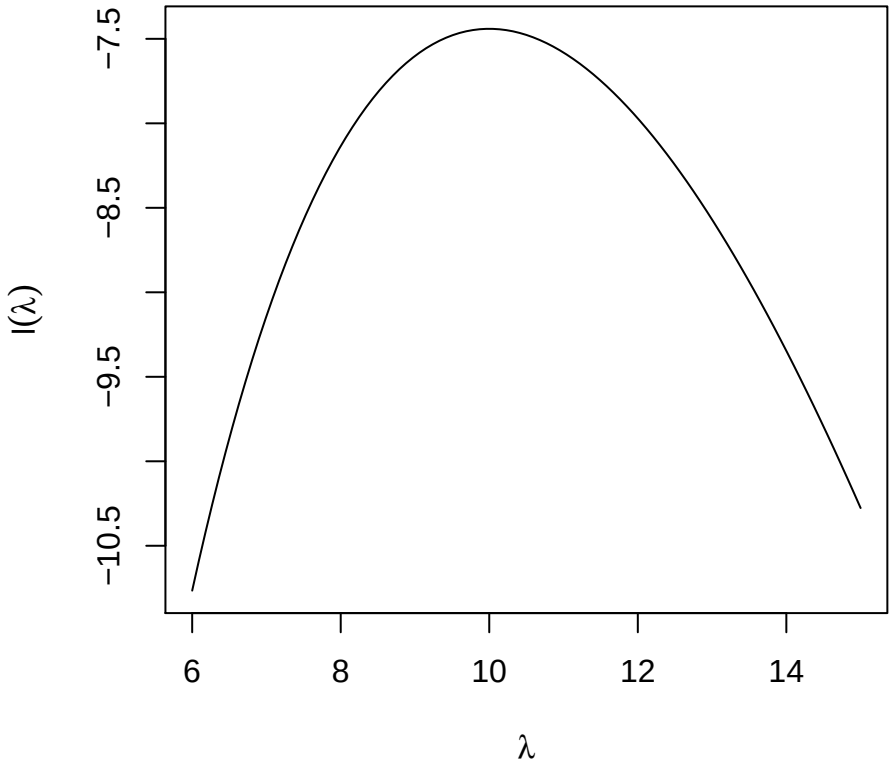

Figure 2: Log-likelihood function. Poisson random sample: 7, 9, 14.

- **Uniform distribution**: A random sample $X_1, \ldots, X_n$ from a uniform distribution on $(0, \theta]$ with p.d.f. given by

$$f(x;\theta) = \begin{cases} \frac{1}{\theta}, & \text{if } 0 < x \leq \theta, \\ 0, & \text{otherwise.} \end{cases}$$

For independent observations $x_1, \ldots, x_n$ the likelihood is

$$L(\theta) = \begin{cases} \frac{1}{\theta^n}, & \text{if } 0 < x_1, \ldots, x_n \leq \theta, \\ 0, & \text{otherwise,} \end{cases}$$

which can be written as

$$L(\theta; x) = \begin{cases} 0, & \text{if } \theta < x_{(n)}, \\ \frac{1}{\theta^n}, & \text{if } \theta \geq x_{(n)}, \end{cases}$$

where $x_{(n)}$ is the largest order statistic, i.e. the largest value in the data set.

In this example we cannot differentiate the (log) likelihood to determine the MLE. (Try it.) However, by plotting the likelihood it is clear that the MLE of $\widehat{\theta}$ is given by the largest observation, i.e. $x_{(n)}$, and we can also see that the likelihood is not differentiable at $\theta = x_{(n)}$; see Figure 3.

To plot this likelihood function in `R` (cf Figure 3) use the following code:

```r
## R code for producing Fig. 3
x <- c(1.3, 2.4, 3.2)
xn <- max(x)
n <- length(x)
L <- function(theta)
    ifelse(theta < xn, 0, 1 / theta^n)
curve(L, from = 0, to = 10, xlab = expression(theta),
      ylab = expression(L(theta)))
points(x, rep(0, 3))
```

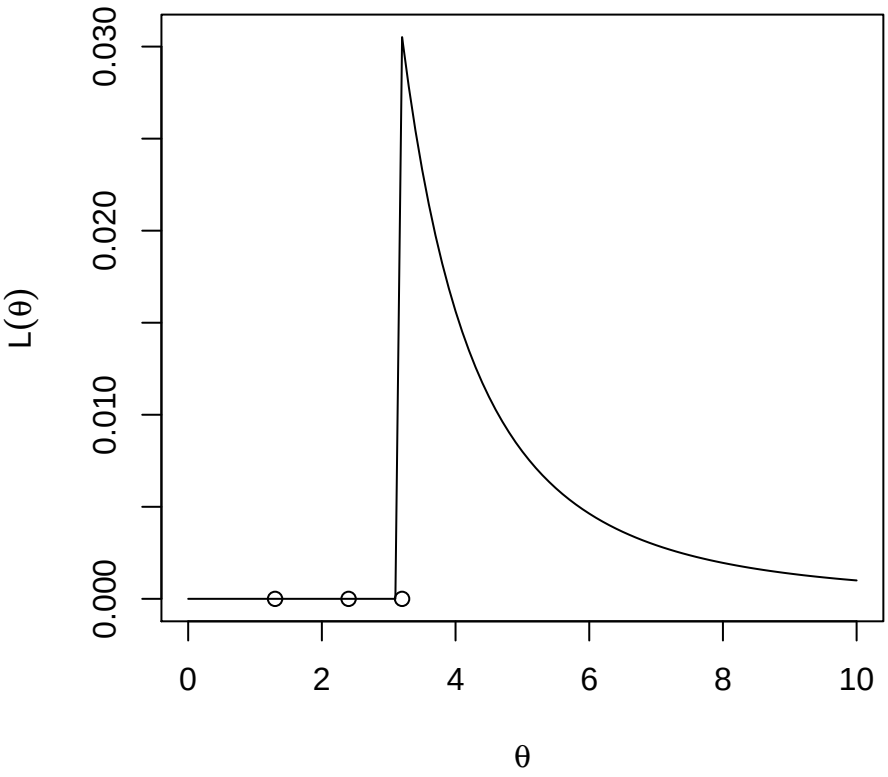

Figure 3: Likelihood function. Uniform random sample: 1.3, 2.4, 3.2.

- **Normal distribution, I**: A random sample $X_1, \ldots, X_n$ from a normal $N(\theta, \sigma^2)$ distribution with p.d.f.

$$f(x;\theta) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\} \qquad (\sigma^2 \text{ known}).$$

For independent observations $x_1, \ldots, x_n$, the likelihood and log-likelihood are given by

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

and

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

Setting the derivative of the log likelihood (with respect to θ) equal to zero gives an equation that determines the MLE

$$\frac{\partial l(\theta)}{\partial\theta} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\theta) = 0.$$

This gives the MLE of θ as $\widehat{\theta} = 1/n\sum_{i=1}^{n} x_i = \bar{x}$, the sample mean.

- **Normal distribution, II**: As previous $N(\theta,\sigma^2)$ example but suppose now that σ is also to be estimated from the observations $x_1,\ldots,x_n$.

  Setting the derivative of the log likelihood (with respect to σ this time) equal to zero gives

$$\frac{\partial l(\theta,\sigma)}{\partial\sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\theta)^2 = 0.$$

  Replacing θ by its MLE, $\widehat{\theta} = \bar{x}$ (obtained in previous example), gives the MLE of σ as $\widehat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\widehat{\theta})^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}$.

- **Gamma distribution**: A random sample $X_1, \ldots, X_n$ from a gamma distribution with p.d.f.

$$f(x;\lambda,r) = \frac{\lambda^r x^{r-1}e^{-\lambda x}}{\Gamma(r)} \qquad (\lambda \text{ known}).$$

  For independent observations $x_1,\ldots,x_n$, the likelihood for r is

$$L(r) = \frac{\lambda^{nr}(x_1\ldots x_n)^{r-1}e^{-\lambda\sum_{i=1}^{n}x_i}}{[\Gamma(r)]^n},$$

  and the log-likelihood is

$$l(r) = nr\log\lambda + (r-1)\sum\log x_i - \lambda\sum_{i=1}^{n}x_i - n\log\Gamma(r).$$

  Then, the derivative of $l(r)$ is given by

$$\frac{\partial l(r)}{\partial r} = n\log\lambda + \sum\log x_i - n\frac{\Gamma'(r)}{\Gamma(r)}.$$

  Iterative methods would be needed to solve $\frac{\partial l(r)}{\partial r} = 0$ for r to obtain the MLE. [Note that this involves using the digamma function.]

Likelihood principle:

For a statistical model $f(x;\theta)$ two data sets x and y such that

$$L(\theta;x) \propto L(\theta;y)$$

must lead to the same inferential conclusions.

e.g. Five trials of a coin toss giving a sample HHHTT should lead to the same conclusions as a similar experiment which produced a sample TTHHH. In both cases the likelihood is proportional to $\theta^3(1-\theta)^2$, if $\Pr(\text{Head}) = \theta$.

Strong likelihood principle:

Given an observation x from a statistical model $f(x;\theta)$ and an observation z from a statistical model $g(z;\theta)$ such that

$$L_f(\theta;x) \propto L_g(\theta;z)$$

they must lead to the same inferential conclusions.

e.g. Five Bernoulli [model $f(x;\theta)$] trials of a coin toss giving a sample HHHTT should lead to the same conclusions as an experiment which resulted in a binomial [model $g(z;\theta)$] realization of 3 heads out of 5 trials. Again, in both cases the likelihood is proportional to $\theta^3(1-\theta)^2$.

1.3 Approximate (asymptotic) variance of the MLE

For **large samples** the (asymptotic) variance of the maximum likelihood estimator $\widehat{\theta}$ is given by

$$\text{var}(\widehat{\theta}) \approx I_\theta^{-1},$$

(under very general conditions), where

$$I_\theta = \text{E}\left(\left[\frac{dl}{d\theta}\right]^2\right) = -\text{E}\left(\frac{d^2l}{d\theta^2}\right).$$

I_θ is called the Fisher information, and $\frac{dl}{d\theta}$ is called the score function. The score is also denoted as $U(\theta)$ and $l'(\theta)$, and the Fisher information can be written as $-\text{E}(U')$ or $\text{E}(U^2)$. When determining I_θ it is usually easier to use $-\text{E}(U')$ rather than $\text{E}(U^2)$. I_θ may be estimated by $I_{\widehat{\theta}}$ in the expression for $\text{var}(\widehat{\theta})$, and this may be denoted by

$$\widehat{\text{var}}(\widehat{\theta}) \approx I_{\widehat{\theta}}^{-1}.$$

Example:

- For a Poisson (λ) sample $x_1, \ldots, x_n$, the derivatives of the log-likelihood $l(\lambda)$ are given by
$$\frac{dl}{d\lambda} = \frac{\sum_{i=1}^n x_i}{\lambda} - n \qquad \text{and} \qquad \frac{d^2l}{d\lambda^2} = -\frac{\sum_{i=1}^n x_i}{\lambda^2}.$$
Thus, since $\text{E}(X_i) = \lambda$ for each $i = 1, \ldots, n$, we have the Fisher information
$$I_\lambda = -\text{E}\left(\frac{d^2l}{d\lambda^2}\right) = \frac{n}{\lambda},$$

and the (large sample) asymptotic variance of the MLE $\widehat{\lambda} = \bar{X}$ is

$$\text{var}(\widehat{\lambda}) \approx I_{\lambda}^{-1} = \frac{\lambda}{n}.$$

[In fact, in this case, the variance is also correct for small samples as $\text{var}(\bar{X}) = \frac{\lambda}{n}$.] An estimate of the asymptotic variance may be obtained by substituting $\widehat{\lambda}$ for λ.

The **estimated standard error** of the maximum likelihood estimator, $\widehat{\theta}$, of θ is given by

$$ESE(\widehat{\theta}) \approx \sqrt{I_{\widehat{\theta}}^{-1}}.$$

Example:

- For a Poisson (λ) model, the estimated standard error of the maximum likelihood estimator, $\widehat{\lambda}$, of λ is

$$ESE(\widehat{\lambda}) = \sqrt{I_{\widehat{\lambda}}^{-1}} = \sqrt{\frac{\widehat{\lambda}}{n}},$$

and, as $\widehat{\lambda} = \bar{x}$, we have

$$ESE(\widehat{\lambda}) = \sqrt{\frac{\bar{x}}{n}}.$$

2 Likelihood-based inference

This section considers methods of statistical inference based on the likelihood function, or to be more precise the log likelihood function. These include **hypothesis tests** and **confidence regions**. Although these are closely related, confidence regions may be considered to be more informative as they provide a measure of the precision with which inferences can be made.

2.1 Likelihood and MLE as random quantities

The properties of the likelihood function and the maximum likelihood estimator may be obtained by treating them as random quantities, based on a random sample $Y_1, \ldots, Y_n$.

Example:

- $Y \sim$ Bernoulli(θ). $y_1, \ldots, y_{10}$ consist of 3 observations $y = 1$ and 7 observations $y = 0$. Then $L(\theta; \mathbf{y}) = \theta^3(1-\theta)^7$. Figure 4 shows a plot of the likelihood function.

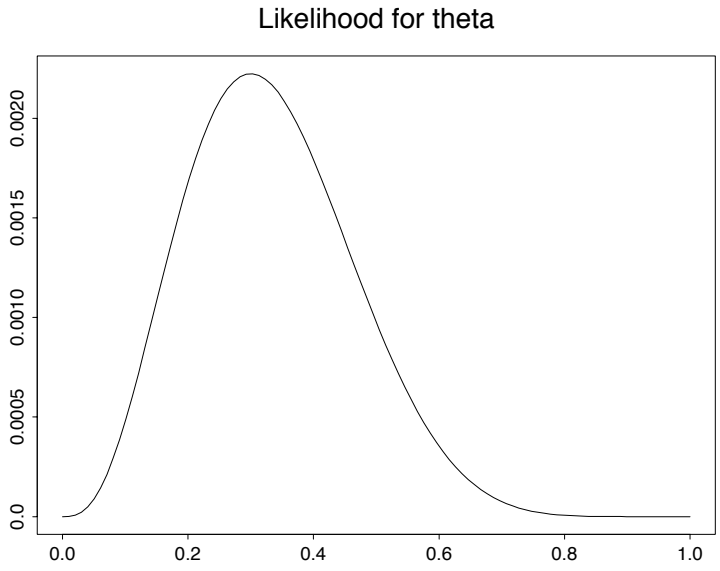

Figure 4: $L(\theta; \mathbf{y}) = \theta^3(1-\theta)^7$

For a random sample $\mathbf{Y} = (Y_1, \ldots, Y_n)$ from a Bernoulli(θ) distribution we can view $L(\theta; \mathbf{Y})$ and $\widehat{\theta}(\mathbf{Y})$ as random variables with distributions dependent on the distribution of $\mathbf{Y}$,
i.e. $L(\theta; R) = \theta^R(1-\theta)^{10-R}$, where $R = \sum Y_i$, and $\widehat{\theta} = \frac{R}{10}$ are random variables.

2.2 Moments of the score function

It is well known that MLEs are usually found more easily by solving $\frac{d \log L}{d\theta} = 0$. The **score function** $\frac{d \log L}{d\theta}$ plays a central role in statistical theory. For independent observations $y_1, \ldots, y_n$ from a distribution with pdf $f(y; \theta)$, the log likelihood is $l = \log L(\theta; y) =$

$\sum_{i=1}^{n} \log f(y_i;\theta)$ and score is

$$\frac{dl}{d\theta} = \sum_{i=1}^{n} \frac{d \log f(y_i;\theta)}{d\theta}.$$

Now for a random variable Y (a particular Y_i), we view

$$U = \frac{d \log f(Y;\theta)}{d\theta}$$

as a random variable with the distribution determined by the distribution of Y. If $Y_1, \ldots, Y_n$ are iid, then the corresponding random variables $U_1, \ldots, U_n$ are also iid, and thus $\frac{dl}{d\theta} = \frac{d \log L}{d\theta}$, when thought of as a random variable, can be expressed as a sum of iid random variables. [This will allow us to use arguments based on the central limit theorem to establish the properties of MLEs.]

Mean of U: Consider $E(U)$ where the expectation is relative to the distribution of Y.

$$\begin{aligned} E(U) &= \int \frac{d \log f(y;\theta)}{d\theta} f(y;\theta) dy \\ &= \int \frac{1}{f(y;\theta)} \frac{d f(y;\theta)}{d\theta} f(y;\theta) dy \\ &= \int \frac{d f(y;\theta)}{d\theta} dy \end{aligned}$$

(the range of the integration is the range of y). For most well-behaved functions the analytic conditions required for reversing the operations of integration with respect to y and differentiation with respect to θ are satisfied. In such cases we have

$$E(U) = \frac{d}{d\theta} \int f(y;\theta) dy = \frac{d}{d\theta}[1] = 0.$$

Variance of U: Now consider

$$\begin{aligned} \frac{d^2 \log f(y;\theta)}{d\theta^2} &= \frac{d}{d\theta}\left[\frac{1}{f(y;\theta)} \frac{df(y;\theta)}{d\theta}\right] \\ &= -\frac{1}{(f(y;\theta))^2}\left(\frac{df(y;\theta)}{d\theta}\right)^2 + \frac{1}{f(y;\theta)} \frac{d^2 f(y;\theta)}{d\theta^2} \\ &= -\left(\frac{d \log f(y;\theta)}{d\theta}\right)^2 + \frac{1}{f(y;\theta)} \frac{d^2 f(y;\theta)}{d\theta^2}. \end{aligned}$$

Think of each side of the equation as a random variable and take expectations

$$\begin{aligned} E\left[\frac{d^2 \log f(Y;\theta)}{d\theta^2}\right] &= -E\left[U^2\right] + \int \frac{1}{f(y;\theta)} \frac{d^2 f(y;\theta)}{d\theta^2} f(y;\theta) dy \\ &= -E\left[U^2\right] + \frac{d^2}{d\theta^2} \int f(y;\theta) dy \\ &= -E\left[U^2\right] + \frac{d^2}{d\theta^2}[1] = -E\left[U^2\right]. \end{aligned}$$

(Again assuming reversal of integration with respect to y and differentiation with respect to θ is justified.)

Since $E(U) = 0$, we have

$$\text{var}(U) = E(U^2) = -E\left[\frac{d^2 \log f(Y;\theta)}{d\theta^2}\right] = -E\left[\frac{dU}{d\theta}\right].$$

All this has been for a **single observation** of Y. We generalize to a random sample $Y_1, \ldots, Y_n$ by defining U as the total score

$$U = \sum_{i=1}^{n} \frac{d \log f(Y_i;\theta)}{d\theta} = \frac{dl(\theta;\mathbf{Y})}{d\theta}.$$

We still have $E(U) = 0$ since it is the sum of n iid random variables $\frac{d \log f(Y_i;\theta)}{d\theta}$ each with expectation 0. Also

$$\begin{aligned}
\text{var}(U) &= \sum_{i=1}^{n} \text{var}\left[\frac{d \log f(Y_i;\theta)}{d\theta}\right] \\
&= \sum_{i=1}^{n} -E\left[\frac{d^2 \log f(Y_i;\theta)}{d\theta^2}\right] \\
&= -E\sum_{i=1}^{n}\left[\frac{d^2 \log f(Y_i;\theta)}{d\theta^2}\right] \\
&= -E\left[\frac{d^2}{d\theta^2}\left(\sum_{i=1}^{n} \log f(Y_i;\theta)\right)\right] \\
&= -E\left(\frac{d^2 l}{d\theta^2}\right) = -E\left(\frac{dU}{d\theta}\right).
\end{aligned}$$

Information: The following terminology and symbols are commonly used:

$$\begin{aligned}
-\tfrac{d^2 l}{d\theta^2} &= \text{observed information} = J(\theta, \mathbf{y}) \\
-E\left(\tfrac{d^2 l}{d\theta^2}\right) &= \text{expected information} = I(\theta)
\end{aligned}$$

Often $I(\theta)$ is just called 'the information' or 'Fisher's information'. Sometimes it is useful to have a separate notation for the information from a single observation. We can define:

$$i(\theta) = -E\left(\frac{d^2 \log f(Y;\theta)}{d\theta^2}\right)$$

then $I(\theta) = n\, i(\theta)$.

Examples:

- $Y_i \overset{iid}{\sim} \text{Binomial}(1,\pi)$ $i = 1, \ldots, n$; $\quad R = \sum_1^n Y_i \sim \text{Binomial}(n,\pi)$.

$$\begin{aligned}
l(\pi;\mathbf{y}) &= \log(\pi^r(1-\pi)^{n-r}) = r\log\pi + (n-r)\log(1-\pi) \\
U &= \frac{dl}{d\pi} = \frac{R}{\pi} - \frac{n-R}{1-\pi} = \frac{R-n\pi}{\pi(1-\pi)} = \frac{\widehat{\pi}-\pi}{\text{var}(\widehat{\pi})} = \frac{\bar{Y}-\pi}{\text{var}(\bar{Y})} \\
I(\pi) &= \text{var}(U) = \frac{1}{[\pi(1-\pi)]^2}\text{var}(R) = \frac{n}{\pi(1-\pi)} = \frac{1}{\text{var}(\bar{Y})}
\end{aligned}$$

For a single y observation ($n = 1$) $i(\pi) = \frac{1}{\pi(1-\pi)} = \frac{1}{\text{var}(Y)}$.

- $Y_i \overset{iid}{\sim} \text{Poisson}(\mu)$ $i = 1, \ \ldots \ , n$.

$$\begin{aligned} l(\mu; \mathbf{y}) &= \log\left[\frac{e^{-n\mu}\mu^{\sum y_i}}{\prod y_i!}\right] = -n\mu + (\sum y_i)\log\mu + \text{ constant} \\ U &= \frac{dl}{d\mu} = -n + \frac{\sum Y_i}{\mu} = \frac{n(\bar{Y} - \mu)}{\mu} \\ I(\mu) &= \text{var}(U) = \frac{1}{\mu^2}\text{var}(n\bar{Y}) = \frac{n}{\mu} = \frac{1}{\text{var}(\bar{Y})} \end{aligned}$$

  For a single observation $i(\mu) = \frac{1}{\mu} = \frac{1}{\text{var}(Y)}$.

For these simple cases the information about the unknown parameter in a single observation is the inverse variance, and the information in a random sample of n observations is the inverse variance of $\bar{Y}$. Also, in both examples the MLE of the parameter of interest is $\bar{Y}$.

The larger n is the more precision we have for estimating the unknown parameter, i.e. the more 'information' (in the specific sense of inverse variance). It may be shown that $I(\theta)^{-1}$ in fact provides a lower bound for the variance of any unbiased estimator of θ.

2.3 Asymptotic distribution of the score: Score test

For a random sample $Y_1, \ \ldots \ , Y_n$ the total score

$$U = \frac{dl}{d\theta}$$

is the sum of n iid random variables. By the central limit theorem it follows that as $n \to \infty$

$$\frac{U(\theta) - 0}{\sqrt{I(\theta)}} \sim N(0, 1)$$

and

$$\frac{U^2(\theta)}{I(\theta)} \sim \chi^2_1.$$

But we should be careful when using this result to distinguish the value of θ at which $U(\theta)$ and $I(\theta)$ are evaluated. We should really write $U(\theta_0)$ and $I(\theta_0)$ where $\theta = \theta_0$ is the true value and where the expectations in U and I are relative to the distribution $f(y; \theta_0)$.

$$\frac{U(\theta_0)}{\sqrt{I(\theta_0)}} \sim N(0, 1)$$

if $\theta = \theta_0$. This can be used as an asymptotic test of $H_0 : \theta = \theta_0$—it is called a **score test**. The test may be viewed geometrically as comparing the **slope** of the tangent to $l(\theta)$ at θ_0 with the horizontal.

[$\widehat{\theta}$ is the value of θ best supported by the data. If the hypothesized value θ_0 is far from $\widehat{\theta}$, then we would expect the score test statistic to be large.]

In fact the denominator of this statistic

$$\sqrt{\text{var}[U(\theta_0)]} = \sqrt{I(\theta_0)}$$

can be replaced by any consistent estimator of it without affecting the asymptotic normality. Thus

$$\frac{U(\theta_0)}{\sqrt{I(\widehat{\theta})}}, \quad \frac{U(\theta_0)}{\sqrt{J(\theta_0)}} \text{ and } \frac{U(\theta_0)}{\sqrt{J(\widehat{\theta})}}$$

all tend asymptotically to $N(0,1)$ (even though the denominators as well as the numerators are random variables). Indeed for some problems using $J(\widehat{\theta})$ turns out to be an improvement!

Example:

- For $Y_i \overset{iid}{\sim}$ Binomial$(1,\pi)$ $i = 1, \ldots, n$, we find the score test statistic for applying a hypothesis test on the parameter. If we denote $p = \widehat{\pi}$, then

$$\begin{aligned} U(\pi) &= I(\pi)(p-\pi) \\ \frac{U(\pi)}{\sqrt{I(\pi)}} &= \sqrt{I(\pi)}(p-\pi) = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}. \end{aligned}$$

  We know this $\to N(0,1)$ but so does

$$\frac{U(\pi)}{\sqrt{I(\widehat{\pi})}} = \frac{p-\pi}{\sqrt{\frac{p(1-p)}{n}}}.$$

Example:

- $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\theta, \sigma^2)$ $\quad f(y;\theta) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{(y-\theta)^2}{2\sigma^2}\right\}$ $\quad$ (σ^2 known).

  Test $\mathrm{H}_0 : \theta = \theta_0$ against $\mathrm{H}_1 : \theta \neq \theta_0$

Likelihood :	$L(\theta) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{\sum_{i=1}^n (y_i-\theta)^2}{2\sigma^2}\right\}$
Log likelihood :	$l(\theta) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\theta)^2$
Score :	$U(\theta) = l'(\theta) = \frac{1}{\sigma^2}\sum_{i=1}^n (y_i-\theta)$
MLE :	$U(\widehat{\theta}) = 0 \Rightarrow \widehat{\theta} = \frac{\sum_{i=1}^n y_i}{n} \Rightarrow \widehat{\theta} = \bar{y}$
$I(\theta)$:	$U' = l''(\theta) = -\frac{n}{\sigma^2} \Rightarrow I(\theta) = -E(U') = \frac{n}{\sigma^2}$

  Thus, the score test statistic is:

$$z_{\text{score}} = \frac{U(\theta_0)}{\sqrt{I(\theta_0)}} = \frac{\frac{1}{\sigma^2}(\sum_{i=1}^n Y_i - n\theta_0)}{\sqrt{n/\sigma^2}} = \frac{\bar{Y}-\theta_0}{\sqrt{\sigma^2/n}} \sim N(0,1) \quad (\text{under } \mathrm{H}_0).$$

2.4 Asymptotic distribution of the MLE: Wald test

The essential result is that as $n \to \infty$,

$$\widehat{\theta} \sim N(\theta_0, I^{-1}(\theta_0))$$

where θ_0 is the true value of the parameter. I shall only give a heuristic proof; proper proof requires careful specification of regularity conditions. (In particular take care if the maximum of $l(\theta)$ occurs on a boundary of the parameter space or if the number of parameters increases as n increases.)

We have

$$0 = U(\widehat{\theta}) \simeq U(\theta_0) - (\widehat{\theta} - \theta_0)J(\theta_0).$$

Now it can be shown that $\widehat{\theta}$ is a **consistent** estimator of θ_0 (i.e. for any $\epsilon > 0$, $\Pr(|\widehat{\theta} - \theta_0| > \epsilon) \to 0$ as $n \to \infty$) so that $\widehat{\theta} - \theta_0$ is small with high probability, and terms ignored in the Taylor expansion above are 'second order'.

From the equation above

$$\widehat{\theta} - \theta_0 \simeq \frac{U(\theta_0)}{J(\theta_0)}.$$

Now as $n \to \infty$ we know that

$$U(\theta_0) \sim N(0, I(\theta_0))$$

and

$$J(\theta_0) \to I(\theta_0) \quad \text{(Strong Law of Large Numbers)}.$$

Thus

$$\text{var}(\widehat{\theta} - \theta_0) \simeq \frac{\text{var}(U(\theta_0))}{J^2(\theta_0)} \sim \frac{I(\theta_0)}{I^2(\theta_0)} = I^{-1}(\theta_0)$$

and

$$\widehat{\theta} - \theta_0 \sim N(0, I^{-1}(\theta_0))$$

or

$$\sqrt{I(\theta_0)}(\widehat{\theta} - \theta_0) \sim N(0, 1).$$

Thus, asymptotically, the MLE $\widehat{\theta}$ has a normal distribution with mean θ and variance $I^{-1}(\theta)$.

We can test $H_0 : \theta = \theta_0$ by comparing $\sqrt{I(\theta_0)}(\widehat{\theta} - \theta_0)$ with $N(0, 1)$ or $I(\theta_0)(\widehat{\theta} - \theta_0)^2$ with χ^2_1 for a 2-sided test. This test is called an **ML test**, or a **Wald test**. In terms of the log likelihood, we are using the **horizontal** distance $\widehat{\theta} - \theta_0$, appropriately scaled.

Clearly to the order of approximation used

$$\frac{U(\theta_0)}{\sqrt{I(\theta_0)}} \simeq (\widehat{\theta} - \theta_0)\sqrt{I(\theta_0)}$$

so that the score test and the Wald test of $H_0 : \theta = \theta_0$ are asymptotically equivalent. However in small samples they will differ.

It can also be shown that

$$\left.\begin{array}{l} \sqrt{I(\widehat{\theta})}(\widehat{\theta} - \theta_0) \\ \sqrt{J(\theta_0)}(\widehat{\theta} - \theta_0) \\ \sqrt{J(\widehat{\theta})}(\widehat{\theta} - \theta_0) \end{array}\right\} \text{ all } \sim N(0, 1)$$

as $n \to \infty$. In some situations one of these other forms is easier to calculate and may also be a better approximation to $N(0, 1)$.

Example:

- $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\theta, \sigma^2)$ $\qquad f(y;\theta) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{(y-\theta)^2}{2\sigma^2}\right\}$ $\qquad$ (σ^2 known).

  Test $\mathrm{H}_0 : \theta = \theta_0$ against $\mathrm{H}_1 : \theta \neq \theta_0$,

$$\text{Score}: \quad U(\theta) = l'(\theta) = \tfrac{1}{\sigma^2}\textstyle\sum_{i=1}^{n}(y_i - \theta)$$
$$I(\theta): \quad U' = l''(\theta) = -\tfrac{n}{\sigma^2} \Rightarrow I(\theta) = -E(U') = \tfrac{n}{\sigma^2}$$

  Thus, the Wald test statistic is:

$$z_{\text{Wald}} = \frac{\widehat{\theta} - \theta_0}{\sqrt{I^{-1}(\theta_0)}} = \frac{\bar{Y} - \theta_0}{\sqrt{\sigma^2/n}} \sim N(0,1) \quad (\text{under } \mathrm{H}_0).$$

2.5 Likelihood Ratio test for a simple hypothesis

A Taylor expansion of $l(\widehat{\theta})$ gives

$$l(\widehat{\theta}) \simeq l(\theta_0) + (\widehat{\theta} - \theta_0)U(\theta_0) - \frac{1}{2}(\widehat{\theta} - \theta_0)^2 J(\theta_0). \qquad (*)$$

To see why we need to expand to the second term, consider the orders of magnitude. We have

$$U(\theta_0) \sim N(0, I(\theta_0)) \sim N(0, n\,i(\theta_0))$$

so that $U(\theta_0) \sim O(n^{1/2})$.

$$\widehat{\theta} - \theta_0 \sim N(0, I^{-1}(\theta_0)) \sim N\left(0, \frac{1}{ni(\theta_0)}\right)$$

so $\widehat{\theta} - \theta_0 \sim O(n^{-1/2})$ and

$$J(\theta_0) \sim I(\theta_0) \sim O(n).$$

Each of the last 2 terms in $(*)$ is $O(1)$; the next term in the expansion is actually $O(n^{-1})$ [See: McCullagh and Nelder, 1983, Generalized Linear Models, Chapman and Hall, Appendix A]

Since we saw in Section 2.4 that

$$\widehat{\theta} - \theta_0 \sim \frac{U(\theta_0)}{J(\theta_0)}$$

$(*)$ gives

$$l(\widehat{\theta}) - l(\theta_0) \simeq \frac{1}{2}(\widehat{\theta} - \theta_0)^2 J(\theta_0) \simeq \frac{1}{2}\frac{U^2(\theta_0)}{J(\theta_0)}$$
$$2(l(\widehat{\theta}) - l(\theta_0)) \simeq \frac{(\widehat{\theta} - \theta_0)^2}{I^{-1}(\theta_0)} \simeq \frac{U^2(\theta_0)}{I(\theta_0)}.$$

All three statistics are **asymptotically distributed as** χ^2_1 and are equivalent if we ignore terms of order $n^{-1/2}$. Usually the likelihood ratio for testing the simple null hypothesis $\theta = \theta_0$ against the composite alternative hypothesis $\theta \neq \theta_0$ is defined by

$$\text{Likelihood Ratio} = LR = \frac{L(\theta_0)}{L(\widehat{\theta})}$$

and

$$-2\log(LR) = -2(l(\theta_0) - l(\widehat{\theta})).$$

The **likelihood ratio test** of $\theta = \theta_0$ rejects H_0 if

$$-2(l(\theta_0) - l(\widehat{\theta})) > \chi^2_1(\alpha)$$

for the suitable critical value α. This is equivalent to using the **vertical** deviation of the log likelihood.

Example:

- $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\theta, \sigma^2)$ $\qquad f(y;\theta) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{(y-\theta)^2}{2\sigma^2}\right\}$ $\qquad$ (σ^2 known).

  Test $\mathrm{H}_0 : \theta = \theta_0$ against $\mathrm{H}_1 : \theta \neq \theta_0$

  $$\begin{aligned} &\text{Likelihood :} && L(\theta) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{\sum_{i=1}^n (y_i-\theta)^2}{2\sigma^2}\right\} \\ &\text{Log likelihood :} && l(\theta) = -\tfrac{n}{2}\log(2\pi\sigma^2) - \tfrac{1}{2\sigma^2}\textstyle\sum_{i=1}^n (y_i-\theta)^2 \\ &\text{MLE :} && U(\widehat{\theta}) = 0 \Rightarrow \widehat{\theta} = \frac{\sum_{i=1}^n y_i}{n} \Rightarrow \widehat{\theta} = \bar{y} \end{aligned}$$

  Thus, the likelihood ratio test statistic is:

  $$-2\log(LR) = -2\log\left(\frac{L(\theta_0)}{L(\widehat{\theta})}\right) = -2\log\left[\frac{(2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{\sum_{i=1}^n (y_i-\theta_0)^2}{2\sigma^2}\right\}}{(2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{\sum_{i=1}^n (y_i-\widehat{\theta})^2}{2\sigma^2}\right\}}\right]$$

  or, since $\widehat{\theta} = \bar{y}$,

  $$-2(l(\theta_0) - l(\widehat{\theta})) = \frac{1}{\sigma^2}\{2n\bar{y}(\widehat{\theta} - \theta_0) - n\widehat{\theta}^2 + n\theta_0^2\} = \frac{n}{\sigma^2}(\bar{y} - \theta_0)^2.$$

  This should be compared with a χ^2_1 distribution.

When $l(\theta)$ is a quadratic function of θ (as with the normal distribution with σ^2 known), all the approximations above are **exact** and we have exact small sample equality between the 3 statistics. All 3 statistics may be used to obtain approximate $(1-\alpha)$ confidence intervals for an unknown parameter. e.g. 95% confidence intervals defined by

$$\left\{\theta : 2\left(l\left(\widehat{\theta}\right) - l(\theta)\right) < 3.84\right\} \text{ or } \left\{\theta : \left(\widehat{\theta} - \theta\right)^2 < 3.84\, I^{-1}\left(\widehat{\theta}\right)\right\} \text{ or } \left\{\theta : \frac{U^2(\theta)}{I(\theta)} < 3.84\right\}.$$

Example:

- Poisson. Suppose $y_1, \ldots, y_n$ are independent observations from a Poisson(μ).

  $$\begin{aligned} l(\mu) &= -n\mu + \left(\sum y_i\right)\log(\mu) - \sum \log(y_i!) = n(-\mu + \bar{y}\log(\mu)) + \text{constant} \\ U(\mu) &= \frac{n}{\mu}(\bar{Y} - \mu) = I(\mu)(\bar{Y} - \mu) \quad \text{and} \quad I(\mu) = \frac{n}{\mu} \end{aligned}$$

  We set $\widehat{\mu} = \bar{Y}$

  $$\begin{aligned} -2[l(\mu_0) - l(\widehat{\mu})] &= 2n\left[-(\widehat{\mu} - \mu_0) + \bar{y}\log\left(\frac{\widehat{\mu}}{\mu_0}\right)\right] = 2n\left[-(\bar{y} - \mu_0) + \bar{y}\log\left(\frac{\bar{y}}{\mu_0}\right)\right] \\ \frac{U^2(\mu_0)}{I(\mu_0)} &= \frac{(\bar{y} - \mu_0)^2}{\frac{\mu_0}{n}} \quad \text{and also} \quad \frac{(\widehat{\mu} - \mu_0)^2}{I^{-1}(\mu_0)} = \frac{(\bar{y} - \mu_0)^2}{\frac{\mu_0}{n}}. \end{aligned}$$

Suppose $n = 10$, $\bar{y} = 8.5$ and we want to test $H_0 : \mu = \mu_0 = 6$.

$$-2[l(\mu_0) - l(\widehat{\mu})] = 20\left[-2.5 + 8.5\log\left(\frac{8.5}{6}\right)\right] = 9.212$$

and

$$\frac{(\bar{y} - \mu_0)^2}{\frac{\mu_0}{n}} = \frac{2.5^2}{0.6} = 10.417.$$

Both statistics are highly significant compared with χ^2_1.

95% confidence interval using LR statistic

$$\left\{\mu : 20\left[\mu - 8.5 + 8.5\log\left(\frac{8.5}{\mu}\right)\right] < 3.84\right\}.$$

This can be solved numerically yielding 95% confidence interval for $\mu = (6.819, 10.437)$.

95% confidence interval using the score statistic

$$\left\{\mu : \frac{(8.5 - \mu)^2}{\frac{\mu}{10}} < 3.84\right\}.$$

Solving the quadratic

$$\begin{aligned} \mu^2 - 17\mu + 72.25 &= 0.384\mu \\ \mu^2 - 17.384\mu + 72.25 &= 0 \end{aligned}$$

$$95\%\text{ CI for } \mu = \frac{17.384 \pm \sqrt{17.384^2 - 4 \times 1 \times 72.25}}{2} = \frac{17.384 \pm \sqrt{13.2035}}{2} = (6.875, 10.509).$$

95% confidence interval using the Wald (ML) statistic with $I(\widehat{\theta})$ in place of $I(\theta_0)$

$$\left\{\mu : \frac{(8.5 - \mu)^2}{0.85} < 3.84\right\}.$$

Thus

$$95\%\text{ CI for } \mu = 8.5 \pm \sqrt{3.84 \times 0.85} = 8.5 \pm 1.807 = (6.693, 10.307).$$

(You could use these limits as an initial approximation for the first method.)

2.6 Tests for multiparameter problems

Suppose random variable Y has p.d.f. $f(y; \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is a p-dimensional parameter. For a sample $\mathbf{y} = (y_1, \ldots, y_n)$ we define

- Likelihood $L(\boldsymbol{\theta}; \mathbf{y}) = \prod_{i=1}^n f(y_i; \boldsymbol{\theta})$
- Log likelihood $l(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i=1}^n \log f(y_i; \boldsymbol{\theta})$
- Score vector $\mathbf{U}(\boldsymbol{\theta})$ with jth element $\dfrac{dl}{d\theta_j}$

- Expected information matrix $I(\boldsymbol{\theta})$ with (j,k)th element $-E\left[\frac{\partial^2 l}{\partial\theta_j\partial\theta_k}\right]$
- Observed information matrix $J(\boldsymbol{\theta})$ with (j,k)th element $-\frac{\partial^2 l}{\partial\theta_j\partial\theta_k}$

The following results are straightforward extensions of the 1-parameter results.

1. $E(\mathbf{U}) = \mathbf{0}$.
2. The variance-covariance matrix of the vector random variable $\mathbf{U}$ is $\text{var}(\mathbf{U}) = I(\boldsymbol{\theta})$, i.e.,
$$\text{cov}\left(\frac{\partial l}{\partial\theta_j}, \frac{\partial l}{\partial\theta_k}\right) = E\left(\frac{\partial l}{\partial\theta_j}\frac{\partial l}{\partial\theta_k}\right) = -E\left(\frac{\partial^2 l}{\partial\theta_j\partial\theta_k}\right).$$
3. As $n\to\infty$ the distribution of $\mathbf{U}$ tends to a p-variate normal distribution
$$\mathbf{U}(\boldsymbol{\theta}_0) \sim N_p(\mathbf{0}, I(\boldsymbol{\theta}_0))$$
and
$$\mathbf{U}(\boldsymbol{\theta}_0)^T I^{-1}(\boldsymbol{\theta}_0)\mathbf{U}(\boldsymbol{\theta}_0) \sim \chi^2_p$$
where $\boldsymbol{\theta}_0$ is the true value of $\boldsymbol{\theta}$, the expectations are relative to the distribution $f(y;\boldsymbol{\theta}_0)$ and the derivatives are evaluated at $\boldsymbol{\theta}_0$.
4. If $\widehat{\boldsymbol{\theta}}$ is the MLE, i.e. the solution of the system of the p equations $\mathbf{U}(\boldsymbol{\theta}) = \mathbf{0}$, then as $n\to\infty$
$$\widehat{\boldsymbol{\theta}} \sim N_p(\boldsymbol{\theta}_0, I^{-1}(\boldsymbol{\theta}_0))$$
and
$$(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T I(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \sim \chi^2_p.$$
5. The asymptotic results (3 and 4) above still hold if $I(\boldsymbol{\theta}_0)$ is replaced by $I(\widehat{\boldsymbol{\theta}})$, $J(\boldsymbol{\theta}_0)$ or $J(\widehat{\boldsymbol{\theta}})$.
6. If $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ is true, then
$$-2\log\left\{\frac{L(\boldsymbol{\theta}_0)}{L(\widehat{\boldsymbol{\theta}})}\right\} = -2[l(\boldsymbol{\theta}_0) - l(\widehat{\boldsymbol{\theta}})] \sim \chi^2_p.$$

Hypothesis tests:

Tests of the simple null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ which are asymptotically correct can be based on any of the results in 3, 4 or 6. Reject H_0, at the $100\alpha\%$ level of significance, if

Likelihood ratio test	$-2[l(\boldsymbol{\theta}_0) - l(\widehat{\boldsymbol{\theta}})] > \chi^2_p(\alpha)$
Score test	$\mathbf{U}(\boldsymbol{\theta}_0)^T I^{-1}(\boldsymbol{\theta}_0)\mathbf{U}(\boldsymbol{\theta}_0) > \chi^2_p(\alpha)$
Wald test (ML test)	$(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T I(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) > \chi^2_p(\alpha)$

for some suitable α (e.g. 0.05 or 0.01).

Confidence regions:

p-dimensional confidence regions for $\boldsymbol{\theta}$ are obtained from the set of values of $\boldsymbol{\theta}$ not rejected by the corresponding test at a given level α,

Likelihood ratio	$\{\boldsymbol{\theta} : -2[l(\boldsymbol{\theta}) - l(\widehat{\boldsymbol{\theta}})] < \chi^2_p(\alpha)\}$
Score	$\{\boldsymbol{\theta} : \mathbf{U}(\boldsymbol{\theta})^T I^{-1}(\boldsymbol{\theta})\mathbf{U}(\boldsymbol{\theta}) < \chi^2_p(\alpha)\}$
Wald (ML)	$\{\boldsymbol{\theta} : (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T I(\boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) < \chi^2_p(\alpha)\}$

2.7 The LR test for composite hypothesis

Most frequently in multiparameter problems we want to test hypotheses about one parameter without specifying values of other parameters (e.g. in simple linear regression we may want to test $\beta = 0$ without specifying α or σ.)

Let Ω denote the whole parameter space (say R^p if $\boldsymbol{\theta}$ is p-dimensional and its components have unrestricted range). In general terms we have

$$\begin{aligned} &H_0 : \boldsymbol{\theta} \in \omega \quad \text{for some subspace } \omega \subset \Omega \\ &H_1 : \boldsymbol{\theta} \notin \omega. \end{aligned}$$

The **general likelihood ratio test** compares the maximum likelihood attainable if $\boldsymbol{\theta}$ is restricted to ω (i.e. under the 'reduced' model) with the maximum attainable if $\boldsymbol{\theta}$ is unrestricted (i.e. under the 'full' model):

$$\begin{aligned} LR &= \frac{\max_{\boldsymbol{\theta}\in\omega} L(\boldsymbol{\theta})}{\max_{\boldsymbol{\theta}\in\Omega} L(\boldsymbol{\theta})} \\ &= \frac{L(\widehat{\boldsymbol{\theta}}_\omega)}{L(\widehat{\boldsymbol{\theta}})} \end{aligned}$$

where $\widehat{\boldsymbol{\theta}}$ is the unrestricted MLE of $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}_\omega$ is the restricted MLE of $\boldsymbol{\theta}$.

(Some writers (e.g. Dobson, A. J., An Introduction to Generalized Linear Models, 1990, Chapman and Hall) define the likelihood ratio the other way around or as $LR = \frac{\max_{\boldsymbol{\theta}\in\omega} L(\boldsymbol{\theta})}{\max_{\boldsymbol{\theta}\notin\omega} L(\boldsymbol{\theta})}$, but these definitions give monotonic functions of our definition and lead to equivalent tests.)

With log likelihoods we consider

$$\lambda = -2\log(LR) = -2[l(\widehat{\boldsymbol{\theta}}_\omega) - l(\widehat{\boldsymbol{\theta}})]$$

Theorem: Suppose dim $\omega = r$. It can be shown, under suitable regularity conditions, that as $n \to \infty$, $-2\log(LR)$ has the distribution χ^2_{p-r} if H_0 is true.

Here the degrees of freedom $= (p - r) =$ the difference in the dimensions between the space of the full model and the space of the reduced model = the number of independent linear restrictions imposed on components of $\boldsymbol{\theta}$ by H_0.

Example:

- Normal sample. $\quad Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$

  Test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$,

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

Under ω ($H_0 : \mu = \mu_0$): MLE $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\mu_0)^2$.
Under Ω ($H_0 \cup H_1$: μ and σ^2 can take any value):

$$\text{MLEs} \quad \widehat{\widehat{\mu}} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\widehat{\widehat{\sigma}}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2.$$

Thus, the likelihood ratio is

$$LR = \frac{\max_{\boldsymbol{\theta}\in\omega} L(\boldsymbol{\theta})}{\max_{\boldsymbol{\theta}\in\Omega} L(\boldsymbol{\theta})} = \frac{(2\pi\widehat{\sigma}^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\widehat{\sigma}^2}\sum_{i=1}^{n}(y_i-\mu_0)^2\right\}}{(2\pi\widehat{\widehat{\sigma}}^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\widehat{\widehat{\sigma}}^2}\sum_{i=1}^{n}(y_i-\bar{y})^2\right\}} = \left(\frac{\widehat{\sigma}^2}{\widehat{\widehat{\sigma}}^2}\right)^{-\frac{n}{2}}$$

$$= \left(\frac{\sum_{i=1}^{n}(y_i-\mu_0)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}\right)^{-\frac{n}{2}}.$$

But

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\mu_0)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 + (\bar{y}-\mu_0)^2,$$

and thus

$$LR = \left(\frac{\widehat{\widehat{\sigma}}^2 + (\bar{y}-\mu_0)^2}{\widehat{\widehat{\sigma}}^2}\right)^{-\frac{n}{2}} = \left(1 + \frac{t^2}{n-1}\right)^{-\frac{n}{2}},$$

where

$$t = \frac{\sqrt{n}(\bar{y}-\mu_0)}{s} \quad \text{and} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2.$$

Rejecting for small values of LR is equivalent to rejecting for large values of t^2, i.e. this test is equivalent to a two-tailed t-test.

Here we can determine an exact distribution for LR, or rather the equivalent test statistic t, i.e.

$$t \sim t_{n-1} \quad (\text{under} \quad H_0 : \mu = \mu_0).$$

However, in general, the distribution of LR (under H_0) is difficult to determine and we need to use the approximate χ^2 method.

2.8 Bayesian inference [2]

An alternative to MLE, which also utilizes the information in the likelihood function, is Bayesian estimation. Bayesian estimation differs from classical methods of point estimation, including MLE, by treating the parameter θ in a probability model as a random variable; in the classical or frequentist approach to point estimation, θ is treated as a fixed unknown real number, not a random variable.

[2] From Christensen, R. et al., Bayesian Ideas and Data Analysis, An Introduction for Scientists and Statisticians, Chapman and Hall, 2011.

2.8.1 Bayes theorem

For two events A and B the conditional probability of A given B is defined as

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)},$$

in which $A \cap B$ denotes the intersection of A and B, i.e., the event that both A and B occur.

Let A^c denote the complement of A, that is, all the outcomes that are not part of A. Then **Bayes Theorem** allows us to compute $Pr(A|B)$ via

$$Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(B|A)Pr(A) + Pr(B|A^c)Pr(A^c)}.$$

using the Law of Total Probability in the denominator, we have

$$Pr(B) = Pr([A \cap B]) + Pr([A^c \cap B]) = Pr(B|A)Pr(A) + Pr(B|A^c)Pr(A^c).$$

Theorem: Let our sample space S be partitioned into disjoint events $A_k, k = 1, 2, ..., n$. Then for $1 \leqslant j \leqslant n$,

$$Pr(A_j|B) = \frac{Pr(B|A_j)Pr(A_j)}{\sum_{k=1}^{n} Pr(B|A_k)Pr(A_k)}.$$

In which, using the law of total probability:

$$Pr(B) = Pr(B|A_1)Pr(A_1) + ... + Pr(B|A_n)Pr(A_n).$$

Example:

- Drug testing. Let D indicate a drug user and C indicate someone who is clean of drugs. Let + indicate that someone tests positive for a drug, and - indicates testing negative. The overall prevalence of drug use in the population is $Pr(D) = 0.01$, so in this population drug use is relatively rare. The *sensitivity* of the drug test is, say, $Pr(+|D) = 0.98$. This is how good the test is at identifying people who use drugs. The *specificity* of the drug test is $Pr(-|C) = 0.95$, which is how good the test is at correctly identifying nonusers. We want to find the probability that someone uses drugs given that they tested positive.

  From Bayes Theorem, the probability is

$$\begin{aligned} Pr(D|+) &= \frac{Pr(+|D)Pr(D)}{Pr(+|D)Pr(D) + Pr(+|C)Pr(C)} \\ &= \frac{0.98 \times 0.01}{[0.98 \times 0.01] + [(1 - 0.95) \times (1 - 0.01)]} \\ &= \frac{0.98 \times 0.01}{[0.98 \times 0.01] + [0.05 \times 0.99]} \\ &= \frac{0.0098}{0.0098 + 0.0495} \\ &= 0.165. \end{aligned}$$

Even after testing positive for drug use using a very good test, there is an $1 - 0.165 = 83\%$ chance, that the person does not use drugs. The overwhelming probability is that a positive test outcome is incorrect. This result is driven by the very low initial probability of drug use, so even after incorporating the positive test, the probability remains low.

2.8.2 Bayesian modelling

Bayesian statistics starts by using (prior) probabilities to describe your current state of knowledge. It then incorporates information through the collection of data. This results in new (posterior) probabilities to describe your state of knowledge after combining the prior probabilities with the data. In Bayesian statistics, all uncertainty and all information are incorporated through the use of probability distributions, and all conclusions obey the laws of probability theory.

Bayesian statistics typically begins with prior information about the state of nature $\boldsymbol{\theta}$ that is embodied in the prior density $p(\boldsymbol{\theta})$. It then uses Bayes Theorem and the random data $\mathbf{y}$, with sampling density $f(\mathbf{y}|\boldsymbol{\theta})$, to update this information into a posterior density $p(\boldsymbol{\theta}|\mathbf{y})$ that incorporates both the prior information and the data. Specifically, Bayes Theorem tells us that

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}, \boldsymbol{\theta})}{f(\mathbf{y})} = \frac{f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}.$$

To a Bayesian, the best information one can ever have about $\boldsymbol{\theta}$ is to know the posterior density $p(\boldsymbol{\theta}|\mathbf{y})$. Nonetheless, it is often convenient to summarize the posterior information (most summaries involve integration, which we will typically perform by computer simulation).

When θ is a scalar with a continuous distribution, the posterior median, say $\tilde{\theta}$, satisfies

$$\frac{1}{2} = \int_{-\infty}^{\tilde{\theta}} p(\theta|\mathbf{y})\mathrm{d}\theta.$$

The posterior mode is the value of θ , say θ_M, such that

$$\theta_M = \arg\{\max_{\theta}[p(\theta|\mathbf{y})]\}.$$

The posterior mean is

$$E(\theta|\mathbf{y}) = \int \theta\, p(\theta|\mathbf{y})\mathrm{d}\theta.$$

The posterior variance is

$$\mathrm{Var}(\theta|\mathbf{y}) = \int [\theta - E(\theta|\mathbf{y})]^2\, p(\theta|\mathbf{y})\mathrm{d}\theta.$$

Example:

- Binomial Data. Suppose we are interested in assessing the proportion of transportation industry workers who use drugs on the job. Let θ denote this proportion and assume that a random sample of n workers is to be taken while they are actually on the job. Each individual will be strapped down and forced to deliver a blood sample which will

be analysed for a variety of legal and illegal drugs. The number of positive tests is denoted y.

Data obtained like this follow a Binomial distribution, as, $y|\theta \sim \text{Bin}(n, \theta)$, with discrete density function for $y = 0, 1, ..., n$,

$$f(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}.$$

It will be convenient, but by no means necessary, to use a Beta distribution as a model for the prior. The Beta distribution is conjugate to the Binomial distribution in the sense that the densities have similar functional forms. Let $\theta \sim \text{Beta}(a, b)$, with density function for $\theta \in (0, 1)$,

$$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}.$$

The *hyperparameters* a and b are selected to reflect the researchers beliefs and uncertainty.

To obtain the posterior, apply Bayes Theorem to the densities

$$\begin{aligned}
p(\theta|y) &= \frac{f(y|\theta)p(\theta)}{\int f(y|\theta)p(\theta)\mathrm{d}\theta} \\
&= \frac{\binom{n}{y}\theta^y(1-\theta)^{n-y}\,\dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}}{\displaystyle\int_0^1 \binom{n}{y}\theta^y(1-\theta)^{n-y}\,\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\mathrm{d}\theta} \\
&= \frac{\theta^y(1-\theta)^{n-y}\,\theta^{a-1}(1-\theta)^{b-1}}{\displaystyle\int_0^1 \theta^y(1-\theta)^{n-y}\,\theta^{a-1}(1-\theta)^{b-1}\mathrm{d}\theta} \\
&= \frac{\theta^{y+a-1}(1-\theta)^{n-y+b-1}}{\displaystyle\int_0^1 \theta^{y+a-1}(1-\theta)^{n-y+b-1}\mathrm{d}\theta} \\
&= \frac{\Gamma(n+a+b)}{\Gamma(y+a)\Gamma(n-y+b)}\theta^{y+a-1}(1-\theta)^{n-y+b-1}.
\end{aligned}$$

The posterior distribution turns out to be another Beta distribution, in particular the distribution $\text{Beta}(y + a, n - y + b)$. An estimate for θ could be

$$E(\theta \mid y) = \frac{y+a}{n+a+b}.$$

Prior to the widespread use of computers to approximate posterior distributions, Bayesian statistics could only be performed when the calculus involved in finding the posterior distribution could be performed. This calculus is often relatively simple when, in terms of the parameters $\boldsymbol{\theta}$, the prior density $p(\boldsymbol{\theta})$ has the same functional form as the sampling density $f(\mathbf{y}|\boldsymbol{\theta})$ so that, after applying Bayes Theorem, the posterior has the same functional form as the prior. Such pairs of sampling densities and prior densities are called **conjugate families**.

Example:

- Poisson Data. Let $y_1, ..., y_n$ be a sample of iid Poisson(θ) with $\theta > 0$ and suppose that the prior distribution of θ is Exp($\frac{1}{\lambda}$).

  To obtain the posterior distribution, Bayes Theorem is applied to the densities

$$\begin{aligned}
p(\theta|\mathbf{y}) &= \frac{f(\mathbf{y}|\theta)p(\theta)}{\int f(\mathbf{y}|\theta)p(\theta)\mathrm{d}\theta} \\
&= \frac{\frac{e^{-n\theta}\theta^{\sum y_i}}{\prod y_i!} \times \frac{1}{\lambda}e^{-\frac{\theta}{\lambda}}}{\int_0^\infty \frac{e^{-n\theta}\theta^{\sum y_i}}{\prod y_i!} \times \frac{1}{\lambda}e^{-\frac{\theta}{\lambda}}\mathrm{d}\theta} \\
&= \frac{\frac{1}{\lambda\prod y_i!}e^{-n\theta}\theta^{\sum y_i} \times e^{-\frac{\theta}{\lambda}}}{\frac{1}{\lambda\prod y_i!}\int_0^\infty e^{-n\theta}\theta^{\sum y_i} \times e^{-\frac{\theta}{\lambda}}\mathrm{d}\theta} \\
&= \frac{e^{-\theta(n+\frac{1}{\lambda})}\theta^{\sum y_i}}{\frac{\Gamma(\sum y_i+1)}{(n+\frac{1}{\lambda})^{(\sum y_i+1)}}\int_0^\infty \frac{(n+\frac{1}{\lambda})^{(\sum y_i+1)}}{\Gamma(\sum y_i+1)}e^{-\theta(n+\frac{1}{\lambda})}\theta^{\sum y_i}\mathrm{d}\theta} \\
&= \frac{(n+\frac{1}{\lambda})^{(\sum y_i+1)}e^{-\theta(n+\frac{1}{\lambda})}\theta^{\sum y_i}}{\Gamma(\sum y_i+1)}
\end{aligned}$$

  Thus the posterior distribution is Gamma($\sum y_i + 1, n + \frac{1}{\lambda}$), and

$$E(\theta \mid \mathbf{y}) = \frac{\sum y_i + 1}{n + \frac{1}{\lambda}}.$$

3 Maximum likelihood estimation

3.1 Single parameter problems: direct estimates

We have already seen several examples where we can obtain an explicit expression for the MLE in terms of a random sample by solving $U(\theta) = 0$ (and verifying that it corresponds to a maximum of the (log) likelihood by checking the observed information at $\widehat{\theta}$, $J(\widehat{\theta}) = -l''(\theta)|_{\theta=\widehat{\theta}}$, is positive), e.g., the MLE for a parameter θ based on a normal $N(\theta, \sigma^2)$, where σ^2 is known, sample $y_1, \ldots, y_n$ is obtained by solving the score function set equal to zero, and gives the MLE as $\widehat{\theta} = \bar{y}$.

3.2 Multiparameter problems: direct estimates

If we have a parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$, then to maximize the likelihood, or the log likelihood, we often solve

$$U(\boldsymbol{\theta}) = \mathbf{0},$$

where $U(\boldsymbol{\theta})$ is the score vector. We should check that this corresponds to a maximum by investigating the Hessian matrix $H(\boldsymbol{\theta}) = -J(\boldsymbol{\theta})$ (based on the log likelihood function). For the MLE $\widehat{\boldsymbol{\theta}}$ to correspond to a maximum of the log likelihood we should have the observed information matrix at $\widehat{\boldsymbol{\theta}}$, $J(\widehat{\boldsymbol{\theta}}) = -H(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$, positive definite.

Example:

- $y_1, \ldots, y_n \overset{iid}{\sim} N(\mu, \sigma^2)$, $p = 2$. Parameter vector $\boldsymbol{\theta} = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$ to be estimated.

  Log likelihood: $l(\boldsymbol{\theta}) = \text{const} - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum(y_i - \mu)^2$

  Score vector: $U(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial l}{\partial \mu} \\ \frac{\partial l}{\partial \sigma^2} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2}\sum(y_i - \mu) \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum(y_i - \mu)^2 \end{pmatrix}$

  To determine the MLE for $\boldsymbol{\theta}$ solve $U(\boldsymbol{\theta}) = \mathbf{0}$, i.e. $\begin{pmatrix} \frac{\partial l}{\partial \mu} \\ \frac{\partial l}{\partial \sigma^2} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$.

  This gives the MLEs of μ and σ^2 as

$$\sum(y_i - \widehat{\mu}) = 0 \quad \Rightarrow \quad n\bar{y} - n\widehat{\mu} = 0 \quad \Rightarrow \quad \widehat{\mu} = \bar{y}$$

$$-\frac{n}{2\widehat{\sigma}^2} + \frac{1}{2\widehat{\sigma}^4}\sum(y_i - \widehat{\mu})^2 = 0 \quad \Rightarrow \quad \widehat{\sigma}^2 = \frac{1}{n}\sum(y_i - \widehat{\mu})^2 = \frac{1}{n}\sum(y_i - \bar{y})^2$$

  [Note that $\widehat{\sigma}^2$ is a biased estimator for σ^2. $s^2 = \frac{1}{n-1}\sum(y_i - \bar{y})^2$ is the unbiased estimator.]

$$\frac{\partial^2 l}{\partial \mu^2} = -\frac{n}{\sigma^2}, \quad \frac{\partial^2 l}{\partial \mu \partial \sigma^2} = \frac{\partial^2 l}{\partial \sigma^2 \partial \mu} = -\frac{1}{\sigma^4}\sum_{i=1}^{n}(y_i - \mu), \quad \frac{\partial^2 l}{(\partial \sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(y_i - \mu)^2$$

  Since $\mathrm{E}(Y_i - \mu) = 0$ and $\mathrm{E}[\sum(Y_i - \mu)^2] = n\sigma^2$, Fisher's information matrix is

$$I(\boldsymbol{\theta}) = \mathrm{E}(J(\boldsymbol{\theta})) = -\mathrm{E}(H(\boldsymbol{\theta})) = -\mathrm{E}\begin{pmatrix} \frac{\partial^2 l}{\partial \mu^2} & \frac{\partial^2 l}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 l}{\partial \mu \partial \sigma^2} & \frac{\partial^2 l}{(\partial \sigma^2)^2} \end{pmatrix} = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$$

For large samples, the variance-covariance matrix of $\widehat{\boldsymbol{\theta}}$ is $\text{var}(\widehat{\boldsymbol{\theta}}) = I^{-1}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix}$.

Estimated standard errors: $\text{ESE}(\widehat{\mu}) = \frac{\widehat{\sigma}}{\sqrt{n}}$ $\quad$ $\text{ESE}(\widehat{\sigma}^2) = \sqrt{\frac{2\widehat{\sigma}^4}{n}}$

Asymptotic distribution of MLE: $\widehat{\boldsymbol{\theta}} \dot{\sim} \text{BVN}\left(\begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}, \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix}\right)$

Note that $-l(\boldsymbol{\theta})$ (i.e., minus log likelihood) is a minimum at $\widehat{\boldsymbol{\theta}}$ since $J(\widehat{\boldsymbol{\theta}}) = -H(\widehat{\boldsymbol{\theta}}) = \begin{pmatrix} \frac{n}{\widehat{\sigma}^2} & 0 \\ 0 & \frac{n}{2\widehat{\sigma}^4} \end{pmatrix}$ is positive definite (as $\frac{n}{\widehat{\sigma}^2} > 0$ and $\left|-H(\widehat{\theta})\right| > 0$). Thus, $\widehat{\boldsymbol{\theta}}$ corresponds to a maximum of $l(\boldsymbol{\theta})$.

3.3 Iterative estimation of the MLE

Often there is no explicit solution to the p equations

$$U_j(\theta) = \frac{dl}{d\theta_j} = 0 \quad j = 1, \ \ldots \ , p$$

since they are non-linear in $\boldsymbol{\theta}$. In such situations we need iterative methods to determine the MLEs, e.g., Newton-Raphson. Suppose $\boldsymbol{\theta}^{(0)}$ is an approximate solution of the equations. By a Taylor expansion we have for $j = 1, \ \ldots \ , p$

$$\begin{aligned} 0 &= U_j(\widehat{\boldsymbol{\theta}}) \simeq U_j(\boldsymbol{\theta}^{(0)}) + \sum_{s=1}^{p} \frac{\partial U_j(\boldsymbol{\theta}^{(0)})}{\partial \theta_s}(\widehat{\theta}_s - \theta_s^{(0)}) \text{ where } U_j = \frac{\partial l}{\partial \theta_j} \\ \mathbf{0} &\simeq \mathbf{U}(\boldsymbol{\theta}^{(0)}) - J(\boldsymbol{\theta}^{(0)})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{(0)}) \end{aligned}$$

Thus the next approximation is

$$\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(0)} + J^{-1}(\boldsymbol{\theta}^{(0)})\mathbf{U}(\boldsymbol{\theta}^{(0)})$$

Actually it is often simpler to think in terms of the vector of 'corrections'

$$\boldsymbol{\delta}_{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)},$$

which is the solution of the system of equations

$$J(\boldsymbol{\theta}^{(0)})\boldsymbol{\delta}_{\boldsymbol{\theta}} = \mathbf{U}(\boldsymbol{\theta}^{(0)}).$$

This process usually converges quite rapidly; however a more convenient and possibly faster iteration is obtained by replacing J by I. $I(\theta)$ is often easier to calculate; also for many problems it is either constant or easier to recalculate at each iteration. This modification of Newton-Raphson is Fisher's method of scoring.

3.4 Fisher's method of scoring: single parameter case

We can apply Newton-Raphson to determine the maximum of the log likelihood. The **Newton-Raphson** iterative formula is given by:

$$\theta_{r+1} = \theta_r - \frac{U(\theta_r)}{l''(\theta_r)} = \theta_r + \frac{U(\theta_r)}{J(\theta_r)} \qquad (r = 0, 1, \ldots)$$

Fisher's method of scoring replaces $l''(\theta)$ by its expectation $\mathrm{E}(l''(\theta))$, i.e.

$$\theta_{r+1} = \theta_r - \frac{U(\theta_r)}{\mathrm{E}(l''(\theta_r))} = \theta_r + \frac{U(\theta_r)}{I(\theta_r)} \qquad (r = 0, 1, \ldots)$$

For both iterative methods we need to specify an initial value of θ, this value is denoted by θ_0.

Example:

- Estimating the parameter θ of the truncated Poisson with p.m.f.

$$f(y;\theta) = \frac{\theta^y e^{-\theta}}{y!(1-e^{-\theta})} \qquad (y = 1, 2, \ldots).$$

(Poisson-type of distribution for the case where it is not possible to observe zeros).

Log likelihood:

$$l(\theta) = \sum_{i=1}^{n} y_i \log\theta - n\theta - n\log(1-e^{-\theta}) = n\bar{y}\log\theta - n\theta - n\log(1-e^{-\theta})$$

Score:

$$U(\theta) = l'(\theta) = \frac{n\bar{y}}{\theta} - n - n\frac{e^{-\theta}}{1-e^{-\theta}}$$

$U(\widehat{\theta}) = 0$ cannot be solved to give an explicit expression for the MLE of θ — try solving it! The observed information is

$$J(\theta) = -l''(\theta) \qquad \text{where} \qquad l''(\theta) = -\frac{n\bar{y}}{\theta^2} + n\frac{e^{-\theta}}{(1-e^{-\theta})^2}.$$

The expected information is

$$I(\theta) = \mathrm{E}(J(\theta)) = -\mathrm{E}(U') = n\left(\frac{1}{\theta(1-e^{-\theta})} - \frac{e^{-\theta}}{(1-e^{-\theta})^2}\right) = \frac{n(1-e^{-\theta}-\theta e^{-\theta})}{\theta(1-e^{-\theta})^2}.$$

A data set has $n = 2423$ observations with a sample mean of $\bar{y} = \frac{3663}{2423} = 1.511762$.

Taking the initial value as $\theta_0 = \bar{y}$, we have for Newton-Raphson,

r	θ_r	$U(\theta_r)$	$J(\theta_r)$
0	1.511762	−685.487	723.349
1	0.564105	873.369	4095.485
2	0.777357	228.246	2247.976
3	0.878891	24.150	1799.634
4	0.892310	0.325	1751.489
5	0.892496	0.000	1750.837

and, for Fisher's methods of scoring,

r	θ_r	$U(\theta_r)$	$I(\theta_r)$
0	1.511762	−685.487	1176.784
1	0.929254	−62.082	1696.292
2	0.892655	−0.279	1750.591
3	0.892496	0.000	1750.837

Both iterative methods yield the MLE of θ as $\widehat{\theta} = 0.892496$. The estimated standard error is $\text{ESE}(\widehat{\theta}) = I^{-1/2}(\widehat{\theta}) = \frac{1}{\sqrt{1750.837}} = 0.024$. In this example the observed and expected information are very similar. Why?

Although Newton-Raphson converges very quickly when near the optimum, scoring usually converges for a wider range of initial values θ_0. If $J(\theta)$ does not involve the random sample Newton-Raphson and scoring are the same.

3.5 Fisher's method of scoring: multiparameter case

Newton-Raphson iterative formula:

$$\boldsymbol{\theta}_{r+1} = \boldsymbol{\theta}_r - [H(\boldsymbol{\theta}_r)]^{-1}U(\boldsymbol{\theta}_r) = \boldsymbol{\theta}_r + J^{-1}(\boldsymbol{\theta}_r)U(\boldsymbol{\theta}_r) \qquad (r = 0, 1, \ldots)$$

Fisher's method of scoring iterative formula:

$$\boldsymbol{\theta}_{r+1} = \boldsymbol{\theta}_r - [E(H(\boldsymbol{\theta}_r))]^{-1}U(\boldsymbol{\theta}_r) = \boldsymbol{\theta}_r + I^{-1}(\boldsymbol{\theta}_r)U(\boldsymbol{\theta}_r) \qquad (r = 0, 1, \ldots)$$

Example:

- $y_1, \ldots, y_n \overset{iid}{\sim}$ Cauchy(α, β) with p.d.f.

$$f(y; \alpha, \beta) = \frac{\beta}{\pi\{\beta^2 + (y-\alpha)^2\}} \qquad (-\infty < y < \infty; \beta > 0).$$

$p = 2$ and $\boldsymbol{\theta} = (\alpha, \beta)^T$.

Log likelihood:

$$l(\alpha, \beta) = n \log \beta - \sum_{i=1}^{n} \log\{\beta^2 + (y_i - \alpha)^2\}$$

Score vector:

$$U(\alpha, \beta) = \begin{pmatrix} \dfrac{\partial l}{\partial \alpha} \\ \dfrac{\partial l}{\partial \beta} \end{pmatrix} = \begin{pmatrix} 2\displaystyle\sum_{i=1}^{n} \frac{y_i - \alpha}{\beta^2 + (y_i - \alpha)^2} \\ \dfrac{n}{\beta} - \displaystyle\sum_{i=1}^{n} \frac{2\beta}{\beta^2 + (y_i - \alpha)^2} \end{pmatrix}$$

Hessian matrix:

$$H(\alpha, \beta) = \begin{pmatrix} \dfrac{\partial^2 l}{\partial \alpha^2} & \dfrac{\partial^2 l}{\partial \alpha \partial \beta} \\ \dfrac{\partial^2 l}{\partial \alpha \partial \beta} & \dfrac{\partial^2 l}{\partial \beta^2} \end{pmatrix} = \begin{pmatrix} 2\displaystyle\sum_{i=1}^{n} \frac{(y_i - \alpha)^2 - \beta^2}{(\beta^2 + (y_i - \alpha)^2)^2} & -4\beta \displaystyle\sum_{i=1}^{n} \frac{y_i - \alpha}{(\beta^2 + (y_i - \alpha)^2)^2} \\ -4\beta \displaystyle\sum_{i=1}^{n} \frac{y_i - \alpha}{(\beta^2 + (y_i - \alpha)^2)^2} & -\dfrac{n}{\beta^2} + 2\displaystyle\sum_{i=1}^{n} \frac{\beta^2 - (y_i - \alpha)^2}{(\beta^2 + (y_i - \alpha)^2)^2} \end{pmatrix}.$$

Observed information matrix: $J(\alpha, \beta) = -H(\alpha, \beta)$

Expected information matrix is given as $I(\alpha, \beta) = -\mathrm{E}[H(\alpha, \beta)] = \frac{n}{2\beta^2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

Thus, for large samples, the variance-covariance matrix of $\widehat{\boldsymbol{\theta}}$ is

$$\mathrm{var}(\widehat{\boldsymbol{\theta}}) = I^{-1}(\boldsymbol{\theta}) = \frac{2\beta^2}{n} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Estimated standard errors: $\mathrm{ESE}(\widehat{\alpha}) = \mathrm{ESE}(\widehat{\beta}) = \frac{\sqrt{2}\widehat{\beta}}{\sqrt{n}}$ Also $\mathrm{cov}(\widehat{\alpha}, \widehat{\beta}) = 0$.

Asymptotic distribution of MLE: $\widehat{\boldsymbol{\theta}} \dot{\sim} \mathrm{BVN}\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \begin{pmatrix} \frac{2\beta^2}{n} & 0 \\ 0 & \frac{2\beta^2}{n} \end{pmatrix} \right)$.

This is the distribution on which we would base a Wald test of $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, i.e.

$$H_0 : \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix}$$

A random sample is available: $y_1 = 1.09, y_2 = -0.23, y_3 = 0.79, y_4 = 2.31, y_5 = -0.81$, and yields MLEs $\widehat{\alpha} = 0.728$ and $\widehat{\beta} = 0.718$.

4 Simple linear regression

4.1 Motivation

Let's consider the following example with data readily available from R. The data are about eruptions of the Old Faithful Geyser, Yellowstone National Park, Wyoming, US. Below we plot durations of eruption 'as a function' of waiting times; the figure was produced using the following code:

```
data(faithful)
attach(faithful)
plot(eruptions ~ waiting, pch = 16,
     xlab = "Waiting time between eruptions",
     ylab = "Duration of eruption")
abline(lm(eruptions ~ waiting), lwd = 3, col = "red")
```

Figure 5: Durations of eruption vs waiting times.

The goal of this section will be on developing methods that can be used to fit a (regression) line to these data, such as the one in red above. The main `R` command to put the theory below into practice is `lm`.

4.2 Assumptions

The assumptions of the simple linear regression model are that, given the values $x_1, ..., x_n$ of an explanatory variable x, the random variables $Y_1, ..., Y_n$

(a) are uncorrelated,

(b) have common variance σ^2,

(c) have expectations of the form $\mathrm{E}(Y_i \,|\, x_i) = \beta_0 + \beta_1 \, x_i$ $(i = 1, \ldots, n)$,

where β_0, β_1 and σ are unknown parameters. Here y is called the *response* (or *dependent*) variable and x is called the *explanatory* (or *predictor, regressor* or *independent*) variable.

4.3 Least squares estimation

For a given set of data (x_i, y_i) $(i = 1, \ldots, n)$, the *Method of Least Squares* estimates the unknown parameters β_0 and β_1 in $\mathrm{E}(Y_i \,|\, x_i)$ by the values of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ which minimize the sum of squares

$$Q = \sum_{i=1}^{n} \{y_i - \mathrm{E}(Y_i \,|\, x_i)\}^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 \, x_i)^2 \; . \tag{2.1.1}$$

One reason for minimizing Q (and not, say, the sum of moduli $\sum |y_i - \beta_0 - \beta_1 x_i|$) is the connection with maximum likelihood (ML) estimation: if the Y_i are Normally distributed as well as satisfying (a), (b) and (c) then the ML estimates minimize Q.

To minimize Q with respect to β_0 and β_1, we can solve for β_0 and β_1 the pair of equations given by $\frac{\partial Q}{\partial \beta_0} = 0$ and $\frac{\partial Q}{\partial \beta_1} = 0$: the Least Squares (LS) estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ thus satisfy

$$\sum_i (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 \, x_i) \;\; = \;\; 0 \, , \tag{2.1.2}$$

$$\sum_i (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 \, x_i) \, x_i \;\; = \;\; 0 \, . \tag{2.1.3}$$

Equation (2.1.2) gives

$$\sum_i y_i - n \, \widehat{\beta}_0 - \widehat{\beta}_1 \sum_i x_i = 0$$

or, dividing by n and rearranging,

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \, \overline{x} \, . \tag{2.1.4}$$

Subtracting $\overline{x}$ times (2.1.2) from (2.1.3) gives

$$\sum_i (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 \, x_i)(x_i - \overline{x}) = 0$$

or

$$\widehat{\beta}_1 \sum_i (x_i - \overline{x}) \, x_i = \sum_i (x_i - \overline{x}) \, y_i$$

(using $\sum (x_i - \overline{x}) = 0$), so that

$$\widehat{\beta}_1 = \sum_i (x_i - \overline{x}) \, y_i \, / \sum_i (x_i - \overline{x}) \, x_i \; . \tag{2.1.5}$$

We introduce a more convenient notation for sums of squares and products as follows:

$$S_{xx} = \sum(x_i - \overline{x})^2 \equiv \sum(x_i - \overline{x})x_i \equiv \sum x_i^2 - n\,\overline{x}^2 \equiv \sum x_i^2 - \left(\sum x_i\right)^2/n, \tag{2.1.6}$$

$$S_{xy} = \sum(x_i - \overline{x})(y_i - \overline{y}) \equiv \sum(x_i - \overline{x})y_i \equiv \sum x_i y_i - n\,\overline{x}\,\overline{y} \equiv \sum x_i y_i - \sum x_i \sum y_i/n \tag{2.1.7}$$

(with S_{yy} defined in a similar way). Then (2.1.5) becomes simply

$$\widehat{\beta}_1 = S_{xy} \,/\, S_{xx}\,. \tag{2.1.8}$$

The estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ can be regarded as the values taken by random variables called *estimators*; these are also denoted by $\widehat{\beta}_0$ and $\widehat{\beta}_1$ but are based on $Y_1, \ldots, Y_n$ rather than $y_1, \ldots, y_n$. Thus we have

$$\widehat{\beta}_1 = \textstyle\sum_i (x_i - \overline{x})Y_i \,/ S_{xx} \qquad \text{and} \qquad \widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1\,\overline{x}\,. \tag{2.1.9}$$

The chunk of code below revists the Old Faithful data and shows that both `lm` and the expressions we have just derived yield the same results.

```
y <- eruptions
x <- waiting
lm(y ~ x)

##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##    -1.87402      0.07563

## Using the expressions we have just derived
Sxx <- sum((x - mean(x))^2)
Sxy <- sum((x - mean(x)) * (y - mean(y)))
beta1 <- Sxy / Sxx
beta0 <- mean(eruptions) - mean(waiting) * beta1

round(beta0, 5)

## [1] -1.87402

round(beta1, 5)

## [1] 0.07563
```

4.4 Some properties of the least squares estimators

1. *Linearity:* The LS estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are *linear* functions of the responses $y_1, ..., y_n$ since $\widehat{\beta}_0$ and $\widehat{\beta}_1$ may be expressed as $\sum_i c_i\, y_i$ and $\sum_i d_i\, y_i$ with coefficients given by

$$c_i = \frac{1}{n} - \frac{\overline{x}\,(x_i - \overline{x})}{S_{xx}}\,, \quad d_i = \frac{x_i - \overline{x}}{S_{xx}}\,. \tag{2.2.1}$$

2. *Normality:* If the responses $Y_1, ..., Y_n$ are Normally distributed then $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are jointly Normal because they are linear functions of the Y_i.

3. *Expectation:* Since expectation is a linear operator and $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are linear functions of the Y_i, we have (using $\mathbf{x}$ to denote the vector $[x_1 \dots x_n]^T$)

$$\begin{aligned}
\mathrm{E}(\widehat{\beta}_1 \,|\, \mathbf{x}) &= \mathrm{E}\left[S_{xx}^{-1} \sum (x_i - \overline{x})\, Y_i \,|\, \mathbf{x}\right] \qquad \text{(from (2.1.9))} \\
&= S_{xx}^{-1} \sum (x_i - \overline{x})\, \mathrm{E}(Y_i \,|\, x_i) \\
&= S_{xx}^{-1} \sum (x_i - \overline{x})\, (\beta_0 + \beta_1 x_i) \\
&= \beta_1 \qquad \text{(using (2.1.6) and } \sum (x_i - \overline{x}) = 0)\,,
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}(\widehat{\beta}_0 \,|\, \mathbf{x}) &= \mathrm{E}\left[\sum \left\{n^{-1} - \overline{x}\, S_{xx}^{-1}\, (x_i - \overline{x})\right\} Y_i \,|\, \mathbf{x}\right] \qquad \text{(from (2.1.4))} \\
&= \sum \left\{n^{-1} - \overline{x}\, S_{xx}^{-1}\, (x_i - \overline{x})\right\}\, (\beta_0 + \beta_1\, x_i) \\
&= n^{-1} \sum \beta_0 - \overline{x}\, S_{xx}^{-1} \sum (x_i - \overline{x}) \beta_0 + n^{-1} \beta_1 \sum x_i - \overline{x}\, \beta_1 \\
&= \beta_0\,.
\end{aligned}$$

   Because $\widehat{\beta}_0$ and $\widehat{\beta}_1$ have expectations equal to the corresponding parameters, they are said to be *unbiased*.

4. *Variances and Covariances:* We next use the results that if $Y_1, ..., Y_n$ are uncorrelated random variables and $c_1, \dots, c_n$, $d_1, \dots, d_n$ are non-random then

$$\mathrm{var}\left(\sum c_i\, Y_i\right) = \sum c_i^2\ \mathrm{var}(Y_i)\,, \qquad \mathrm{cov}\left(\sum c_i\, Y_i, \sum d_i\, Y_i\right) = \sum c_i\, d_i\ \mathrm{var}(Y_i)\,.$$

Applying these results to the coefficients given in (2.2.1) yields:

$$\mathrm{var}\left(\widehat{\beta}_1 \,|\, \mathbf{x}\right) = \sum \left(\frac{x_i - \overline{x}}{S_{xx}}\right)^2 \sigma^2 = S_{xx}^{-1}\, \sigma^2\,, \tag{2.2.2}$$

$$\mathrm{var}\left(\widehat{\beta}_0 \,|\, \mathbf{x}\right) = \sum \left\{\frac{1}{n} - \frac{\overline{x}\,(x_i - \overline{x})}{S_{xx}}\right\}^2 \sigma^2 = \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right) \sigma^2\,, \tag{2.2.3}$$

$$\mathrm{cov}\left(\widehat{\beta}_0, \widehat{\beta}_1 \,|\, \mathbf{x}\right) = \sum \left\{\frac{1}{n} - \frac{\overline{x}(x_i - \overline{x})}{S_{xx}}\right\} \frac{x_i - \overline{x}}{S_{xx}}\, \sigma^2 = -\frac{\overline{x}}{S_{xx}}\, \sigma^2. \tag{2.2.4}$$

Also, putting $c_i = n^{-1}$,

$$\mathrm{cov}\left(\overline{Y}, \widehat{\beta}_1 \,|\, \mathbf{x}\right) = \sum \frac{1}{n}\, \frac{x_i - \overline{x}}{S_{xx}}\, \sigma^2 = 0\,. \tag{2.2.5}$$

4.5 An Alternative Formulation of Simple Linear Regression

Instead of writing the expected response in a linear regression model in the form

$$\mathrm{E}(Y_i \,|\, x_i) = \beta_0 + \beta_1\, x_i\,,$$

it is sometimes useful to express x_i as a deviation from the mean $\overline{x}$ and write

$$\mathrm{E}(Y_i \,|\, x_i) = \gamma + \beta_1(x_i - \overline{x}) \quad (i = 1, \ldots, n)\,, \tag{2.3.1}$$

so that γ is the expected response at $x = \overline{x}$, rather than at $x = 0$. Hence γ equals $\beta_0 + \beta_1 \overline{x}$. The LS estimates $\widehat{\gamma}$ of γ and $\widehat{\beta}_1$ of β_1 minimize the sum of squares

$$Q \;=\; \sum_{i=1}^{n} \{y_i - \mathrm{E}(Y_i \,|\, x_i)\}^2 \;=\; \sum_{i=1}^{n} \{y_i - \gamma - \beta_1(x_i - \overline{x})\}^2\,, \tag{2.3.2}$$

and therefore satisfy

$$\sum \{y_i - \widehat{\gamma} - \widehat{\beta}_1(x_i - \overline{x})\} \;=\; 0\,, \tag{2.3.3}$$

$$\sum \{y_i - \widehat{\gamma} - \widehat{\beta}_1(x_i - \overline{x})\}\,(x_i - \overline{x}) \;=\; 0\,, \tag{2.3.4}$$

instead of (2.1.2) and (2.1.3). Hence (2.3.3) and (2.3.4) give respectively

$$\widehat{\gamma} = \overline{y}\,, \qquad \widehat{\beta}_1 = S_{xy} \,/\, S_{xx}\,. \tag{2.3.5}$$

One advantage of this formulation is that the estimate for γ is simpler than (2.1.4); another is that the estimators $\widehat{\gamma} = \overline{Y}$ and $\widehat{\beta}_1$ are uncorrelated (from (2.2.5)), so that

$$\mathrm{var}\,(\widehat{\gamma} \,|\, \mathbf{x}) = n^{-1}\,\sigma^2\,, \qquad \mathrm{cov}(\widehat{\gamma}, \widehat{\beta}_1 \,|\, \mathbf{x}) = 0\,. \tag{2.3.6}$$

This is useful for finding the variance of the estimated response $\widehat{\beta}_0 + \widehat{\beta}_1\, x_*$ at a value x_* of x. The estimated response may also be written as $\widehat{\gamma} + \widehat{\beta}_1(x_* - \overline{x})$, so its variance is expressible as

$$\mathrm{var}\left(\widehat{\gamma} + \widehat{\beta}_1(x_* - \overline{x}) \,|\, \mathbf{x}\right) = \mathrm{var}\,(\widehat{\gamma} \,|\, \mathbf{x}) + \mathrm{var}\left(\widehat{\beta}_1 \,|\, \mathbf{x}\right)(x_* - \overline{x})^2 = \sigma^2 \left\{\frac{1}{n} + \frac{(x_* - \overline{x})^2}{S_{xx}}\right\}. \tag{2.3.7}$$

```
y <- eruptions
x <- waiting
xdemean <- x - mean(x)
lm(y ~ xdemean)

##
## Call:
## lm(formula = y ~ xdemean)
##
## Coefficients:
## (Intercept)      xdemean
##     3.48778      0.07563

## Using the expressions we have just derived
Sxx <- sum((x - mean(x))^2)
Sxy <- sum((x - mean(x)) * (y - mean(y)))
beta1 <- Sxy / Sxx
gamma <- mean(eruptions)

round(gamma, 5)

## [1] 3.48778

round(beta1, 5)

## [1] 0.07563
```

4.6 Residual and Regression Sums of Squares

An important statistic in simple linear regression is the minimum value of the sum of squares Q in (2.1.1) or (2.3.2), which may be written

$$\sum (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 \qquad \text{or} \qquad \sum \left\{ y_i - \overline{y} - \widehat{\beta}_1 (x_i - \overline{x}) \right\}^2 , \tag{2.4.1}$$

since $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are (by definition) the values of β_0 and β_1 minimizing Q. This is called the *residual sum of squares* (*RSS*). It measures that part of the variation in the responses $y_1, ..., y_n$ which is *not* accounted for by the regression model.

Corresponding to the ith response, y_i, there is a *fitted value,*

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \,, \tag{2.4.2}$$

which is the LS estimate of the ith expected response, $\mathrm{E}(Y_i \,|\, x_i)$, under the model, and a *residual*,

$$\begin{aligned} e_i &= y_i - \widehat{y}_i \\ &= y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \\ &= (y_i - \overline{y}) - \widehat{\beta}_1 (x_i - \overline{x}). \end{aligned} \tag{2.4.3}$$

The residuals can be used for checking the assumptions of the model: see subsection 4.9.

Note that the residuals for simple linear regression satisfy the two constraints

$$\sum e_i = 0 \qquad \text{and} \qquad \sum x_i \, e_i = S_{xy} - \widehat{\beta}_1 \, S_{xx} = 0 \,. \tag{2.4.4}$$

The relation $\widehat{\beta}_1 = S_{xy}/S_{xx}$ (from (2.1.8)) can be used to derive various equivalent expressions for the residual sum of squares:

$$\begin{aligned} RSS &= \sum e_i^2 \\ &= \sum \{ (y_i - \overline{y}) - \widehat{\beta}_1 (x_i - \overline{x}) \}^2 \qquad \text{(using (2.4.3))} \\ &= S_{yy} - 2\,\widehat{\beta}_1 \, S_{xy} + \widehat{\beta}_1^{\,2} \, S_{xx} \\ &= S_{yy} - \widehat{\beta}_1 \, S_{xy} && (2.4.5) \\ &= S_{yy} - \widehat{\beta}_1^{\,2} \, S_{xx} && (2.4.6) \\ &= S_{yy} - {S_{xy}}^2 / S_{xx} && (2.4.7) \end{aligned}$$

Equations (2.4.5), (2.4.6) and (2.4.7) express the residual sum of squares as the difference between the total sum of squares about the mean, S_{yy}, and the *regression sum of squares*, which is given by $\widehat{\beta}_1 \, S_{xy}$, $\widehat{\beta}_1^{\,2} \, S_{xx}$ or S_{xy}^2/S_{xx}. The regression sum of squares represents the amount of variation in the y_i which *is* accounted for (or *explained*) by the regression model.

The proportion of the total sum of squares explained by the model is denoted by R^2 and sometimes called the *coefficient of determination.* For *simple linear regression*,

$$R^2 \;=\; \frac{\text{regression SS}}{\text{total SS about } \overline{y}} \;=\; \frac{S_{xy}^2/S_{xx}}{S_{yy}} \;=\; \frac{S_{xy}^2}{S_{xx} S_{yy}} \;=\; r^2 \,, \tag{2.4.8}$$

where $r = S_{xy}/\sqrt{S_{xx} S_{yy}}$ is the sample (Pearson) correlation between the x's and y's. The chunk of code below can be used to obtain the coefficient of determination from R on the Old Faithful data.

```
y <- eruptions
x <- waiting
fit <- lm(y ~ x)
summary(fit)$r.squared

## [1] 0.8114608
```

It will be seen later that (subsection 4.8), under the model defined at the beginning of §4, the RSS has expectation $(n-2)\sigma^2$ and $n-2$ degrees of freedom. To estimate σ^2, we use the *residual mean square*, defined by

$$\widehat{\sigma}^2 = \frac{RSS}{n-2}\,, \tag{2.4.9}$$

which is an unbiased estimator of σ^2.

```
data(faithful)
attach(faithful)
y <- eruptions
x <- waiting
n <- length(eruptions)
fit <- lm(y ~ x)
rss <- sum(fit$residuals^2)
sigma2 <- rss / (n - 2)
```

4.7 Analysis of Variance in Simple Linear Regression

Many types of statistical analysis may be summarized in an *analysis of variance table* (or *ANOVA table*). Such tables appear, for example, in the `anova` output from R's regression `lm` function. A more descriptive term than 'analysis of variance' might be 'analysis of sums of squares' or 'decomposition of sums of squares'. For simple linear regression, the analysis of variance table adds nothing to the information in the residual mean square, the least squares estimates and their standard errors, but this type of table is useful for summarizing more complicated analyses.

Analysis of variance tables provide a convenient way to display a decomposition of the total sum of squares of the responses (usually the total sum of squares *about the mean*). We can measure the variation in a set of observed responses $y_1, \ldots, y_n$ by using their sample variance, or equivalently the sum of squares about the mean $\overline{y}$, which is

$$S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2\,.$$

This sum of squares can be regarded as the minimum with respect to β_0 of $\sum(y_i - \beta_0)^2$, the minimum being attained when β_0 equals $\overline{y}$. Hence S_{yy} is the residual sum of squares under the simple model

$$\mathrm{E}(Y_i \,|\, x_i) = \beta_0 \quad (i = 1, ..., n)\,. \tag{2.5.1}$$

Model (2.5.1) ignores any dependence of $\mathrm{E}(Y_i \,|\, x_i)$ on the explanatory variable x. If we generalize this model to include dependence on x_i, using the simple linear regression

$$\mathrm{E}(Y_i \,|\, x_i) = \beta_0 + \beta_1\, x_i \quad (i = 1, ..., n)\,, \tag{2.5.2}$$

then, from §4.6, the minimum sum of squares is the residual sum of squares,

$$\sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 \;=\; S_{yy} - S_{xy}^2/S_{xx}\,.$$

The term S_{xy}^2/S_{xx} is the regression sum of squares (or regression SS), as defined in §4.6, and is equivalent to $\widehat{\beta}_1 S_{xy}$ and $\widehat{\beta}_1^2 S_{xx}$. It equals the reduction in the residual sum of squares due to including $\beta_1\, x_i$ as well as β_0 in $\mathrm{E}(Y_i \,|\, x_i)$. Hence we have the decomposition

$$S_{yy} \qquad \equiv \quad S_{xy}^2/S_{xx} \quad + \quad S_{yy} - S_{xy}^2/S_{xx} \tag{2.5.3}$$

or

$$\text{total SS about } \overline{y} \quad \equiv \quad \text{regression SS} \quad + \quad \text{residual SS}\,.$$

The form of an analysis of variance table for simple linear regression based on n responses is shown below. [Note that the 'Total' sum of squares is calculated about the mean response $\overline{y}$.] The explanation of the columns in the table is as follows.

- The column headed 'SS' shows the decomposition given in (2.5.3) of the total sum of squares about the mean $\overline{y}$ into the regression and residual sums of squares.
- The 'DF' column gives the *degrees of freedom* corresponding to these sum of squares.
- The entries in the 'MS' column are the regression and residual *mean squares*, given by the sums of squares divided by their degrees of freedom.
- The value in the 'F' column denotes the ratio of the regression mean square to the residual mean square. This statistic can be used to test whether the expected response depends on the explanatory variable. For simple linear regression, it equals the square of the t-statistic formed by dividing $\widehat{\beta}_1$ by its estimated standard error $\sqrt{\widehat{\sigma}^2/S_{xx}}$.
- The column headed 'p' gives the significance probability in the above test.

SOURCE	DF	SS	MS	F	p
Regression	1	S_{xy}^2/S_{xx}	S_{xy}^2/S_{xx}	$S_{xy}^2/(S_{xx}\,\widehat{\sigma}^2)$	
Residual	$n-2$	$S_{yy} - S_{xy}^2/S_{xx}$	$\widehat{\sigma}^2$		
Total	$n-1$	S_{yy}			

4.8 Inferences about β_1 and Expected and Future Responses when the Responses are Normally Distributed

We now add an extra assumption to the three made at Section 4.2:

(d) given $x_1, \ldots, x_n$, the responses $Y_1, ..., Y_n$ are Normally distributed.

Assumptions (a) to (d) imply that (given $x_1, \ldots, x_n$) $Y_1, \ldots, Y_n$ are independent with distributions $N(\beta_0 + \beta_1\, x_i, \sigma^2)$ $(i = 1, ..., n)$. It will be seen later that this implies that

(i) $\widehat{\beta}_1$ has the distribution $N(\beta_1,\, \sigma^2/S_{xx})$;

(ii) $\overline{Y}$ has the distribution $N(\beta_0 + \beta_1\, \overline{x}, \sigma^2/n)$ independently of $\widehat{\beta}_1$;

(iii) $RSS/\sigma^2 = (n-2)\widehat{\sigma}^2/\sigma^2$ has the distribution $\chi^2(n-2)$ independently of $\widehat{\beta}_1$ and $\overline{Y}$;

(iv) $(\widehat{\beta}_1 - \beta_1) \Big/ \sqrt{\sigma^2/S_{xx}}$ has the distribution $N(0,1)$;

(v) $(\widehat{\beta}_1 - \beta_1) \Big/ \sqrt{\widehat{\sigma}^2/S_{xx}}$ has the Student's distribution $t(n-2)$.

4.8.1 Inferences about β_1

To test a hypothesis that β_1 takes a specified value β_{10}, say, we can compare the value of

$$t = (\widehat{\beta}_1 - \beta_{10}) \Big/ \sqrt{\widehat{\sigma}^2/S_{xx}} \tag{2.6.1}$$

with Student's distribution $t(n-2)$. Depending on the alternative hypothesis, we might look for large values of t, $-t$ or $|t|$. Let's see how to do this in R recycling the code from above.

```
data(faithful)
attach(faithful)
y <- eruptions
x <- waiting
n <- length(eruptions)
fit <- lm(y ~ x)
Sxx <- sum((x - mean(x))^2)
rss <- sum(fit$residuals^2)
sigma2 <- rss / (n - 2)

beta1 <- fit$coefficients[2]
beta1 / sqrt(sigma2 / Sxx)

##        x
## 34.08904

summary(fit)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.29917 -0.37689  0.03508  0.34909  1.19329
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.874016   0.160143  -11.70   <2e-16 ***
## x            0.075628   0.002219   34.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4965 on 270 degrees of freedom
## Multiple R-squared:  0.8115,Adjusted R-squared:  0.8108
## F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

A 95% confidence interval for β_1 is given by

$$\widehat{\beta}_1 \pm t_{0.025} \sqrt{\widehat{\sigma}^2 / S_{xx}}\,, \tag{2.6.2}$$

where $t_{0.025}$ denotes the upper 2.5% point of $t(n-2)$. Here is a way of constructing such intervals in R recycling the code above

```
t <- qt(0.025, n - 2)
fit$coefficients[2] + t * sqrt(sigma2 / Sxx)

##          x
## 0.07126011

fit$coefficients[2] - t * sqrt(sigma2 / Sxx)

##          x
## 0.07999579

## R built-in function yields
confint(fit)

##                    2.5 %      97.5 %
## (Intercept) -2.18930436 -1.55872761
## x            0.07126011  0.07999579
```

4.8.2 Inferences about an expected response

The *expected response* under a simple linear regression model when x takes a value x_* is

$$\mathrm{E}(Y \,|\, x_*) = \beta_0 + \beta_1\, x_*\,;$$

the least squares estimator of this expectation is

$$\widehat{\beta}_0 + \widehat{\beta}_1\, x_* = \overline{Y} + \widehat{\beta}_1 (x_* - \overline{x})\,, \tag{2.6.3}$$

which has variance $\sigma^2 \left\{ n^{-1} + (x_* - \overline{x})^2 / S_{xx} \right\}$, from (2.3.7).

A 95% confidence interval for $\mathrm{E}(Y \,|\, x_*)$ is therefore

$$\widehat{\beta}_0 + \widehat{\beta}_1\, x_* \pm t_{0.025}\, \widehat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_* - \overline{x})^2}{S_{xx}}}\,. \tag{2.6.4}$$

4.8.3 Inferences about a future response

The *future response* Y_* to be observed at a value x_* of x is a random variable which may be expressed as

$$Y_* = \beta_0 + \beta_1 x_* + \varepsilon\ ,$$

where ε has the distribution $N(0, \sigma^2)$ independently of $Y_1, ..., Y_n$ and hence of $\widehat{\beta}_0$ and $\widehat{\beta}_1$. We use the same estimator (given by (2.6.3)) for Y_* as for the expected response $\beta_0 + \beta_1\, x_*$ in §4.6.2, but now consider the variance of the *difference* between the estimator and the response,

$$\widehat{\beta}_0 + \widehat{\beta}_1\, x_* - Y_*\,.$$

Using (2.3.5) and the independence of $\widehat{\beta}_0 + \widehat{\beta}_1 x_*$ and Y_*, we have

$$\mathrm{var}\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_* - Y_* \,|\, \mathbf{x}, x_*\right) \;=\; \sigma^2 \left\{ \frac{1}{n} + \frac{(x_* - \overline{x})^2}{S_{xx}} \right\} + \sigma^2 = \sigma^2 \left\{ 1 + \frac{1}{n} + \frac{(x_* - \overline{x})^2}{S_{xx}} \right\}; \tag{2.6.5}$$

the extra σ^2 (relative to (2.3.7)) comes from $\mathrm{var}(Y_* \,|\, x_*)$.

A 95% *prediction interval* for Y_* is thus

$$\widehat{\beta}_0 + \widehat{\beta}_1 x_* \;\pm\; t_{0.025}\, \widehat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_* - \overline{x})^2}{S_{xx}}}\,. \tag{2.6.6}$$

4.9 Analysis of Residuals

The *residuals* $e_1, \ldots, e_n$ from the least squares fit of a simple linear regression were defined in equation (2.4.3). They are used for checking whether the assumptions of the model are valid.

For a simple linear regression model, if we define the random variables

$$\varepsilon_i \;=\; Y_i - \mathrm{E}(Y_i \,|\, x_i) \;=\; Y_i - \beta_0 - \beta_1 x_i \quad (i = 1, \ldots, n)$$

then, given the x_i and assumptions (a) to (c) at the beginning of §2,

(a) the ε_i are uncorrelated, i.e. $\mathrm{cov}(\varepsilon_i, \varepsilon_j \,|\, \mathbf{x}) = 0 \quad (i \neq j)$,

(b) the ε_i have common variance σ^2, i.e. $\mathrm{var}(\varepsilon_i \,|\, \mathbf{x}) = \sigma^2$,

(c) $\mathrm{E}(\varepsilon_i \,|\, \mathbf{x}) = 0$,

(d) if the Y_i are Normal, so are the ε_i (in fact $N(0, \sigma^2)$ independently).

The ε_i cannot be calculated because they depend on the unknown values of β_0 and β_1, but we *can* evaluate the residuals defined in (2.4.3) as

$$e_i \;=\; y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \;=\; y_i - \overline{y} - \widehat{\beta}_1 (x_i - \overline{x}) \quad (i = 1, \ldots, n)\,.$$

The corresponding random variables

$$E_i = Y_i - \overline{Y} - \widehat{\beta}_1 (x_i - \overline{x}) \quad (i = 1, \ldots, n) \tag{2.7.1}$$

have similar properties to the ε_i if n is large and none of the x_i is extreme. In particular, if we define

$$p_{ij} = \frac{1}{n} + \frac{(x_i - \overline{x})(x_j - \overline{x})}{S_{xx}} \quad (i, j = 1, ..., n) \tag{2.7.2}$$

then the E_i satisfy

(a′) $\mathrm{cov}(E_i, E_j \,|\, \mathbf{x}) = -p_{ij}\, \sigma^2 \quad (i \neq j)$,

(b′) $\mathrm{var}(E_i \,|\, \mathbf{x}) = (1 - p_{ii})\, \sigma^2 = \left[1 - \frac{1}{n} - \frac{(x_i - \overline{x})^2}{S_{xx}} \right] \sigma^2$,

(c′) $\mathrm{E}(E_i \,|\, \mathbf{x}) = 0$,

(d′) if the Y_i are Normal, so are the E_i since they are linear functions of the Y_i.

Note that (c′) and (d′) are like (c) and (d) above. If n is large and none of the quantities $|x_i - \overline{x}|$ is large then the p_{ij} are small, so that (a′), (b′) approximate to (a), (b); hence the correlations between the E_i are small.

The variances of the residuals differ (depending on $|x_i - \overline{x}|$), so we define *standardized residuals,* r_i , with a common estimated variance of 1 by

$$\begin{aligned} r_i &= \frac{e_i}{\text{estimated standard error of } E_i} \\ &= \frac{e_i}{\sqrt{\widehat{\sigma}^2 \left(1 - \frac{1}{n} - \frac{(x_i - \overline{x})^2}{S_{xx}}\right)}} \quad (i = 1, \ldots, n)\,, \end{aligned} \tag{2.7.3}$$

where $\widehat{\sigma}^2$ is the residual mean square defined in (2.4.9). To examine how well a simple linear regression model fits we can

(i) plot the r_i against the x_i and look for any sign that E $(R_i \,|\, x_i)$ depends on the x_i (indicated by curvature of the plot) or that var$(R_i \,|\, x_i)$ depends on the x_i (e.g. increases with x);

(ii) see if there are any unusually large values of $|r_i|$, say greater than 2.5 or 3.

(iii) make a Normal probability plot of the r_i and look for evidence of non-Normality.

4.10 The 'Normal Linear Model'

Let $\mathbf{Y}$ be a random n-vector of responses, $\mathbf{X}$ an $n \times p$ matrix (with $n > p$) whose elements are known values x_{ij}, and $\boldsymbol{\beta}$ a p-vector of unknown parameters. For the *Normal Linear Model,* we assume

$$\mathrm{E}(\mathbf{Y} \,|\, \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}\,, \tag{3.3.1}$$

$$\mathrm{var}(\mathbf{Y} \,|\, \mathbf{X}) = \sigma^2 \mathbf{I}_n\,. \tag{3.3.2}$$

The n elements $\mathrm{E}(Y_i \,|\, \mathbf{X})$ of $\mathrm{E}(\mathbf{Y} \,|\, \mathbf{X})$, are then given by

$$\mathrm{E}(Y_i \,|\, \mathbf{X}) = \sum_{j=1}^{p} x_{ij}\,\beta_j\,, \tag{3.3.3}$$

which is a *linear* function of the coefficients β_j . Thus the Normal Linear Model is linear in the β's. Even a quadratic regression model with $\mathrm{E}(Y_i \,|\, x_i) = \beta_0 + \beta_1\, x_i + \beta_2\, x_i^2$ is linear in this sense, as is a model $\mathrm{E}(Y_i \,|\, x_i) = \beta_0 + \beta_1\, \ln x_i$.

If the model contains an 'intercept' or 'constant term', such as β_0 in (3.1.1) and (3.2.1) or γ in (3.1.4), this corresponds to a column of 1's in $\mathbf{X}$. If $\mathbf{X}$ has full rank p then the $p \times p$ matrix $\mathbf{X}^T\mathbf{X}$ is non-singular, and the least squares estimates are unique. It is sometimes convenient to consider a model which is not of full rank, and to define the least squares estimates using generalized inverses. For making inferences about $\boldsymbol{\beta}$, we also assume that $\mathbf{Y}$ has a multivariate Normal distribution (given $\mathbf{X}$). The distribution of $\mathbf{Y}$ is then $N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\,\mathbf{I}_n)$.

4.11 Least Squares Estimation

As in Section 4.3, we find the values of the β's which minimize the sum of squares

$$Q = \sum_{i=1}^{n} \{ y_i - \mathrm{E}(Y_i \,|\, \mathbf{X}) \}^2 \tag{3.4.1}$$

for the observed responses $y_1, \ldots, y_n$. In terms of vectors and matrices, the function to be minimized is

$$\begin{aligned} Q &= \{ \mathbf{y} - \mathrm{E}(\mathbf{Y} \,|\, \mathbf{X}) \}^T \{ \mathbf{y} - \mathrm{E}(\mathbf{Y} \,|\, \mathbf{X}) \} \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}^T\mathbf{y} - 2\,\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\,. \end{aligned} \tag{3.4.2}$$

Using matrix results, the vector of partial derivatives of Q with respect to the β's is given by

$$\frac{\partial Q}{\partial \boldsymbol{\beta}} = 2\,\left(\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - \mathbf{X}^T\mathbf{y}\right)\,. \tag{3.4.3}$$

Equating this vector to $\mathbf{0}$, the vector $\widehat{\boldsymbol{\beta}}$ of least squares estimates satisfies the p normal equations

$$\mathbf{X}^T\mathbf{X}\,\widehat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}\,. \tag{3.4.4}$$

These may also be written

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\,\widehat{\boldsymbol{\beta}}) = \mathbf{0}\,. \tag{3.4.5}$$

Note that the jk-th element of $\mathbf{X}^T\mathbf{X}$ and the j-th element of $\mathbf{X}^T\mathbf{y}$ are respectively

$$\sum_{i=1}^{n} x_{ij}\,x_{ik}\,, \quad \sum_{i=1}^{n} x_{ij}\,y_i\,. \tag{3.4.6}$$

If $\mathbf{X}$ has full rank p then $(\mathbf{X}^T\mathbf{X})^{-1}$ exists and there is a unique *least squares estimate* of $\boldsymbol{\beta}$ given by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\,; \tag{3.4.7}$$

otherwise the estimate is not unique. To show that a solution of (3.4.4) gives a minimum of Q, note that

$$\begin{aligned} Q &= \left\{(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) + \mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right\}^T \left\{(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) + \mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right\} \\ &= (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\mathbf{X}^T\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + 2(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\mathbf{X}^T(\mathbf{y} - \mathbf{X}\,\widehat{\boldsymbol{\beta}}) \end{aligned} \tag{3.4.8}$$

The third term is 0 (by (3.4.5)); the first and second terms are non-negative. Hence

$$Q \geq (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})\,, \tag{3.4.9}$$

and the minimum is attained when $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$. The right-hand side of (3.4.9) is called the *residual sum of squares* for the Normal Linear Model, and is considered in §4.7.