

COMPILED BY SHRAVAN VASISHTH

LINEAR MODELING (MSC COGNITIVE SYSTEMS, POTSDAM)

VASISHTH LAB LECTURE NOTES



# *Contents*

<b>1</b>	<i>Preliminaries</i>	<b>9</b>
1.1	<i>What this course is about</i>	9
1.2	<i>Software and source code accompanying these notes</i>	10
1.3	<i>A comment on notation</i>	10
<b>2</b>	<i>Random variables, Expectation and Variance</i>	<b>11</b>
2.1	<i>Discrete random variables</i>	11
2.1.1	<i>Example: The Binomial random variable</i>	12
2.2	<i>Continuous random variables</i>	13
2.2.1	<i>Example: Normal random variable</i>	14
2.3	<i>Expectations and Variances</i>	16
2.3.1	<i>Expectations and variances of discrete RVs</i>	16
2.3.2	<i>Expectations and variances of continuous RVs</i>	16
2.3.3	<i>Example: The expectation and variance of the standard normal RV</i>	17
<b>3</b>	<i>Some useful probability distributions</i>	<b>19</b>
3.0.1	<i>Exponential random variables</i>	19
3.0.2	<i>Weibull distribution</i>	20
3.0.3	<i>Gamma distribution</i>	20
3.0.4	<i>Uniform random variable</i>	23
3.0.5	<i>Beta distribution</i>	23

<b>3.1</b>	<i>Jointly distributed random variables</i>	<b>24</b>
3.1.1	<i>Discrete case</i>	24
3.1.2	<i>Continuous case</i>	25
3.1.3	<i>Marginal probability distribution functions</i>	26
3.1.4	<i>Independent random variables</i>	27
3.1.5	<i>Sums of independent random variables</i>	28
3.1.6	<i>Conditional distributions</i>	28
3.1.7	<i>Joint and marginal expectation</i>	29
3.1.8	<i>Covariance and correlation</i>	30
3.1.9	<i>Conditional expectation</i>	31
<b>4</b>	<i>Maximum Likelihood Estimation</i>	<b>33</b>
4.1	<i>Introduction to Maximum Likelihood Estimation</i>	33
4.1.1	<i>Example 1: MLE of the Binomial distribution</i>	37
4.1.2	<i>Example 2: MLE of the Normal distribution</i>	37
4.1.3	<i>Example 3: MLE of the Exponential distribution</i>	40
4.1.4	<i>Practical implications</i>	40
4.2	<i>Asymptotic properties of MLEs</i>	44
4.2.1	<i>The binomial distribution</i>	47
4.2.2	<i>The normal distribution</i>	51
<b>5</b>	<i>Linear modeling</i>	<b>55</b>
5.1	<i>Basic linear modeling theory</i>	55
5.1.1	<i>Least squares estimation: Geometric argument</i>	56
5.1.2	<i>The expectation and variance of the parameters beta</i>	58
5.1.3	<i>Statistical inference</i>	60
5.1.4	<i>The p-value, and Type S and M errors</i>	62
5.1.5	<i>Hypothesis tests and the sampling distribution of the mean</i>	66
5.1.6	<i>Hypothesis testing using the likelihood ratio</i>	67
5.1.7	<i>Hypothesis testing using Analysis of variance (ANOVA)</i>	70
5.1.8	<i>Multiple regression</i>	71
5.1.9	<i>Checking model assumptions</i>	74
5.1.10	<i>Correcting for multiple testing</i>	81
5.1.11	<i>Transformations: Box-Cox procedure</i>	81

<b>5.2 Generalized Linear Models</b>	<b>84</b>
5.2.1 <i>Introduction: Logistic regression</i>	84
5.2.2 <i>Multiple logistic regression: Example from Hindi data</i>	90
5.2.3 <i>Some theory for GLMs</i>	91
5.2.4 <i>The canonical link</i>	93
5.2.5 <i>Estimation of parameters</i>	93
5.2.6 <i>Deviance</i>	95
5.2.7 <i>Hypothesis testing: Residual deviance</i>	95
5.2.8 <i>Assessing goodness of fit of a fitted model</i>	101
5.2.9 <i>Residuals in GLMs</i>	102
<b>6 Hierarchical linear models: Frequentist and Bayesian approaches</b>	<b>105</b>
6.1 <i>Linear mixed models</i>	105
6.1.1 <i>Informal presentation of LMMs</i>	105
6.1.2 <i>Linear mixed model</i>	112
6.1.3 <i>Some basic types of linear mixed model and their variance components</i>	118
6.1.4 <i>Parameter estimation</i>	122
6.1.5 <i>Computing the BLUPs</i>	123
6.1.6 <i>Correlation of fixed effects</i>	124
6.2 <i>Bayesian data analysis: Some introductory ideas</i>	125
6.2.1 <i>Example 1: Proportions</i>	126
6.2.2 <i>Example 2: Proportions</i>	129
6.2.3 <i>Exercise: The proportion of female births in France</i>	133
6.2.4 <i>The posterior is the weighted mean of the prior mean and the MLE</i>	133
6.2.5 <i>Example: The Poisson-Gamma conjugate case</i>	135
6.2.6 <i>Using JAGS for the Poisson-Gamma conjugate example</i>	136
6.2.7 <i>The posterior in the Poisson-Gamma case as a weighted sum</i>	139
6.2.8 <i>Exercise: Using the posterior as a prior for new data</i>	140
6.3 <i>Fitting Linear Models and Linear Mixed Models in a Bayesian setting</i>	143
<b>7 Bibliography</b>	<b>145</b>



## *Acknowledgements*

Much of the material here is derived from the University of Sheffield lecture notes in the MSc in Statistics. I'm grateful to Lena Jäger and Paul Mätzig for catching numerous errors and unclear paragraphs. Any mistakes are of course mine.



# 1

## Preliminaries

### 1.1 What this course is about

These lecture notes cover the basic theory of linear models. My notes are heavily dependent on the MSc lecture notes in Statistics taught at the University of Sheffield, UK, and on the textbooks mentioned in these notes.<sup>1</sup>

The lecture notes are intended for graduate students in the MSc Cognitive Systems program, but the course can be taken by other students in Linguistics at the University of Potsdam. I assume some basic knowledge of probability theory, calculus, and linear algebra; it is recommended that students take the Introduction to Mathematics taught in winter semester before taking this course. The course presents linear modeling from the statistical, as opposed to the machine learning, perspective; in this respect, it complements the Machine Learning course taught in parallel to this one.

My general philosophy is to try to convey the intuitive idea (graphically if possible), augmented with some proofs. I avoid complex proofs, referring the interested reader to more advanced textbooks. Since this is a graduate level course, I do expect that you will read the relevant chapters from the assigned books, and will do the assigned exercises.

We begin by considering some facts about random variables. Then, we look at how expectation and variance etc. are computed. In subsequent chapters, several typical probability distributions and their properties are discussed. A major topic of interest is maximum likelihood estimation. Then we cover the basic theory of linear models, generalized linear models, and linear mixed models. We close with a tutorial on Bayesian linear modeling.

<sup>1</sup> G. Jay Kerns. *Introduction to Probability and Statistics Using R*. 2010; Sheldon Ross. *A first course in probability*. Pearson Education, 2002; A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, Cambridge, UK, 2007; and Annette J Dobson and Adrian Barnett. *An introduction to generalized linear models*. CRC press, 2011

## 1.2 Software and source code accompanying these notes

Please install RStudio and R on your computer. You can download the source code and data associated with these lecture notes from the course web page:

<http://www.ling.uni-potsdam.de/~vasishth/statistics/LinearModeling.html>

## 1.3 A comment on notation

Throughout, I will define the Normal distribution in terms of  $\mu$  and  $\sigma$ ; this is not standard. In statistics textbooks, you will find that the normal distribution is defined in terms of  $\mu$  and  $\sigma^2$ . The reason I deviate from this convention is that in R the normal distribution is defined in terms of  $\sigma$ .

I will not mark a vector of values (e.g.,  $\beta$ ) any differently from a scalar  $\beta$ ; it will usually be clear from context which is meant.

## 2

# *Random variables, Expectation and Variance*

### *2.1 Discrete random variables*

A random variable  $X$  is a function  $X : S \rightarrow \mathbb{R}$  that associates to each outcome  $\omega \in S$  exactly one number  $X(\omega) = x$ .

$S_X$  is all the  $x$ 's (all the possible values of  $X$ , the support of  $X$ ). I.e.,  $x \in S_X$ .

An example of a **discrete** random variable: number of coin tosses until a heads appears. The number of coin tosses could be 0, 1, 2, .... These are **discrete** because the number of tosses can only be an integer value between zero and infinity; they are not **continuous** because the number of tosses can't be any real number between 0 and infinity: 3.5 is not a possible value.

- $X : \omega \rightarrow x$
- $\omega$ : H, TH, TTH, ... (infinite)
- $x = 0, 1, 2, \dots ; x \in S_X$ . (Note that the function  $X : S \rightarrow \mathbb{R}$  now maps to a subset of  $\mathbb{R}$ , the integers.)

Every discrete random variable  $X$  has associated with it a **probability mass/distribution function (PDF)**, also called **distribution function**.

$$p_X : S_X \rightarrow [0, 1] \quad (2.1)$$

defined by

$$p_X(x) = P(X(\omega) = x), x \in S_X \quad (2.2)$$

[**Note:** Books sometimes abuse notation by overloading the meaning of  $X$ . They usually have:  $p_X(x) = P(X = x), x \in S_X$ ]

The **cumulative distribution function** is

$$F(a) = \sum_{\text{all } x \leq a} p(x) \quad (2.3)$$

### 2.1.1 Example: The Binomial random variable

Suppose that  $n$  independent trials are performed, there are two possible outcomes, success and failure, each with probability  $\theta$  and  $(1 - \theta)$  respectively.

Then, from the binomial theorem, the probability of  $x$  successes out of  $n$  is:

$$P(X = x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x} \quad (2.4)$$

For example, if we toss a coin twice, the probability of one or less successes out of 2 tosses is the sum of

- The probability of 0 successes

$$P(X = 0) = \binom{2}{0} \theta^0 (1 - \theta)^{2-0} = 1 \times (1 - \theta)^2 \quad (2.5)$$

- The probability of 1 success

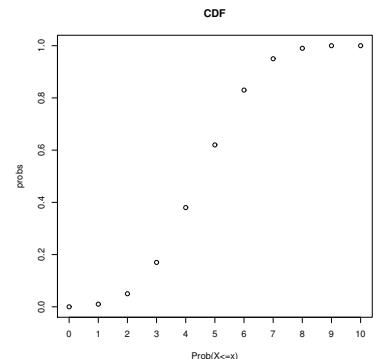
$$P(X = 1) = \binom{2}{1} \theta^1 (1 - \theta)^{2-1} = 1 \times 2 \times \theta (1 - \theta)^1 \quad (2.6)$$

If  $\theta = 0.5$ , we have  $0.5^2 + 2 \times 0.5 \cdot 0.5 = 0.75$ .

This will quickly become cumbersome to do by hand. Consider the case where we have  $n=10$  coin tosses. What's the prob. of 1 or fewer successes? 2 or fewer? We can quickly compute the probability of getting  $x$  or fewer successes where  $x=0$  to 10. For this, we use the built in cumulative distribution function (CDF) function `pbinom`.

```
## sample size
n<-10
## prob of success
p<-0.5
probs<-rep(NA,11)
for(x in 0:10){
  ## Cumulative Distribution Function:
  probs[x+1]<-round(pbinom(x,size=n,prob=p),digits=2)
}
```

```
## Plot the CDF:
plot(1:11,probs,xaxt="n",
      xlab="Prob(X<=x)",
      main="CDF")
axis(1,at=1:11,labels=0:10)
```



The probability of getting exactly 1 success,  $P(X=1)$  can be computed by subtracting the probability of 0 heads using `pbinom` from the probability of getting 1 or 0 heads:

```
pbinom(1,size=10,prob=0.5)-pbinom(0,size=10,prob=0.5)
## [1] 0.009765625
```

Figure 2.1: The CDF of the binomial.

```
choose(10,1) * 0.5 * (1-0.5)^9
## [1] 0.009765625
```

What about the probability density function (PDF)? The built-in function in R for the PDF is `dbinom`:

```
## P(X=0)
dbinom(0,size=10,prob=0.5)

## [1] 0.0009765625
```

To summarize, a discrete random variable  $X$  will be defined by

1. the function  $X : S \rightarrow \mathbb{R}$ , where  $S$  is the discrete set of outcomes (i.e., outcomes are  $\omega \in S$ ).
2.  $X(\omega) = x$ , and  $S_X$  is the **support** of  $X$  (i.e.,  $x \in S_X$ ).
3. A PDF is defined for  $X$ :

$$p_X : S_X \rightarrow [0, 1]$$

4. A CDF is defined for  $X$ :

$$F(a) = \sum_{\text{all } x \leq a} p(x)$$

## 2.2 Continuous random variables

As mentioned above in the discrete case, a random variable  $X$  is a function  $X : S \rightarrow \mathbb{R}$  that associates to each outcome  $\omega \in S$  exactly one number  $X(\omega) = x$ .  $S_X$  is all the  $x$ 's (all the possible values of  $X$ , the support of  $X$ ). I.e.,  $x \in S_X$ .

$X$  is a continuous random variable if there is a non-negative function  $f$  defined for all real  $x \in (-\infty, \infty)$  having the property that for any set  $B$  of real numbers,

$$P\{X \in B\} = \int_B f(x) dx \quad (2.7)$$

$f(x)$  is the probability density function of the random variable  $X$ .

Since  $X$  must assume some value,  $f$  must satisfy

$$1 = P\{X \in (-\infty, \infty)\} = \int_{-\infty}^{\infty} f(x) dx \quad (2.8)$$

If  $B = [a, b]$ , then

$$P\{a \leq X \leq b\} = \int_a^b f(x) dx \quad (2.9)$$

```
## Plot the pdf:
plot(1:11,
      dbinom(0:10,size=10,prob=0.5),
      main="PDF",
      xaxt="n")
axis(1,at=1:11,labels=0:10)
```

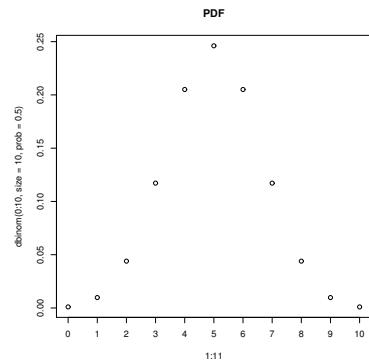


Figure 2.2: The PDF (actually, probability mass function) of the binomial.

If  $a = b$ , we get

$$P\{X = a\} = \int_a^a f(x) dx = 0 \quad (2.10)$$

Hence, for any continuous random variable,

$$P\{X < a\} = P\{X \leq a\} = F(a) = \int_{-\infty}^a f(x) dx \quad (2.11)$$

$F$  is the **cumulative distribution function**. Differentiating both sides in the above equation:

$$\frac{dF(a)}{da} = f(a) \quad (2.12)$$

The density (PDF) is the derivative of the CDF. Ross<sup>1</sup> suggests that it is intuitive to think about it as follows:

$$P\left\{a - \frac{\epsilon}{2} \leq X \leq a + \frac{\epsilon}{2}\right\} = \int_{a-\epsilon/2}^{a+\epsilon/2} f(x) dx \approx \epsilon f(a) \quad (2.13)$$

when  $\epsilon$  is small and when  $f(\cdot)$  is continuous. I.e.,  $\epsilon f(a)$  is the approximate probability that  $X$  will be contained in an interval of length  $\epsilon$  around the point  $a$ .

### 2.2.1 Example: Normal random variable

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty. \quad (2.14)$$

We write  $X \sim \text{norm}(\text{mean} = \mu, \text{sd} = \sigma)$ , and the associated R function for the PDF is `dnorm(x, mean = 0, sd = 1)`, and the one for CDF is `pnorm`.

Note the default values for  $\mu$  and  $\sigma$  as 0 and 1 respectively. Note also that R defines the PDF in terms of  $\mu$  and  $\sigma$ , not  $\mu$  and  $\sigma^2$ .

Computing probabilities using the CDF:

```
pnorm(Inf) - pnorm(-Inf)
## [1] 1

pnorm(2) - pnorm(-2)
## [1] 0.9544997

pnorm(1) - pnorm(-1)
## [1] 0.6826895
```

<sup>1</sup> Sheldon Ross. *A first course in probability*. Pearson Education, 2002

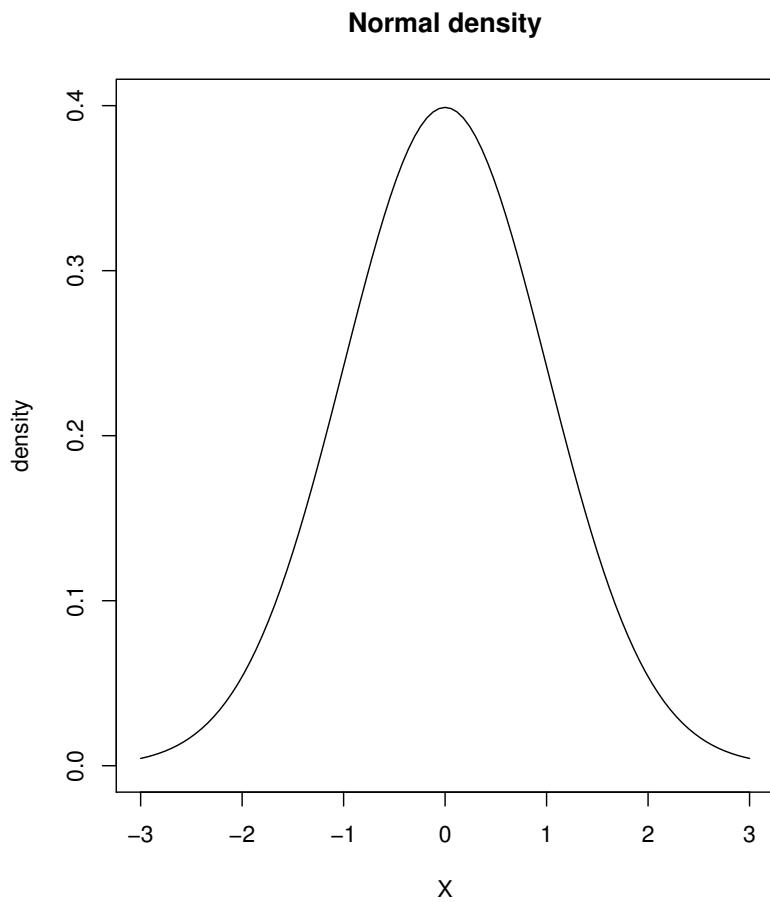


Figure 2.3: Normal distribution.

*Standard or unit normal random variable* If  $X$  is normally distributed with parameters  $\mu$  and  $\sigma$ , then  $Z = (X - \mu)/\sigma$  is normally distributed with parameters  $\mu = 0, \sigma = 1$ .

We conventionally write  $\Phi(a)$  for the CDF:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-y^2/2} dy \quad \text{where } z = (x - \mu)/\sigma \quad (2.15)$$

For example:  $\Phi(2)$ :<sup>2</sup>

<sup>2</sup> How would you compute  $\Phi(-2)$ ?

```
pnorm(2)
## [1] 0.9772499
```

If  $Z$  is a standard normal random variable (SNRV) then

$$p\{Z \leq -x\} = P\{Z > x\}, \quad -\infty < x < \infty \quad (2.16)$$

Since  $Z = ((X - \mu)/\sigma)$  is an SNRV whenever  $X$  is normally distributed with parameters  $\mu$  and  $\sigma^2$ , then the CDF of  $X$  can be expressed as:

$$F_X(a) = P\{X \leq a\} = P\left(\frac{X - \mu}{\sigma} \leq \frac{a - \mu}{\sigma}\right) = \Phi\left(\frac{a - \mu}{\sigma}\right) \quad (2.17)$$

The standardized version of a normal random variable  $X$  is used to compute specific probabilities relating to  $X$  (it's also easier to compute probabilities from different CDFs so that the two computations are comparable).

### 2.3 Expectations and Variances

#### 2.3.1 Expectations and variances of discrete RVs

The expectation can be seen as the long-run average value. Let  $X$  be a discrete random variable. Then, random samples will be designated as the values  $x_1, x_2, \dots, x_n$ . For example:

```
x<-0:10
## expectation in our binomial example:
sum(x*dbinom(x,size=10,prob=0.5))

## [1] 5
```

$$E[X] = \sum_{i=1}^n x_i p(x_i) \quad (2.18)$$

In the binomial case,  $E[X] = np$ , where  $n$  is the sample size, and  $p$  the probability of success.<sup>3</sup> We will refer to  $E[X]$  as  $\mu$ .

The variance of the discrete random variable  $X$  is

$$\text{Var}(X) = E[(X - \mu)^2] \quad (2.19)$$

In the binomial case,  $\text{Var}(X) = np(1 - p)$ .<sup>4</sup>

#### 2.3.2 Expectations and variances of continuous RVs

Let  $X$  be a continuous random variable with PDF  $f(x)$ . Then, the expectation is:

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx = \mu \quad (2.20)$$

The expectation of a function of  $X$ ,  $g(X)$ :

<sup>3</sup> Proof: see [https://proofwiki.org/wiki/Expectation\\_of\\_Binomial\\_Distribution](https://proofwiki.org/wiki/Expectation_of_Binomial_Distribution)

<sup>4</sup> Proof: see [https://proofwiki.org/wiki/Variance\\_of\\_Binomial\\_Distribution](https://proofwiki.org/wiki/Variance_of_Binomial_Distribution)

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx = \mu \quad (2.21)$$

The variance is defined as:

$$\text{Var}[X] = E[(X - E[X])^2] \quad (2.22)$$

An easier way to find the variance is through this equality:

$$\text{Var}[X] = E[X^2] - (E[X])^2 \quad (2.23)$$

That is, to compute variance we just need to find  $E[X]$  and then  $E[X^2]$ . The proof of the above equality goes as follows:

Let  $E[X] = \mu$ . By the definition of variance:

$$\text{Var}[X] = E[(X - E[X])^2] = E[(X - \mu)^2] \quad (2.24)$$

Expanding out the RHS:

$$\text{Var}[X] = E[(X - \mu)^2] = E[X^2 - 2\mu X + \mu^2] \quad (2.25)$$

By the linearity of expectation, we can rewrite this as:

$$\begin{aligned} \text{Var}[X] &= E[X^2] - 2\mu E[X] + \mu^2 \\ &= E[X^2] - 2\mu^2 + \mu^2 \\ &= E[X^2] - \mu^2 \end{aligned} \quad (2.26)$$

■

### 2.3.3 Example: The expectation and variance of the standard normal RV

*Expectation* Let  $X$  be a standard normal random variable.

$$E[X] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} xe^{-x^2/2} dx$$

Let  $u = -x^2/2$ .

Then,  $du/dx = -2x/2 = -x$ . I.e.,  $du = -x dx$  or  $-du = x dx$ .

We can rewrite the integral as:

$$E[X] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^u x dx$$

Replacing  $x dx$  with  $-du$  we get:

$$-\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^u du$$

which yields:

$$-\frac{1}{\sqrt{2\pi}}[e^u]_{-\infty}^{\infty}$$

Replacing  $u$  with  $-x^2/2$  we get:

$$-\frac{1}{\sqrt{2\pi}}[e^{-x^2/2}]_{-\infty}^{\infty} = 0$$

*Variance* We know that

$$\text{Var}(X) = E[X^2] - (E[X])^2$$

Since  $(E[X])^2 = 0$  (see immediately above), we just have to compute  $E[g(X)] = E[X^2]$ . Here, we use the earlier definition, see Equation 2.21, of the expectation of a function of a random variable.

$$\text{Var}(X) = E[X^2] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2} dx$$

This is  $g(X)$ .

Write  $x^2$  as  $x \times x$  and use integration by parts:<sup>5</sup>

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x x e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} x \underset{u}{\uparrow} - e^{-x^2/2} \underset{v}{\uparrow} - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} -e^{-x^2/2} \underset{v}{\uparrow} 1 \underset{du/dx}{\uparrow} dx = 1$$

[Explained in p. 274 of Grinstead and Snell<sup>6</sup>: “The first summand above can be shown to equal 0, since as  $x \rightarrow \pm\infty$ ,  $e^{-x^2/2}$  gets small more quickly than  $x$  gets large. The second summand is just the standard normal density integrated over its domain, so the value of this summand is 1. Therefore, the variance of the standard normal density equals 1.”]

<sup>5</sup> Recall how integration by parts works:

$$\frac{d(uv)}{dx} = u \frac{dv}{dx} + \int v \frac{du}{dx} \quad (2.27)$$

$$uv = \int u \frac{dv}{dx} dx + \int v \frac{du}{dx} dx \quad (2.28)$$

$$\int u \frac{dv}{dx} dx = uv - \int v \frac{du}{dx} dx \quad (2.29)$$

<sup>6</sup> C.M. Grinstead and J.L. Snell. *Introduction to probability*. American Mathematical Society, 1997

# 3

## *Some useful probability distributions*

### 3.0.1 Exponential random variables

For some  $\lambda > 0$ ,

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

A continuous random variable with the above PDF is an exponential random variable (or is said to be exponentially distributed).

The CDF:

$$\begin{aligned} F(a) &= P(X \leq a) \\ &= \int_0^a \lambda e^{-\lambda x} dx \\ &= \left[ -e^{-\lambda x} \right]_0^a \\ &= 1 - e^{-\lambda a} \quad a \geq 0 \end{aligned}$$

[Note: the integration requires the u-substitution:  $u = -\lambda x$ , and then  $du/dx = -\lambda$ , and then use  $-du = \lambda dx$  to solve.]

*Expectation and variance of an exponential random variable* For some  $\lambda > 0$  (called the rate), if we are given the PDF of a random variable X:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

Find  $E[X]$ .

[This proof seems very strange and arbitrary—one starts really generally and then scales down, so to speak. The standard method can equally well be used, but this is more general, it allows for easy calculation of the second moment, for example. Also, it's an example of how reduction formulae are used in integration.]

$$E[X^n] = \int_0^\infty x^n \lambda e^{-\lambda x} dx$$

Use integration by parts (see equation 2.29 on page 18):

Let  $u = x^n$ , which gives  $du/dx = nx^{n-1}$ . Let  $dv/dx = \lambda e^{-\lambda x}$ , which gives  $v = -e^{-\lambda x}$ . Therefore:

$$\begin{aligned} E[X^n] &= \int_0^\infty x^n \lambda e^{-\lambda x} dx \\ &= \left[ -x^n e^{-\lambda x} \right]_0^\infty + \int_0^\infty e^{\lambda x} n x^{n-1} dx \\ &= 0 + \frac{n}{\lambda} \int_0^\infty \lambda e^{-\lambda x} n x^{n-1} dx \end{aligned}$$

Thus,

$$E[X^n] = \frac{n}{\lambda} E[X^{n-1}]$$

If we let  $n = 1$ , we get  $E[X]$ :

$$E[X] = \frac{1}{\lambda}$$

Note that when  $n = 2$ , we have

$$E[X^2] = \frac{2}{\lambda} E[X] = \frac{2}{\lambda^2}$$

Variance is, as usual,

$$\text{var}(X) = E[X^2] - (E[X])^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$$

### 3.0.2 Weibull distribution

$$f(x | \alpha, \beta) = \alpha \beta (\beta x)^{\alpha-1} \exp(-(\beta x)^\alpha) \quad (3.1)$$

When  $\alpha = 1$ , we have the exponential distribution.

### 3.0.3 Gamma distribution

[The text is an amalgam of Kerns<sup>1</sup> and Ross<sup>2</sup>. I don't put it in double-quotes as a citation because it would look ugly.]

This is a generalization of the exponential distribution. We say that  $X$  has a gamma distribution and write  $X \sim \text{gamma}(\text{shape} = \alpha, \text{rate} = \lambda)$ , where  $\alpha > 0$  (called shape) and  $\lambda > 0$  (called rate). It has PDF

$$f(x) = \begin{cases} \frac{\lambda e^{-\lambda x} (\lambda x)^{\alpha-1}}{\Gamma(\alpha)} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

$\Gamma(\alpha)$  is called the gamma function:

<sup>1</sup> G. Jay Kerns. *Introduction to Probability and Statistics Using R*. 2010

<sup>2</sup> Sheldon Ross. *A first course in probability*. Pearson Education, 2002

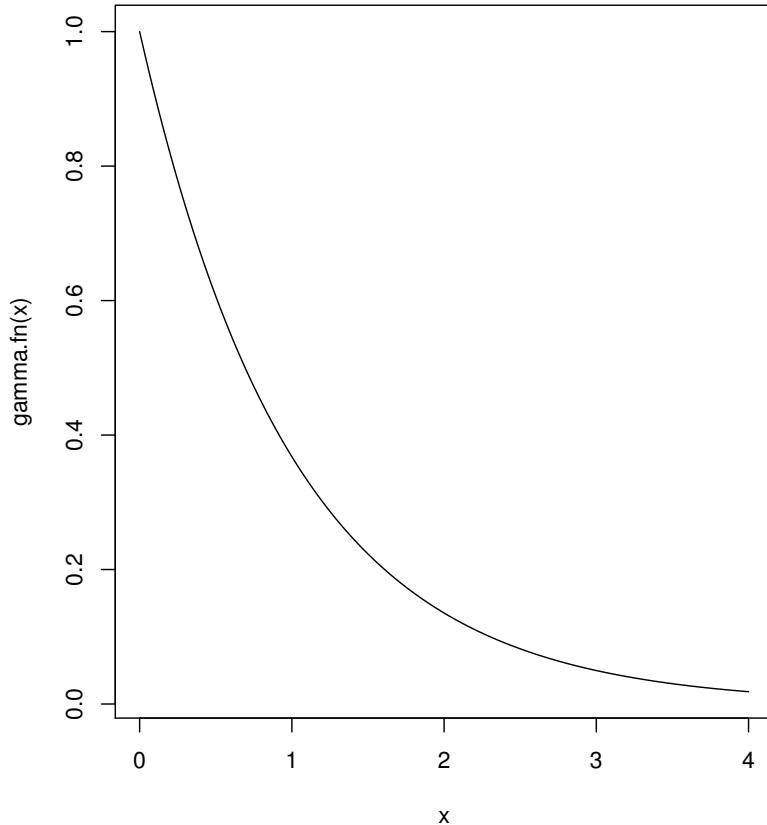
$$\Gamma(\alpha) = \int_0^{\infty} e^{-y} y^{\alpha-1} dy = (\alpha - 1) \underset{\text{integration by parts}}{\uparrow} \Gamma(\alpha - 1)$$

Note that for integral values of  $n$ ,  $\Gamma(n) = (n - 1)!$  (follows from above equation).

The associated R functions are `gamma(x, shape, rate = 1)`, `pgamma`, `qgamma`, and `rgamma`, which give the PDF, CDF, quantile function, and simulate random variates, respectively. If  $\alpha = 1$  then  $X \sim \exp(\text{rate} = \lambda)$ . The mean is  $\mu = \alpha/\lambda$  and the variance is  $\sigma^2 = \alpha/\lambda^2$ .

To motivate the gamma distribution recall that if  $X$  measures the length of time until the first event occurs in a Poisson process with rate  $\lambda$  then  $X \sim \exp(\text{rate} = \lambda)$ . If we let  $Y$  measure the length of time until the  $\alpha^{\text{th}}$  event occurs then  $Y \sim \text{gamma}(\text{shape} = \alpha, \text{rate} = \lambda)$ . When  $\alpha$  is an integer this distribution is also known as the **Erlang** distribution.

Figure 3.1: The gamma distribution.



The Chi-squared distribution is the gamma distribution with  $\lambda = 1/2$  and  $\alpha = n/2$ , where  $n$  is an integer:

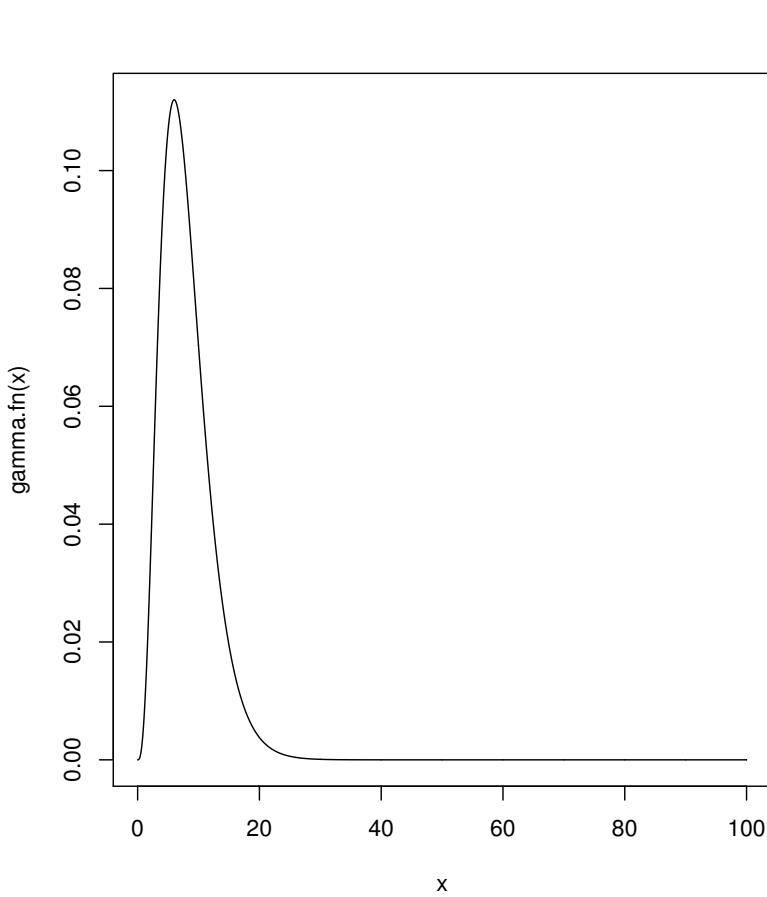


Figure 3.2: The chi-squared distribution.

*Mean and variance of gamma distribution* Let  $X$  be a gamma random variable with parameters  $\alpha$  and  $\lambda$ .

$$\begin{aligned} E[X] &= \frac{1}{\Gamma(\alpha)} \int_0^\infty x \lambda e^{-\lambda x} (\lambda x)^{\alpha-1} dx \\ &= \frac{1}{\lambda \Gamma(\alpha)} \int_0^\infty e^{-\lambda x} (\lambda x)^\alpha dx \\ &= \frac{\Gamma(\alpha+1)}{\lambda \Gamma(\alpha)} \\ &= \frac{\alpha}{\lambda} \end{aligned}$$

(See derivation of  $\Gamma(\alpha)$ , p. 215 of Ross<sup>3</sup>.)

It is easy to show (exercise) that

<sup>3</sup> Sheldon Ross. *A first course in probability*. Pearson Education, 2002

$$Var(X) = \frac{\alpha}{\lambda^2}$$

### 3.0.4 Uniform random variable

A random variable ( $X$ ) with the continuous uniform distribution on the interval  $(\alpha, \beta)$  has PDF

$$f_X(x) = \begin{cases} \frac{1}{\beta-\alpha}, & \alpha < x < \beta, \\ 0, & \text{otherwise} \end{cases} \quad (3.2)$$

The associated R function is `dunif(min = a, max = b)`. We write  $X \sim \text{unif}(\min = a, \max = b)$ . Due to the particularly simple form of this PDF we can also write down explicitly a formula for the CDF  $F_X$ :

$$F_X(a) = \begin{cases} 0, & a < 0, \\ \frac{a-\alpha}{\beta-\alpha}, & \alpha \leq a < \beta, \\ 1, & a \geq \beta. \end{cases} \quad (3.3)$$

$$E[X] = \frac{\beta + \alpha}{2} \quad (3.4)$$

$$Var(X) = \frac{(\beta - \alpha)^2}{12} \quad (3.5)$$

### 3.0.5 Beta distribution

This is a generalization of the continuous uniform distribution.

$$f(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1} (1-x)^{b-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

where

$$B(a, b) = \int_0^1 x^{a-1} (1-x)^{b-1} dx$$

There is a connection between the beta and the gamma:

$$B(a, b) = \int_0^1 x^{a-1} (1-x)^{b-1} dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

which allows us to rewrite the beta PDF as

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1}, \quad 0 < x < 1. \quad (3.6)$$

The mean and variance are

$$E[X] = \frac{a}{a+b} \text{ and } Var(X) = \frac{ab}{(a+b)^2 (a+b+1)}. \quad (3.7)$$

### 3.1 Jointly distributed random variables

#### 3.1.1 Discrete case

[This section is an extract from Kerns. I omit quotes as that would make the text harder to read.]

Consider two discrete random variables  $X$  and  $Y$  with PMFs  $f_X$  and  $f_Y$  that are supported on the sample spaces  $S_X$  and  $S_Y$ , respectively. Let  $S_{X,Y}$  denote the set of all possible observed pairs  $(x,y)$ , called the **joint support set** of  $X$  and  $Y$ . Then the **joint probability mass function** of  $X$  and  $Y$  is the function  $f_{X,Y}$  defined by

$$f_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y), \quad \text{for } (x,y) \in S_{X,Y}. \quad (3.8)$$

Every joint PMF satisfies

$$f_{X,Y}(x,y) > 0 \text{ for all } (x,y) \in S_{X,Y}, \quad (3.9)$$

and

$$\sum_{(x,y) \in S_{X,Y}} f_{X,Y}(x,y) = 1. \quad (3.10)$$

It is customary to extend the function  $f_{X,Y}$  to be defined on all of  $\mathbb{R}^2$  by setting  $f_{X,Y}(x,y) = 0$  for  $(x,y) \notin S_{X,Y}$ .

In the context of this chapter, the PMFs  $f_X$  and  $f_Y$  are called the **marginal PMFs** of  $X$  and  $Y$ , respectively. If we are given only the joint PMF then we may recover each of the marginal PMFs by using the Theorem of Total Probability:

$$f_X(x) = \mathbb{P}(X = x), \quad (3.11)$$

$$= \sum_{y \in S_Y} \mathbb{P}(X = x, Y = y), \quad (3.12)$$

$$= \sum_{y \in S_Y} f_{X,Y}(x,y). \quad (3.13)$$

By interchanging the roles of  $X$  and  $Y$  it is clear that

$$f_Y(y) = \sum_{x \in S_X} f_{X,Y}(x,y). \quad (3.14)$$

Given the joint PMF we may recover the marginal PMFs, but the converse is not true. Even if we have **both** marginal distributions they are not sufficient to determine the joint PMF; more information is needed.

Associated with the joint PMF is the **joint cumulative distribution function**  $F_{X,Y}$  defined by

$$F_{X,Y}(x,y) = \mathbb{P}(X \leq x, Y \leq y), \quad \text{for } (x,y) \in \mathbb{R}^2.$$

The bivariate joint CDF is not quite as tractable as the univariate CDFs, but in principle we could calculate it by adding up quantities of the form in Equation 3.8. The joint CDF is typically not used in practice due to its inconvenient form; one can usually get by with the joint PMF alone.

*Example: Discrete bivariate case* Roll a fair die twice. Let  $X$  be the face shown on the first roll, and let  $Y$  be the face shown on the second roll. For this example, it suffices to define

$$f_{X,Y}(x,y) = \frac{1}{36}, \quad x = 1, \dots, 6, y = 1, \dots, 6.$$

The marginal PMFs are given by  $f_X(x) = 1/6$ ,  $x = 1, 2, \dots, 6$ , and  $f_Y(y) = 1/6$ ,  $y = 1, 2, \dots, 6$ , since

$$f_X(x) = \sum_{y=1}^6 \frac{1}{36} = \frac{1}{6}, \quad x = 1, \dots, 6,$$

and the same computation with the letters switched works for  $Y$ .

Here, and in many other ones, the joint support can be written as a product set of the support of  $X$  “times” the support of  $Y$ , that is, it may be represented as a cartesian product set, or rectangle,  $S_{X,Y} = S_X \times S_Y$ , where  $S_X \times S_Y = \{(x,y) : x \in S_X, y \in S_Y\}$ . This form is a necessary condition for  $X$  and  $Y$  to be **independent** (or alternatively **exchangeable** when  $S_X = S_Y$ ). But please note that in general it is not required for  $S_{X,Y}$  to be of rectangle form.

### 3.1.2 Continuous case

For random variables  $X$  and  $y$ , the **joint cumulative pdf** is

$$F(a,b) = P(X \leq a, Y \leq b) \quad -\infty < a, b < \infty \quad (3.15)$$

The **marginal distributions** of  $F_X$  and  $F_Y$  are the CDFs of each of the associated RVs:

1. The CDF of  $X$ :

$$F_X(a) = P(X \leq a) = F_X(a, \infty) \quad (3.16)$$

2. The CDF of  $Y$ :

$$F_Y(b) = P(Y \leq b) = F_Y(\infty, b) \quad (3.17)$$

**Definition 1** *Jointly continuous:* Two RVs  $X$  and  $Y$  are jointly continuous if there exists a function  $f(x, y)$  defined for all real  $x$  and  $y$ , such that for every set  $C$ :

$$P((X, Y) \in C) = \iint_{(x,y) \in C} f(x, y) dx dy \quad (3.18)$$

$f(x, y)$  is the **joint PDF** of  $X$  and  $Y$ .

Every joint PDF satisfies

$$f(x, y) \geq 0 \text{ for all } (x, y) \in S_{X,Y}, \quad (3.19)$$

and

$$\iint_{S_{X,Y}} f(x, y) dx dy = 1. \quad (3.20)$$

For any sets of real numbers  $A$  and  $B$ , and if  $C = \{(x, y) : x \in A, y \in B\}$ , it follows from equation 3.18 that

$$P((X \in A, Y \in B) \in C) = \int_B \int_A f(x, y) dx dy \quad (3.21)$$

Note that

$$F(a, b) = P(X \in (-\infty, a], Y \in (-\infty, b])) = \int_{-\infty}^b \int_{-\infty}^a f(x, y) dx dy \quad (3.22)$$

Differentiating, we get the joint pdf:

$$f(a, b) = \frac{\partial^2}{\partial a \partial b} F(a, b) \quad (3.23)$$

One way to understand the joint PDF:

$$P(a < X < a + da, b < Y < b + db) = \int_b^{a+da} \int_a^{a+da} f(x, y) dx dy \approx f(a, b) dadb \quad (3.24)$$

Hence,  $f(x, y)$  is a measure of how probable it is that the random vector  $(X, Y)$  will be near  $(a, b)$ .

### 3.1.3 Marginal probability distribution functions

If  $X$  and  $Y$  are jointly continuous, they are individually continuous, and their PDFs are:

$$\begin{aligned} P(X \in A) &= P(X \in A, Y \in (-\infty, \infty)) \\ &= \int_A \int_{-\infty}^{\infty} f(x, y) dy dx \\ &= \int_A f_X(x) dx \end{aligned} \quad (3.25)$$

where

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy \quad (3.26)$$

Similarly:

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx \quad (3.27)$$

### 3.1.4 Independent random variables

Random variables  $X$  and  $Y$  are independent iff, for any two sets of real numbers  $A$  and  $B$ :

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B) \quad (3.28)$$

In the jointly continuous case:

$$f(x, y) = f_X(x)f_Y(y) \quad \text{for all } x, y \quad (3.29)$$

A necessary and sufficient condition for the random variables  $X$  and  $Y$  to be independent is for their joint probability density function (or joint probability mass function in the discrete case)  $f(x, y)$  to factor into two terms, one depending only on  $x$  and the other depending only on  $y$ .

*Easy-to-understand example from Kerns* Let the joint PDF of  $(X, Y)$  be given by

$$f_{X,Y}(x, y) = \frac{6}{5} (x + y^2), \quad 0 < x < 1, 0 < y < 1.$$

The marginal PDF of  $X$  is

$$\begin{aligned} f_X(x) &= \int_0^1 \frac{6}{5} (x + y^2) dy, \\ &= \frac{6}{5} \left( xy + \frac{y^3}{3} \right) \Big|_{y=0}^1, \\ &= \frac{6}{5} \left( x + \frac{1}{3} \right), \end{aligned}$$

for  $0 < x < 1$ , and the marginal PDF of  $Y$  is

$$\begin{aligned} f_Y(y) &= \int_0^1 \frac{6}{5} (x + y^2) dx, \\ &= \frac{6}{5} \left( \frac{x^2}{2} + xy^2 \right) \Big|_{x=0}^1, \\ &= \frac{6}{5} \left( \frac{1}{2} + y^2 \right), \end{aligned}$$

for  $0 < y < 1$ .

In this example the joint support set was a rectangle  $[0, 1] \times [0, 1]$ , but it turns out that  $X$  and  $Y$  are not independent. This is because  $\frac{6}{5}(x + y^2)$  cannot be stated as a product of two terms ( $f_X(x)f_Y(y)$ ).

### 3.1.5 Sums of independent random variables

[This is taken nearly verbatim from Ross.]

Suppose that  $X$  and  $Y$  are independent, continuous random variables having probability density functions  $f_X$  and  $f_Y$ . The cumulative distribution function of  $X + Y$  is obtained as follows:

$$\begin{aligned}
 F_{X+Y}(a) &= P(X + Y \leq a) \\
 &= \iint_{\substack{x+y \leq a}} f_{XY}(x, y) dx dy \\
 &= \iint_{\substack{x+y \leq a}} f_X(x)f_Y(y) dx dy \\
 &= \int_{-\infty}^{\infty} \int_{-\infty}^{a-y} f_X(x)f_Y(y) dx dy \\
 &= \int_{-\infty}^{\infty} \int_{-\infty}^{a-y} f_X(x) dx f_Y(y) dy \\
 &= \int_{-\infty}^{\infty} F_X(a - y) f_Y(y) dy
 \end{aligned} \tag{3.30}$$

The CDF  $F_{X+Y}$  is the **convolution** of the distributions  $F_X$  and  $F_Y$ .

If we differentiate the above equation, we get the pdf  $f_{X+Y}$ :

$$\begin{aligned}
 f_{X+Y} &= \frac{d}{dx} \int_{-\infty}^{\infty} F_X(a - y) f_Y(y) dy \\
 &= \int_{-\infty}^{\infty} \frac{d}{dx} F_X(a - y) f_Y(y) dy \\
 &= \int_{-\infty}^{\infty} f_X(a - y) f_Y(y) dy
 \end{aligned} \tag{3.31}$$

### 3.1.6 Conditional distributions

*Discrete case* Recall that the conditional probability of  $B$  given  $A$ , denoted  $\mathbb{P}(B | A)$ , is defined by

$$\mathbb{P}(B | A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}, \quad \text{if } \mathbb{P}(A) > 0. \tag{3.32}$$

If  $X$  and  $Y$  are discrete random variables, then we can define the conditional PMF of  $X$  given that  $Y = y$  as follows:

$$\begin{aligned}
p_{X|Y}(x | y) &= P(X = x | Y = y) \\
&= \frac{P(X = x, Y = y)}{P(Y = y)} \\
&= \frac{p(x, y)}{p_Y(y)}
\end{aligned} \tag{3.33}$$

for all values of  $y$  where  $p_Y(y) = P(Y = y) > 0$ .

The **conditional cumulative distribution function** of  $X$  given  $Y = y$  is defined, for all  $y$  such that  $p_Y(y) > 0$ , as follows:

$$\begin{aligned}
F_{X|Y} &= P(X \leq x | Y = y) \\
&= \sum_{a \leq x} p_{X|Y}(a | y)
\end{aligned} \tag{3.34}$$

If  $X$  and  $Y$  are independent then

$$p_{X|Y}(x | y) = P(X = x) = p_X(x) \tag{3.35}$$

See the examples starting p. 264 of Ross.

*Continuous case* [Taken almost verbatim from Ross.]

If  $X$  and  $Y$  have a joint probability density function  $f(x, y)$ , then the conditional probability density function of  $X$  given that  $Y = y$  is defined, for all values of  $y$  such that  $f_Y(y) > 0$ , by

$$f_{X|Y}(x | y) = \frac{f(x, y)}{f_Y(y)} \tag{3.36}$$

We can understand this definition by considering what  $f_{X|Y}(x | y) dx$  amounts to:

$$\begin{aligned}
f_{X|Y}(x | y) dx &= \frac{f(x, y)}{f_Y(y)} \frac{dxdy}{dy} \\
&= \frac{f(x, y) dxdy}{f_Y(y) dy} \\
&= \frac{P(x < X < d + dx, y < Y < y + dy)}{y < Y < y + dy}
\end{aligned} \tag{3.37}$$

### 3.1.7 Joint and marginal expectation

[Taken nearly verbatim from Kerns.]

Given a function  $g$  with arguments  $(x, y)$  we would like to know the long-run average behavior of  $g(X, Y)$  and how to mathematically calculate it. Expectation in this context is computed by integrating (summing) with respect to the joint probability density (mass) function.

*Discrete case*

$$\mathbb{E} g(X, Y) = \sum_{(x,y) \in S_{X,Y}} g(x, y) f_{X,Y}(x, y). \quad (3.38)$$

*Continuous case*

$$\mathbb{E} g(X, Y) = \iint_{S_{X,Y}} g(x, y) f_{X,Y}(x, y) dx dy, \quad (3.39)$$

### 3.1.8 Covariance and correlation

There are two very special cases of joint expectation: the **covariance** and the **correlation**. These are measures which help us quantify the dependence between  $X$  and  $Y$ .

**Definition 2** *The covariance of  $X$  and  $Y$  is*

$$\text{Cov}(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y). \quad (3.40)$$

Shortcut formula for covariance:

$$\text{Cov}(X, Y) = \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y). \quad (3.41)$$

The **Pearson product moment correlation** between  $X$  and  $Y$  is the covariance between  $X$  and  $Y$  rescaled to fall in the interval  $[-1, 1]$ . It is formally defined by

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}. \quad (3.42)$$

The correlation is usually denoted by  $\rho_{X,Y}$  or simply  $\rho$  if the random variables are clear from context. There are some important facts about the correlation coefficient:

1. The range of correlation is  $-1 \leq \rho_{X,Y} \leq 1$ .
2. Equality holds above ( $\rho_{X,Y} = \pm 1$ ) if and only if  $Y$  is a linear function of  $X$  with probability one.

*Continuous example (from Kerns)* Let us find the covariance of the variables  $(X, Y)$  from an example numbered 7.2 in Kerns. The expected value of  $X$  is

$$\mathbb{E}X = \int_0^1 x \cdot \frac{6}{5} \left( x + \frac{1}{3} \right) dx = \frac{2}{5}x^3 + \frac{1}{5}x^2 \Big|_{x=0}^1 = \frac{3}{5},$$

and the expected value of  $Y$  is

$$\mathbb{E}Y = \int_0^1 y \cdot \frac{6}{5} \left( \frac{1}{2} + y^2 \right) dx = \frac{3}{10}y^2 + \frac{3}{20}y^4 \Big|_{y=0}^1 = \frac{9}{20}.$$

Finally, the expected value of  $XY$  is

$$\begin{aligned}\mathbb{E}XY &= \int_0^1 \int_0^1 xy \frac{6}{5} (x + y^2) dx dy, \\ &= \int_0^1 \left( \frac{2}{5}x^3y + \frac{3}{10}xy^4 \right) \Big|_{x=0}^1 dy, \\ &= \int_0^1 \left( \frac{2}{5}y + \frac{3}{10}y^4 \right) dy, \\ &= \frac{1}{5} + \frac{3}{50},\end{aligned}$$

which is  $13/50$ . Therefore the covariance of  $(X, Y)$  is

$$\text{Cov}(X, Y) = \frac{13}{50} - \left(\frac{3}{5}\right) \left(\frac{9}{20}\right) = -\frac{1}{100}.$$

### 3.1.9 Conditional expectation

Recall that

$$f_{X|Y}(x | y) = P(X = x | Y = y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} \quad (3.43)$$

for all  $y$  such that  $P(Y = y) > 0$ .

It follows that

$$\begin{aligned}E[X | Y = y] &= \sum_x x P(X = x | Y = y) \\ &= \sum_x x p_{X|Y}(x | y)\end{aligned} \quad (3.44)$$

$E[X | Y]$  is that **function** of the random variable  $Y$  whose value at  $Y = y$  is  $E[X | Y = y]$ .  $E[X | Y]$  is a random variable.

*Relationship to ‘regular’ expectation* Conditional expectation given that  $Y = y$  can be thought of as being an ordinary expectation on a reduced sample space consisting only of outcomes for which  $Y = y$ . All properties of expectations hold. Two examples (to-do: spell out the other equations):

*Example 1*

$$E[g(X) | Y = y] = \begin{cases} \sum_x g(x)p_{X|Y}(x,y) & \text{in the discrete case} \\ \int_{-\infty}^{\infty} g(x)f_{X|Y}(x | y) dx & \text{in the continuous case} \end{cases}$$

*Example 2*

$$E \left[ \sum_{i=1}^n X_i | Y = y \right] = \sum_{i=1}^n E[X_i | Y = y] \quad (3.45)$$

**Proposition 1** *Expectation of the conditional expectation*

$$E[X] = E[E[X | Y]] \quad (3.46)$$

If  $Y$  is a discrete random variable, then the above proposition states that

$$E[X] = \sum_y E[X | Y = y] P(Y = y) \quad (3.47)$$

# 4

## *Maximum Likelihood Estimation*

### *4.1 Introduction to Maximum Likelihood Estimation*

Suppose we toss a fair coin 10 times, and count the number of heads each time; we repeat this experiment 3 times in all. The observed sample values (the number of heads in each set of 10 tosses) are  $x_1, x_2, x_3$ .

```
(x<-rbinom(3, size=10, prob=0.5))  
## [1] 7 2 2
```

The joint probability of getting all these values (assuming independence) depends on the parameter we set for the probability  $\theta$ . Due to independence between the experiments, we can just multiply out the probabilities for each outcome:

$$f(X_1 = x_1, X_2 = x_2, X_3 = x_3; \theta) = \\ \binom{n}{x_1} \theta^{x_1} (1 - \theta)^{n-x_1} \binom{n}{x_2} \theta^{x_2} (1 - \theta)^{n-x_2} \binom{n}{x_3} \theta^{x_3} (1 - \theta)^{n-x_3} \quad (4.1)$$

So, the above joint probability is a function of  $\theta$ . You can try plugging in the values for  $n, x_1, \dots, x_3$ , and then try out different values of  $\theta$  to see how this joint probability changes. When this joint probability is expressed as a function of  $\theta$ , we call it the likelihood function.

The value of  $\theta$  for which this function has the maximum value is the maximum likelihood estimate.

We can plug in different values of  $\theta$  to look at the likelihood function:

```
## probability parameter fixed at 0.5  
theta<-0.5  
prod(dbinom(x, size=10, prob=theta))  
## [1] 0.0002263114
```

```
## probability parameter fixed at 0.1
theta<-0.1
prod(dbinom(x,size=10,prob=theta))

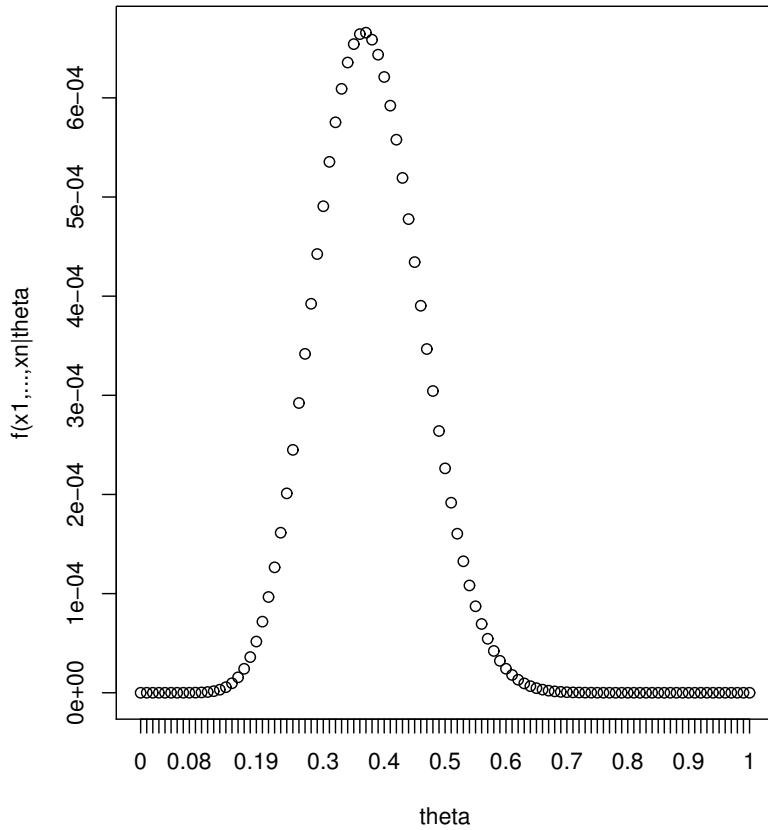
## [1] 3.28257e-07

## probability parameter fixed at 0.9
theta<-0.9
prod(dbinom(x,size=10,prob=theta))

## [1] 7.625597e-15

## let's compute the product for
## a range of probabilities:
theta<-seq(0,1,by=0.01)
store<-rep(NA,length(theta))
for(i in 1:length(theta)){
  store[i]<-prod(dbinom(x,size=10,prob=theta[i]))
}

plot(1:length(store),store,xaxt="n",xlab="theta",
      ylab="f(x1,...,xn|theta")
axis(1,at=1:length(theta),labels=theta)
```



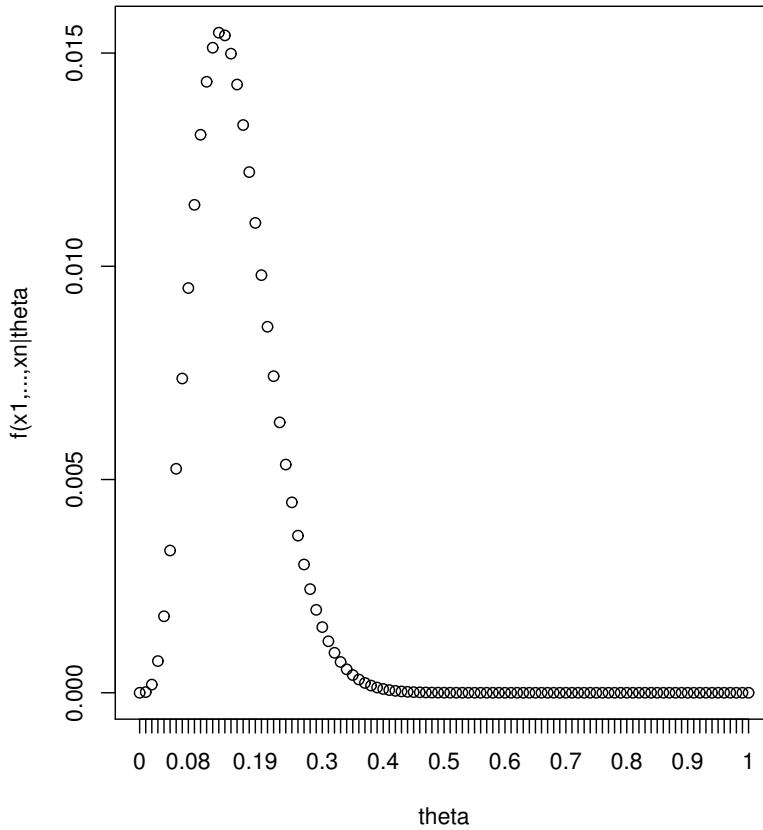
As another example, if the data had been generated by a binomial process with a different true  $\theta$  value, say 0.1, than the one chosen above (0.5):

```
(x<-rbinom(3, size=10, prob=0.1))
## [1] 2 2 0
```

then our likelihood function would look like this:

```
theta<-seq(0,1,by=0.01)
store<-rep(NA,length(theta))
for(i in 1:length(theta)){
  store[i]<-prod(dbinom(x,size=10,prob=theta[i]))
}

plot(1:length(store),store,xlab="theta",
      ylab="f(x1,...,xn|theta)",xaxt="n")
axis(1,at=1:length(theta),labels=theta)
```



Thus, the function  $f$  is the value of the joint probability **distribution** of the random variables  $X_1, \dots, X_n$  at  $X_1 = x_1, \dots, X_n = x_n$ . Since the sample values have been observed and are fixed,  $f(x_1, \dots, x_n; \theta)$  is a function of  $\theta$ . The function  $f$  is called a **likelihood function**. The highest point of this function corresponds to the value of  $\theta$  that is most likely to have generated the data.

#### Continuous case

Here,  $f$  is the joint probability **density**, the rest is the same as above.

**Definition 3** If  $x_1, x_2, \dots, x_n$  are the values of a random sample from a population with parameter  $\theta$ , the **likelihood function** of the sample is given by

$$L(\theta) = f(x_1, x_2, \dots, x_n; \theta) \quad (4.2)$$

for values of  $\theta$  within a given domain. Here,  $f(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n; \theta)$  is the joint probability distribution or density of the random variables  $X_1, \dots, X_n$  at  $X_1 = x_1, \dots, X_n = x_n$ .

So, the method of maximum likelihood consists of maximizing the likelihood function with respect to  $\theta$ . The value of  $\theta$  that maximizes the likelihood function is the **MLE** (maximum likelihood estimate) of  $\theta$ .

#### 4.1.1 Example 1: MLE of the Binomial distribution

From the examples above, we got a visual feel for what the MLE looks like for the binomial case. Now we will derive the MLE analytically. Let's assume again that we have a binomial random variable  $X$  as in the examples above: the observed value  $x$  is the number of heads observed when we toss a coin  $n$  times.

$$L(\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \quad (4.3)$$

Taking logs gives us the log likelihood. It is usually easier to work with the log likelihood because all products become sums and those are easier to deal with:

$$\ell(\theta) = \log \left[ \binom{n}{x} \theta^x (1-\theta)^{n-x} \right] = \log \binom{n}{x} + x \log \theta + (n-x) \log(1-\theta) \quad (4.4)$$

Differentiating and equating to zero to get the maximum:

$$\ell'(\theta) = \frac{x}{\theta} - \frac{n-x}{1-\theta} = 0 \quad (4.5)$$

How to get the second term: let  $u = 1 - \theta$ .

Then,  $du/d\theta = -1$ . Now,  $y = (n-x) \log(1-\theta)$  can be rewritten in terms of  $u$ :  $y = (n-x) \log(u)$ . So,  $dy/du = \frac{n-x}{u}$ .

Now, by the chain rule,  $dy/d\theta = dy/du \times du/d\theta = \frac{n-x}{u} \times (-1) = -\frac{n-x}{1-\theta}$ .

Rearranging terms, we get:

$$\frac{x}{\theta} - \frac{n-x}{1-\theta} = 0 \Leftrightarrow \frac{x}{\theta} = \frac{n-x}{1-\theta} \Leftrightarrow \hat{\theta} = \frac{x}{n} \quad (4.6)$$

Now you should establish that this is a maximum by taking the second derivative ( $\ell''(\theta)$ ) and checking that it is negative. (Exercise)<sup>1</sup>

<sup>1</sup> You should get  $f''(\theta) = \frac{1}{\theta^2} - \frac{(n-x)}{(1-\theta)^2}$ .

#### 4.1.2 Example 2: MLE of the Normal distribution

Let  $X_1, \dots, X_n$  constitute a random variable of size  $n$  from a normal population with mean  $\mu$  and variance  $\sigma^2$ , find joint maximum likelihood estimates of these two parameters.

We can again write the joint probability as a product of  $f(x_1) \times f(x_2) \times \dots \times f(x_n) = \prod_{i=1}^n f(x_i)$ . Another way to write  $f(x_i)$  for the normal distribution is  $N(x_i; \mu, \sigma^2)$ , so we can write:

$$L(\mu; \sigma^2) = \prod_{i=1}^n N(x_i; \mu, \sigma) \quad (4.7)$$

$$= \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right) \quad (4.8)$$

(4.9)

Taking logs and then partial derivatives with respect to  $\mu$  and  $\sigma$ , and equating these to 0, we get:

$$\hat{\mu} = \frac{1}{n} \sum x_i = \bar{x} \quad (4.10)$$

and

$$\hat{\sigma}^2 = \frac{1}{n} \sum (x_i - \bar{x})^2 \quad (4.11)$$

You will normally see the formula:

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 \quad (4.12)$$

The first formula for the variance leads to a biased estimator, and the second one to an unbiased one. Some definitions are in order.

**Definition 4 Estimator:** A point estimator<sup>2</sup> (such as  $\hat{\mu}$  and  $\hat{\sigma}^2$ ) is any function  $f(X_1, \dots, X_n)$  of a sample. I.e., any statistic is a point estimator.

**Definition 5 Bias:** The bias of a point estimator  $\hat{\theta}$  of a parameter  $\theta$  is the difference between the expected value of  $\hat{\theta}$  and  $\theta$ . That is,  $\text{Bias} = E[\hat{\theta}] - \theta$

*Expectation of  $X^2$*  Recall (equation 2.23, page 17) that for a random variable  $X$  with  $E[X] = \mu$  and  $\text{Var}(X) = \sigma^2$ :

$$\sigma^2 = \text{Var}(X) = E[X^2] - E[X]^2 = E[X^2] - \mu^2 \quad (4.13)$$

It follows that

$$E[X^2] = \mu^2 + \sigma^2 \quad (4.14)$$

By a similar logic, for the sampling distribution of  $X$ ,  $E[\bar{X}] = \mu$  and  $\text{Var}(\bar{X}) = \sigma^2/n$  (see equation 5.30, page 67). It follows that:

$$E[\bar{X}^2] = \mu^2 + \frac{\sigma^2}{n} \quad (4.15)$$

*Proof that  $\hat{\sigma}^2$  is biased* Now, we can proceed to show that the **estimator**  $\hat{\sigma}^2$ , defined below, of  $\sigma^2$  is biased. Recall the definition of this estimator: given random variables  $X_1, \dots, X_n$ , which come from  $N(\mu, \sigma^2)$ , we can estimate  $\sigma^2$  by computing:

<sup>2</sup> Note that an **estimate** is different from an estimator. An estimate is the realized value of the estimator when an actual sample is taken. So, in the following proof, note that the estimator is a function of random variables  $X_1, \dots, X_n$  (written as capital  $X$ 's), whereas the corresponding estimate would be a function of the actually sampled values  $x_1, \dots, x_n$  (lower case  $x$ ).

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} \quad (4.16)$$

For this estimator to be **unbiased**, it would have to be the case that  $E[\hat{\sigma}^2] = \sigma^2$ . So it's enough to show that this equality doesn't hold. We prove this below.

$$\begin{aligned} E[\hat{\sigma}^2] &= E\left[\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}\right] \\ &= \frac{1}{n} E\left[\sum_{i=1}^n (X_i - \bar{X})^2\right] \text{ (taking the constant term out)} \end{aligned} \quad (4.17)$$

Now, notice that expanding out the summation of the squared term gives us:

$$\begin{aligned} \sum_{i=1}^n (X_i - \bar{X})^2 &= \sum_{i=1}^n (X_i^2 - 2X_i\bar{X} + \bar{X}^2) \\ &= \sum_{i=1}^n X_i^2 - 2 \sum_{i=1}^n X_i\bar{X} + n\bar{X}^2 \end{aligned} \quad (4.18)$$

Rewriting  $\sum_{i=1}^n X_i$  as  $n\bar{X}$ ,

One useful thing to remember is that whenever you see  $\sum_{i=1}^n X_i$ , this term can be rewritten as:

$$\begin{aligned} \sum_{i=1}^n X_i^2 - 2 \sum_{i=1}^n X_i\bar{X} + n\bar{X}^2 &= \sum_{i=1}^n X_i^2 - 2n\bar{X}\bar{X} + n\bar{X}^2 \\ &= \sum_{i=1}^n X_i^2 - 2n\bar{X}^2 + n\bar{X}^2 \\ &= \sum_{i=1}^n X_i^2 - n\bar{X}^2 \end{aligned} \quad (4.20)$$

$$\sum_{i=1}^n X_i = n \frac{\sum_{i=1}^n X_i}{n} = n\bar{X} \quad (4.19)$$

Now, equation 4.17 can be written as:

$$\begin{aligned} E[\hat{\sigma}^2] &= \frac{1}{n} E\left[\sum_{i=1}^n (X_i - \bar{X})^2\right] \\ &= \frac{1}{n} E\left[\sum_{i=1}^n X_i^2 - n\bar{X}^2\right] \\ &= \frac{1}{n} \left[nE[X^2] - nE[\bar{X}^2]\right] \\ &= \frac{1}{n} \left[n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right] \\ &= \frac{1}{n} \left[(n-1)\sigma^2\right] \\ &= \frac{n-1}{n} \sigma^2 \end{aligned} \quad (4.21)$$

So,  $E[\hat{\sigma}^2] = \frac{n-1}{n}[\sigma^2]$ . That is, it is a biased estimator. It is clear how to make this estimator unbiased: divide by  $n - 1$  instead of  $n$ .

Notice that large  $n$ , dividing by  $n$  or  $n - 1$  isn't going to make much difference.

#### 4.1.3 Example 3: MLE of the Exponential distribution

Here, when I write  $x$ , I mean a vector of data points  $x_1, \dots, x_n$ . And when I write  $\sum$  I mean  $\sum_{i=1}^n$ .

$$f(x; \lambda) = \lambda \exp(-\lambda x) \quad (4.22)$$

Log likelihood:

$$\ell = n \log \lambda - \sum \lambda x_i \quad (4.23)$$

Differentiating and equating to zero:

$$\ell'(\lambda) = \frac{n}{\lambda} - \sum x_i = 0 \quad (4.24)$$

$$\frac{n}{\lambda} = \sum x_i \quad (4.25)$$

I.e.,

$$\frac{1}{\bar{\lambda}} = \frac{\sum x_i}{n} \quad (4.26)$$

#### 4.1.4 Practical implications

Whenever we obtain some data, we make an assumption about the generative process; we have to define the random variable  $X$  that we believe generated the data. The decision about which underlying PDF (and therefore what type of random variable) can be made based on examining the distribution of the data. For example, log reading time data (in log milliseconds) often has a normal distribution. For such data, we would write:

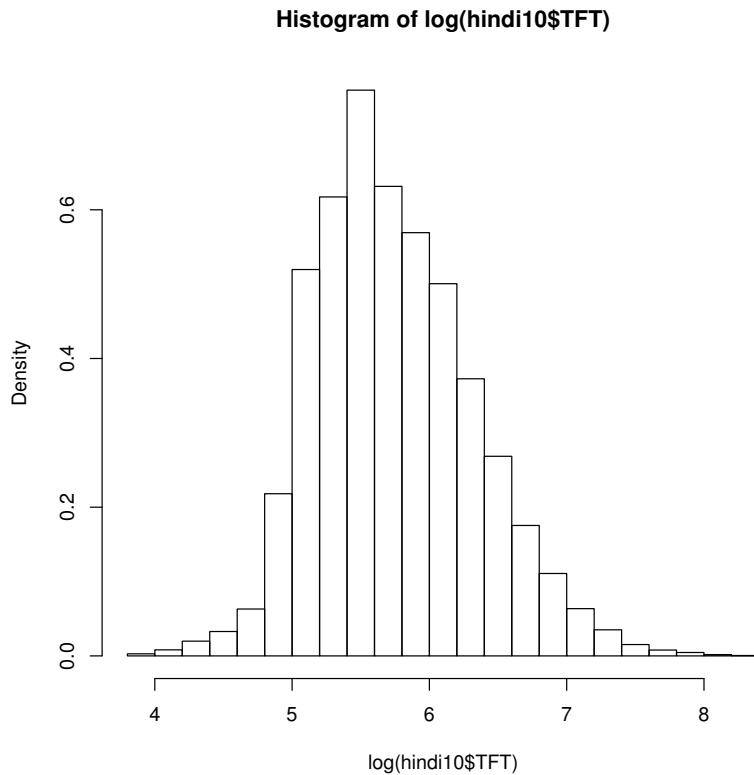
$$X \sim \text{Normal}(\mu, \sigma)$$

Once we've made this assumption, an obvious question arises: what should the values of  $\mu$  and  $\sigma$  be? The MLEs serve to provide a best guess.

As an example, consider the eyetracking data I released earlier. For now we will remove 0 ms reading times from the data-set; for the moment, this can be considered missing data.

A histogram reveals the following distribution of reading times on the log scale. It's a slight stretch, but we start with the guess that this is roughly a normal distribution:

```
hist(log(hindi10$TFT), freq=FALSE)
```

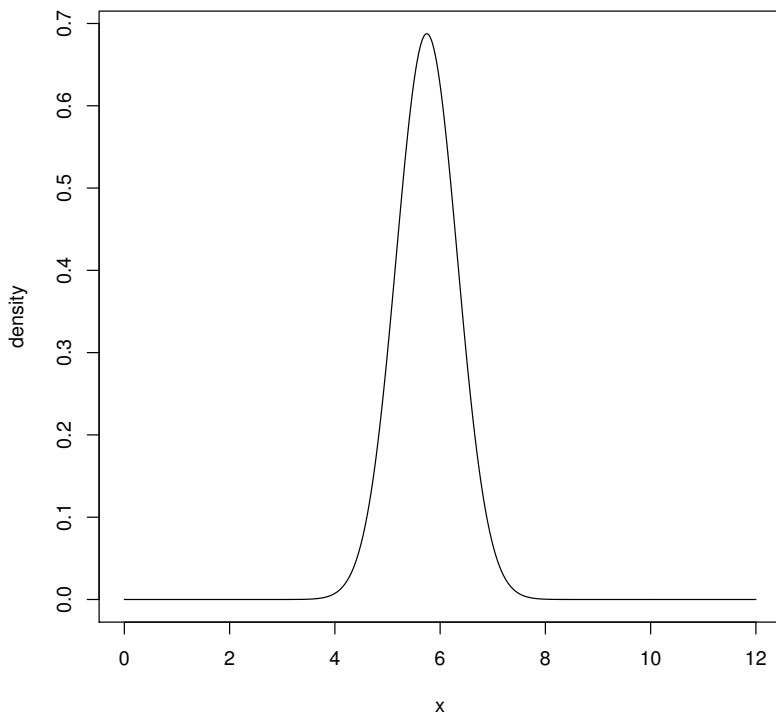


We take the mean and the (bias-corrected) variance estimates as the parameters of the underlying generative distribution. **This is where MLE becomes relevant.**

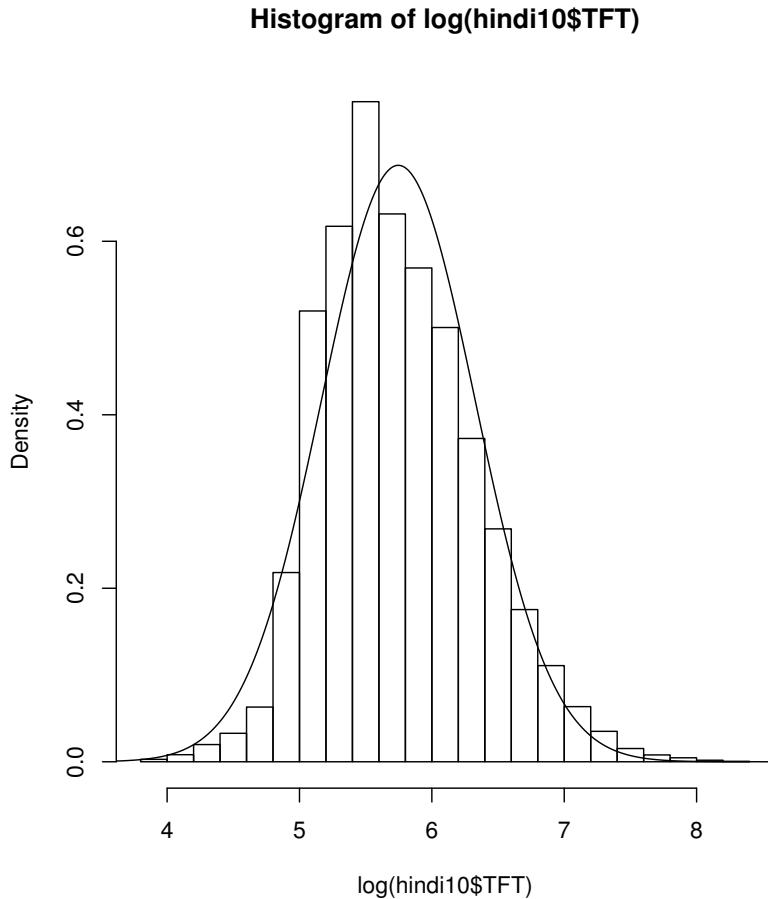
```
(xbar<-mean(log(hindi10$TFT)))
## [1] 5.746552
(xvar<-var(log(hindi10$TFT)))
## [1] 0.336665
```

The MLEs imply that the underlying generative distribution is the following one:

```
xvals<-seq(0,12,by=0.01)
plot(xvals,dnorm(xvals,
                  mean=xbar,
                  sd=sqrt(xvar)),
type="l",ylab="density",xlab="x")
```



```
## The empirical distribution and
## our theoretical distribution:
hist(log(hindi10$TFT), freq=FALSE)
xvals<-seq(0,4000,by=0.01)
lines(xvals,dnorm(xvals,
                   mean=xbar, sd=sqrt(xvar)))
```



**Exercise:** using the raw reading times in ms, compute the mean  $\hat{\mu}$  and variance  $\hat{\sigma}^2$ , and then plot the sample distribution (the histogram) and the theoretical normal distribution on top of it using these estimates, as done in the example above. Comment on the differences between the sample distribution and the theoretical distribution we assume here.

*Computing the MLE using an optimizer* Note that you can use the function `optim` to compute the maximum likelihood estimates, once you define some log-likelihood function whose parameters need to be estimated.

```
## define negative log lik:
nllh.normal<-function(theta,data){
  ## mean and sd
  m<-theta[1]
  s<-theta[2]
  x <- data
```

```

n<-length(x)
logl<- sum(dnorm(x,mean=m,sd=s,log=TRUE))
## return negative log likelihood:
-logl
}

## example output:
nllh.normal(theta=c(40,4),log(hindi10$TFT))

## [1] 766646

## find the MLEs using optim:
## need to specify some starting values:
opt.vals.default<-optim(theta<-c(5,.5),
                         nllh.normal,
                         data=log(hindi10$TFT),
                         hessian=TRUE)

## result of optimization:
(estimate.default<-opt.vals.default$par)

## [1] 5.7465550 0.5801744

## compare with MLE:
xbar

## [1] 5.746552

## bias corrected sd:
sqrt(xvar)

## [1] 0.5802284

```

## 4.2 Asymptotic properties of MLEs

MLEs have the important property that they are asymptotically normally distributed. This means that if we repeatedly takes samples of data, and record the distribution of the sample mean, the distribution will be normal if sample size is large enough, regardless of whether the underlying distribution we are sampling from is normal or not (this is assuming that the underlying distribution has a mean and variance defined for it—the Cauchy distribution does not, for example).

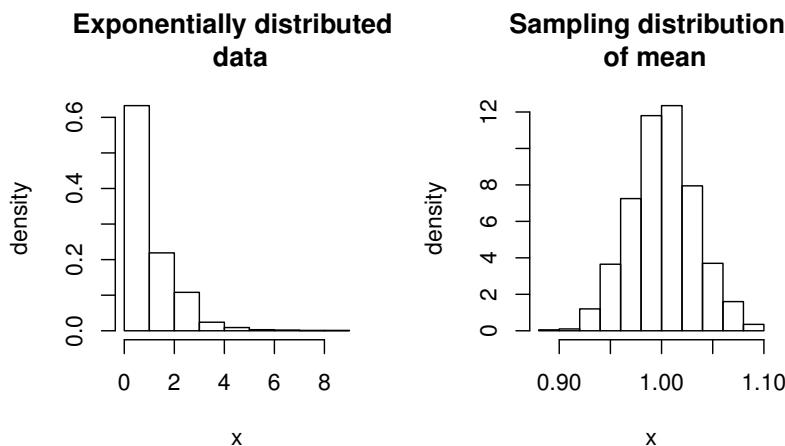
To get an intuition about this, consider the situation where we repeatedly sample from an exponential distribution. Even though the underlying distribution is not normal (see Figure 4.1), the distribu-

tion of the mean under repeated sampling is. Or let's take some other distribution, say the uniform, see Figure 4.2.

```
n_rep<-1000
samp_distrn_mean<- rep(NA,n_rep)
for(i in 1:n_rep){
  x<- rexp(1000)
  samp_distrn_mean[i]<- mean(x)
}

op<- par(mfrow=c(1,2),pty="s")
hist(x,xlab="x",ylab="density",freq=FALSE,
     main="Exponentially distributed \n data")
hist(samp_distrn_mean,xlab="x",ylab="density",freq=FALSE,
     main="Sampling distribution \n of mean")
```

Figure 4.1: Sampling from exponential.



We will now look at this asymptotic property of the distribution of

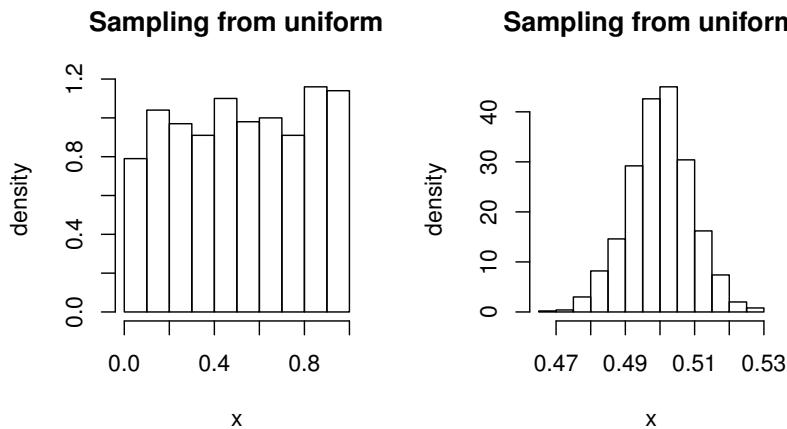
```

n_rep<-1000
samp_distrn_mean<- rep(NA,n_rep)
for(i in 1:n_rep){
  x<-runif(1000)
  samp_distrn_mean[i]<-mean(x)
}

op<-par(mfrow=c(1,2),pty="s")
hist(x,xlab="x",ylab="density",freq=FALSE,
  main ="Sampling from uniform")
hist(samp_distrn_mean,xlab="x",ylab="density",freq=FALSE,
  main="Sampling from uniform")

```

Figure 4.2: Sampling from uniform.



the sample means analytically.

#### 4.2.1 The binomial distribution

Suppose that we have found the maximum likelihood estimate, say of  $p$  in the binomial distribution. Recall that we do this by taking the first derivative  $\ell'(p)$  and then equating it to zero, and then solving for  $p$ .

It turns out that the second derivative of the log likelihood gives you an estimate of the variance of the sampling distribution of the sample mean (SDSM) that I just discussed above. The square root of this variance is called standard error (SE).

Here is an informal explanation for why the second derivative does this. The second derivative is telling us the rate at which the rate of change is happening in the slope, i.e., the rate of curvature of the curve (take a look at Figure 4.3). When the variance of the SDSM is small, then we have a fast rate of change in slope (high value for second derivative), and so if we take the inverse of the second derivative, we get a small value, an estimate of the small variance (small  $SE^2$ ). And when the variance is high, we have a slow rate of change in slope (low value for second derivative). I summarize this in Table 4.1 and a visualization is shown in Figure 4.3.

Variance of SDSM	Rate of slope change	2nd derivative
small	Fast change in slope	large
large	Slow change in slope	small

So if we invert the second derivative, we get a large value, which is an estimate of the large variance (large  $SE^2$ ).

Notice that all these second derivatives would be negative, because we are approaching a maximum as we reach the peak of the curve. So when we take an inverse to estimate the variances, we get negative values. It follows that if we were to take a negative of the inverse, we'd get a positive value.

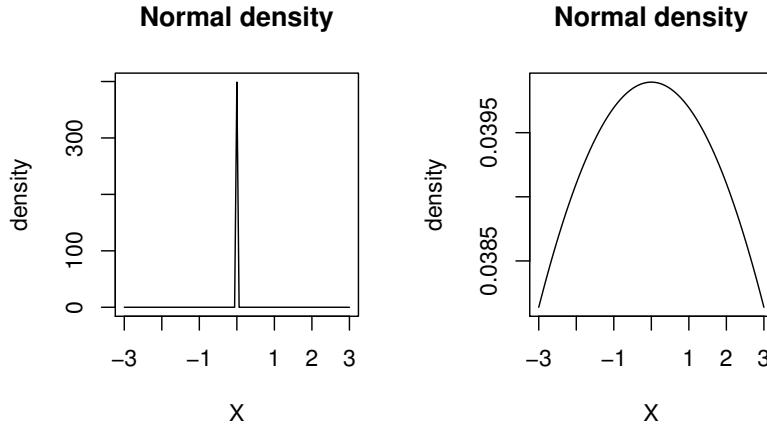
This is the reasoning that leads to the following steps for computing the variance of the SDSM:

1. Take the second partial derivative of the log-likelihood.
2. Compute the negative of the expectation of the second partial derivative. This is called the Information Matrix  $I(\theta)$ .
3. Invert this matrix to obtain estimates of the variances and covariances. To get standard errors take the square root of the diagonal elements in the matrix.

It's better to see this through an example. Let's look at the binomial distribution, which has parameter  $p$ .

Table 4.1: Variance of the SDSM and the relationship with the second derivative.

Figure 4.3: How variance relates to the second derivative.



$$L(p) = \binom{n}{x} p^x (1-p)^{n-x} \quad (4.27)$$

The Log likelihood is:

$$\ell(p) = \log \binom{n}{x} + x \log p + (n-x) \log(1-p) \quad (4.28)$$

Taking the first derivative:

$$\ell'(p) = \frac{x}{p} - \frac{n-x}{1-p} \quad (4.29)$$

Taking the second partial derivative with respect to p:

$$\ell''(p) = -\frac{x}{p^2} - \frac{n-x}{(1-p)^2} \quad (4.30)$$

The quantity  $-\ell''(p)$  is called **observed Fisher information**.

Taking expectations:

$$E(\ell''(p)) = E\left(-\frac{x}{p^2} - \frac{n-x}{(1-p)^2}\right) \quad (4.31)$$

Exploiting that fact the  $E(x/n) = p$  and so  $E(x) = E(n \times x/n) = np$ , we get

$$E(\ell''(p)) = E\left(-\frac{x}{p^2} - \frac{n-x}{(1-p)^2}\right) = -\frac{np}{p^2} - \frac{n-np}{(1-p)^2} \stackrel{\text{exercise}}{\uparrow} = -\frac{n}{p(1-p)} \quad (4.32)$$

Next, we negate and invert the expectation:

$$-\frac{1}{E(\ell''(p))} = \frac{p(1-p)}{n} \quad (4.33)$$

Evaluating this at  $\hat{p}$ , the estimated value of the parameter, we get:

$$-\frac{1}{E(\ell''(\theta))} = \frac{\hat{p}(1-\hat{p})}{n} = \frac{1}{I(p)} \quad (4.34)$$

$I(p)$  is called **expected Fisher Information**. Note that is a  $1 \times 1$  matrix, so we can call it the Information Matrix. If we take the square root of the inverse Information Matrix

$$\sqrt{\frac{1}{I(p)}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad (4.35)$$

we have the **estimated standard error**. This is the standard deviation of the sampling distribution of the sample means. Maybe a little simulation will make this clear.

```
## analytic calculation of SE from a single expt:
## number of heads in 100 coin tosses:
n<-100
p<-0.5
(x<-rbinom(1,n=n,prob=p))

## [1] 1 0 1 1 1 1 0 0 1 0 0 1 1 0 1 0 0 0 0 1 1 0 1 0 1 0 0 1 1 0 1 1 1
## [36] 0 1 1 1 0 1 0 0 1 1 1 1 0 1 0 1 0 0 0 1 1 0 1 1 1 0 1 0 0 1 0 1 0 0 0
## [71] 1 1 1 1 1 0 1 0 1 1 1 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

hat_p <- sum(x)/n
(SE_2<- (hat_p*(1-hat_p))/n)

## [1] 0.002356

(SE<-sqrt(SE_2))

## [1] 0.04853864
```

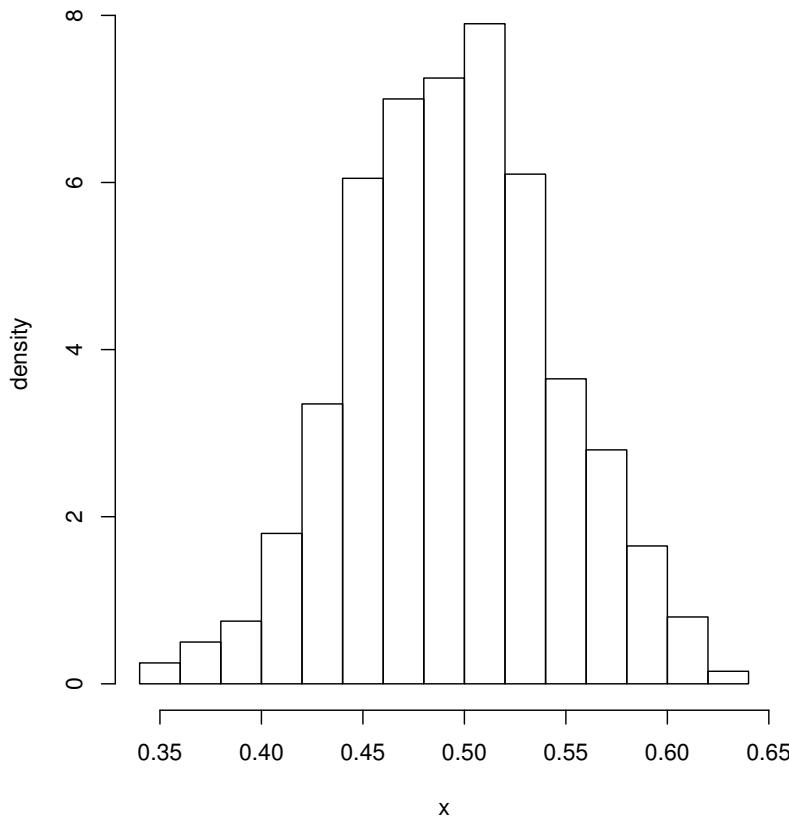
```

## by repeated sampling:
samp_distrn_means<-rep(NA,1000)
for(i in 1:1000){
  x<-rbinom(1,n=n,prob=p)
  samp_distrn_means[i]<-sum(x)/n
}
hist(samp_distrn_means,xlab="x",ylab="density",
  freq=F,main="The sampling distribution (binomial)")
## this is the SE of the SDSM:
sd(samp_distrn_means)

## [1] 0.05067106

```

**The sampling distribution (binomial)**



Here is another example, this time involving the normal distribution.

#### 4.2.2 The normal distribution

Let  $X_1, \dots, X_n$  be a sample of size  $n$  from  $N(\mu, \sigma^2)$ , both parameters of the normal, and both unknown. Let our parameters be defined as  $\theta = (\mu, \sigma^2)$ .

Here, the asymptotic distribution of  $\hat{\mu}$  and  $\hat{\sigma}^2$  is:<sup>3</sup>

$$\begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} \xrightarrow{d} N \left( \begin{pmatrix} \mu \\ \sigma \end{pmatrix}, \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix} \right) \quad (4.36)$$

What this means in practical terms is that when we have a single sample, we compute the sample mean and standard deviation,  $\hat{\mu}$  and  $\hat{\sigma}^2$ , and then compute the standard error of the sampling distribution of  $\hat{\mu}$  (usually, the sampling distribution of  $\hat{\sigma}^2$  is not interesting for inference because we are usually trying to do inference about the true mean).

We can quickly simulate the sampling distribution of the sample mean to get a feel for what the sampling distribution would look like under repeated sampling.

First, run 1000 experiments, each with sample size 100; we sample from a normal distribution with mean 500 and standard deviation 100. Then, run the experiments, and save the sample means and standard deviations or sds (this will vary from one experiment to the next due to sampling variability) and then plot the distribution of these sample means and sds.

```
nsim<-1000
n<-100
mu<-500
sigma<-100

samp_distrn_means<- rep(NA,nsim)
samp_distrn_var<- rep(NA,nsim)

for(i in 1:nsim){
  x<- rnorm(n,mean=mu, sd=sigma)
  samp_distrn_means[i]<- mean(x)
  samp_distrn_var[i]<- var(x)
}
```

The sampling distributions are shown in Figure 4.4.

Compare the standard deviation of the sampling distribution of means from the simulation with the theoretical value:

```
## estimate from simulation:
sd(samp_distrn_means)

## [1] 9.990636
```

<sup>3</sup> For details regarding the derivation, see Khuri (p. 309).

André I Khuri. *Advanced calculus with applications in statistics*, volume 486. Wiley, 2003

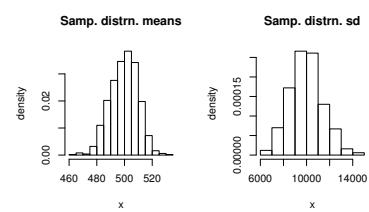


Figure 4.4: The sampling distribution of the sample means and of the standard deviation.

```
## estimate from a single sample of size n:
sigma/sqrt(n)

## [1] 10
```

Similarly, compare the variance of the sampling distribution of  $\sigma^2$  from the simulation with the theoretical value:

```
## estimate from simulation:
sd(samp_distrn_var)

## [1] 1434.255

## theoretical value:
(sqrt(2)*sigma^2)/sqrt(n)

## [1] 1414.214
```

Since we know the asymptotic distribution of the MLE, we can obtain an interval called a 95% confidence interval:

$$\hat{\mu} \pm 2SE(\hat{\mu}) \quad (4.37)$$

So, for the mean, we have a 95% confidence interval as follows:

$$\hat{\mu} \pm 2 \frac{\hat{\sigma}}{\sqrt{n}} \quad (4.38)$$

In our example:

```
## lower bound:
mu-(2*sigma/sqrt(n))

## [1] 480

## upper bound:
mu+(2*sigma/sqrt(n))

## [1] 520
```

This CI has an extremely confusing interpretation: if you were to (hypothetically) repeatedly sample, and compute the CI each time, the true mean  $\mu$  would be contained in 95% of those intervals that we hypothetically calculated each time. The confusing thing is that the single CI that you plot based on a single sample does not give you what you would intuitively expect it to: the range over which you can be 95% sure that the true parameter value  $\mu$  lies. This kind of interval can only be computed in the Bayesian setting; the CI does not have this

interpretation because  $\mu$  has no probability distribution defined over it, it is a point value.

The above simulation can be used to understand the idea of a 95% CI:

```
lower<-rep(NA,nsim)
upper<-rep(NA,nsim)
for(i in 1:nsim){
  x<-rnorm(n,mean=mu,sd=sigma)
  lower[i]<-mean(x) - 2 * sd(x)/sqrt(n)
  upper[i]<-mean(x) + 2 * sd(x)/sqrt(n)
}
## check how many CIs contain mu:
CIs<-ifelse(lower<mu & upper>mu,1,0)
table(CIs)

## CIs
##    0    1
##  54 946

## 95% CIs contain true mean:
table(CIs)[2]/sum(table(CIs))

##      1
## 0.946
```

The reason that we spent so much energy and time understanding the asymptotic properties of MLEs is that in the next chapter we will be looking maximum likelihood estimates of parameters of the linear model:

$$y = \beta_0 + \beta_1 x + \varepsilon \quad (4.39)$$

One of the issues of interest in linear models is an estimate of the uncertainty of the maximum likelihood estimates of  $\beta_0$  and  $\beta_1$ . This estimate of uncertainty is the standard error, and this is what we will depend on for doing statistical inference (null hypothesis significance testing).



# 5

## *Linear modeling*

### *5.1 Basic linear modeling theory*

Consider a deterministic function  $\phi(x, \beta)$  which takes as input some variable values  $x$  and some fixed values  $\beta$ . A simple (if abstract) example would be

$$y = \beta x \quad (5.1)$$

Another example with two fixed values  $\beta_0$  and  $\beta_1$  is:

$$y = \beta_0 + \beta_1 x \quad (5.2)$$

We can rewrite the above equation as follows.

$$\begin{aligned} y &= \beta_0 + \beta_1 x \\ &= \beta_0 \times 1 + \beta_1 x \\ &= \begin{pmatrix} 1 & x \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \\ y &= \phi(x, \beta) \end{aligned} \quad (5.3)$$

In a statistical model, we don't expect an equation like  $y = \phi(x, \beta)$  to fit all the points exactly. For example, we could come up with an equation that, given a person's weight, gives a prediction of their height:

$$\text{predicted height} = \beta_0 + \beta_1 \text{weight} \quad (5.4)$$

Given any single value of the weight of a person, we will probably not get a perfectly correct prediction of the height of that person. This leads us to a non-deterministic version of the above function:

$$y = \phi(x, \beta, \epsilon) = \beta_0 + \beta_1 x + \epsilon \quad (5.5)$$

Here,  $\epsilon$  is an error random variable which is assumed to have some PDF (the normal distribution) associated with it. It is assumed to have expectation (mean)  $0$ , and some standard deviation (to be estimated from the data)  $\sigma$ . We can write this statement in compact form as  $\epsilon \sim N(0, \sigma)$ .

The **general linear model** is a non-deterministic function like the one above:

$$Y = f(x)^T \beta + \epsilon \quad (5.6)$$

The matrix formulation will be written as below.  $n$  refers to the number of data points (that is,  $Y_1, \dots, Y_n$ ), and the index  $j$  ranges from 1 to  $n$ .

$$Y = X\beta + \epsilon \Leftrightarrow y_j = f(x_j)^T \beta + \epsilon_j, j = 1, \dots, n \quad (5.7)$$

To make this concrete, suppose we have three data points, i.e.,  $n=3$ . Then, the matrix formulation is

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \epsilon \quad (5.8)$$

$$Y = X\beta + \epsilon$$

Here,  $f(x_1)^T = (1 \ x_1)$ , and is the first row of the matrix  $X$ ,  $f(x_2)^T = (1 \ x_2)$  is the second row, and  $f(x_3)^T = (1 \ x_3)$  is the third row.

Note that  $E[Y] = X\beta$ .  $\beta$  is a  $p \times 1$  matrix, and  $X$ , the **design matrix**, is  $n \times p$ .<sup>1</sup>

### 5.1.1 Least squares estimation: Geometric argument

When we have a deterministic model  $y = \phi(f(x)^T, \beta) = \beta_0 + \beta_1 x$ , this implies a perfect fit to all data points. This is like solving the equation  $Ax = b$  in linear algebra: we solve for  $\beta$  in  $X\beta = y$  using, e.g., Gaussian elimination. When we have a non-deterministic model  $y = \phi(f(x)^T, \beta, \epsilon)$ , there is no solution. Now, the best we can do is to get  $Ax$  to be as close an approximation as possible to  $b$  in  $Ax = b$ . In other words, we try to minimize  $|b - Ax|$ .

The goal is to estimate  $\beta$ ; we want to find a value of  $\beta$  such that the observed  $Y$  is as close to its expected value  $X\beta$ . In order to be able to identify  $\beta$  from  $X\beta$ , the linear transformation  $\beta \rightarrow X\beta$  should be one-to-one, so that every possible value of  $\beta$  gives a different  $X\beta$ . This in turn requires that  $X$  be of full rank  $p$ . So, if  $X$  is an  $n \times p$  matrix, then it is necessary that  $n \geq p$ . There must be at least as many observations as parameters. If this is not true, then the model is said to be **over-parameterized**.

<sup>1</sup> In this example with three data points, if the model had been

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

what would each row of  $f(x)^T$  look like?

Rank is the number of linearly independent columns or rows. The row rank and column rank of an  $m \times n$  matrix will be the same, so we can just talk of rank of a matrix. An  $m \times n$  matrix  $X$  with  $\text{rank}(X)=\min(m,n)$  is called full rank.

Assuming that  $X$  is of full rank, and that  $n > p$ ,  $Y$  can be considered a point in  $n$ -dimensional space and the set of candidate  $X\beta$  is a  $p$ -dimensional subspace of this space; see my admittedly clunky Figure 5.1. There will be one point in this subspace which is closest to  $Y$  in terms of Euclidean distance. The unique  $\beta$  that corresponds to this point is the **least squares estimator** of  $\beta$ ; we will call this estimator  $\hat{\beta}$ .

Notice that  $\epsilon = (Y - X\hat{\beta})$  and  $X\beta$  are perpendicular to each other. Because the dot product of two perpendicular (orthogonal) vectors is 0, we get the result:

$$(Y - X\hat{\beta})^T X\beta = 0 \Leftrightarrow (Y - X\hat{\beta})^T X = 0 \quad (5.9)$$

Multiplying out the terms, we proceed as follows. One result that we use here is that  $(AB)^T = B^T A^T$ .

$$\begin{aligned} (Y - X\hat{\beta})^T X &= 0 \\ (Y^T - \hat{\beta}^T X^T)X &= 0 \\ \Leftrightarrow Y^T X - \hat{\beta}^T X^T X &= 0 \\ \Leftrightarrow Y^T X &= \hat{\beta}^T X^T X \\ \Leftrightarrow (Y^T X)^T &= (\hat{\beta}^T X^T X)^T \\ \Leftrightarrow X^T Y &= X^T X \hat{\beta} \end{aligned} \quad (5.10)$$

This gives us the important result:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (5.11)$$

$X$  is of full rank, therefore  $X^T X$  is invertible.

**Example:**

```
(X<-matrix(c(rep(1,8),rep(c(-1,1),each=4),
           rep(c(-1,1),each=2,2)),ncol=3))

##      [,1] [,2] [,3]
## [1,]     1   -1   -1
## [2,]     1   -1   -1
## [3,]     1   -1    1
## [4,]     1   -1    1
## [5,]     1    1   -1
## [6,]     1    1   -1
## [7,]     1    1    1
## [8,]     1    1    1

library(Matrix)
## full rank:
rankMatrix(X)
```

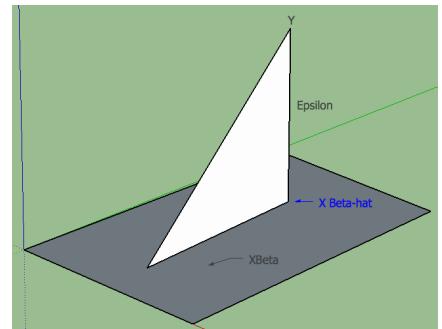


Figure 5.1: Geometric interpretation of least squares.

```

## [1] 3
## attr(,"method")
## [1] "tolNorm2"
## attr(,"useGrad")
## [1] FALSE
## attr(,"tol")
## [1] 1.776357e-15

## det non-zero:
det(t(X) %*% X)

## [1] 512

```

Notice that the inverted matrix is also symmetric. We will use this fact soon.

The matrix  $V = X^T X$  is a symmetric matrix, which means that  $V^T = V$ . The symmetric matrix will be of great interest to us in this course.

### 5.1.2 The expectation and variance of the parameters beta

Our model is:

$$Y = X\beta + \epsilon \quad (5.12)$$

Let  $\epsilon \sim N(0, \sigma^2)$ . In other words, we are assuming that each value generated by the random variable  $\epsilon$  is independent and it has the same distribution, i.e., it is identically distributed. This is sometimes shortened to the iid assumption. So we should technically be writing:

$$Y = X\beta + \epsilon \quad \epsilon \sim N(0, \sigma^2) \quad (5.13)$$

and add that  $Y$  are independent and identically distributed. Note that the independence assumption is grossly violated in our Hindi data—we have multiple measures from each subject, and multiple measures also from each item. So it is not legitimate to fit such a model to our data (this is HW4).

Some consequences of the above statements:

1.  $E(\epsilon) = 0$
2.  $Var(\epsilon) = \sigma^2 I_n$
3.  $E[Y] = X\beta = \mu$
4.  $Var(Y) = \sigma^2 I_n$

We can now derive the expectation and variance of the estimators  $\hat{\beta}$ . We need a fact about variances:  $\text{Var}(aB)$ , where  $a$  is a constant, is  $a^2\text{Var}(B)$ . In the matrix setting,  $\text{Var}(AB)$ , where  $A$  is a constant, is  $A\text{Var}(B)A^T$ .

$$E[\hat{\beta}] = E[(X^T X)^{-1} X^T Y] = (X^T X)^{-1} X^T X \beta = \beta \quad (5.14)$$

Notice that the above shows that  $\hat{\beta}$  is an unbiased estimator of  $\beta$ .

Next, we compute the variance:

$$\text{Var}(\hat{\beta}) = \text{Var}([(X^T X)^{-1} X^T]Y) \quad (5.15)$$

Expanding the right hand side out:

$$\text{Var}([(X^T X)^{-1} X^T]Y) = [(X^T X)^{-1} X^T] \text{Var}(Y) [(X^T X)^{-1} X^T]^T \quad (5.16)$$

Replacing  $\text{Var}(Y)$  with its variance  $\sigma^2 I$ , and unpacking the transpose on the right-most expression  $[(X^T X)^{-1} X^T]^T$ :

$$\text{Var}(\hat{\beta}) = [(X^T X)^{-1} X^T] \sigma^2 I X [(X^T X)^{-1}]^T \quad (5.17)$$

Since  $\sigma^2$  is a scalar we can move it to the left, and any matrix multiplied by  $I$  is the matrix itself, so we ignore  $I$ , getting:

$$\text{Var}(\hat{\beta}) = \sigma^2 [(X^T X)^{-1} X^T X [(X^T X)^{-1}]^T \quad (5.18)$$

Since  $(X^T X)^{-1} X^T X = I$ , we can simplify to

$$\text{Var}(\hat{\beta}) = \sigma^2 [(X^T X)^{-1}]^T \quad (5.19)$$

Now,  $(X^T X)^{-1}$  is symmetric, so  $[(X^T X)^{-1}]^T = (X^T X)^{-1}$ . This gives us:

$$\text{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1} \quad (5.20)$$

An example:

```
y<-as.matrix(hindi10$TFT)
x<-log(hindi10$word_len)
m0<-lm(y~x)

## design matrix:
X<-model.matrix(m0)
head(X, n=4)

##   (Intercept)      x
## 1           1 0.6931472
```

```

## 2          1 1.3862944
## 3          1 1.3862944
## 4          1 2.4849066

## (X^TX)^{-1}
invXTX<-solve(t(X)%%X)
## estimate of beta:
(beta<-invXTX%*%t(X)%*%y)

##          [,1]
## (Intercept) 210.7777
## x          129.4064

## estimated variance (se^2) of the estimate of beta:
(hat_sigma<-summary(m0)$sigma)

## [1] 269.5443

(hat_var<-hat_sigma^2*invXTX)

##          (Intercept)      x
## (Intercept) 31.35647 -21.61119
## x          -21.61119 16.88379

```

What we have here is a bivariate normal distribution as an estimate of the  $\beta$  parameters:

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \sim N\left(\begin{pmatrix} 210.7777 \\ 129.4064 \end{pmatrix}, \begin{pmatrix} 31.35647 & -21.61119 \\ -21.61119 & 16.88379 \end{pmatrix}\right) \quad (5.21)$$

The variance of a bivariate distribution has the variances along the diagonal, and the covariance between  $\beta_0$  and  $\beta_1$  on the off-diagonals. Covariance is defined as:

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \rho \sigma_{\hat{\beta}_0} \sigma_{\hat{\beta}_1} \quad (5.22)$$

where  $\rho$  is the correlation between  $\beta_0$  and  $\beta_1$ .

So  $\hat{\beta}_0 \sim N(210.78, 31.36)$  and  $\hat{\beta}_1 \sim N(129.41, 16.88)$ , and  $\text{Cov}(\beta_0, \beta_1) = -21.61$ . So the correlation between the  $\hat{\beta}$  is

```

## hat rho:
-21.61/(sqrt(31.36)*sqrt(16.88))

## [1] -0.9392485

```

### 5.1.3 Statistical inference

In the model output we see:

```
round(summary(m0)$coefficients[,1:3],
      digits=3)

##           Estimate Std. Error t value
## (Intercept) 210.778     5.600 37.641
## x          129.406     4.109 31.493
```

We know what the first two columns are. The third column is based on the following quantity:

$$t^2 = \frac{(\hat{\beta}_0 - \beta_0)^2}{Var(\hat{\beta}_0)} \quad (5.23)$$

Note the difference between  $\hat{\beta}_0$  and  $\beta_0$ . The former is the estimate from the data, and the latter is the true value (a point value) of the parameter.

This is called the Wald statistic, and the test, which is called a Wald test, says that the following quantity has a chi-squared distribution (with degrees of freedom p, the number of parameters). Consider  $\beta_0$ , i.e., p=1:

$$t^2 = \frac{(\hat{\beta}_0 - \beta_0)^2}{Var(\hat{\beta}_0)} \sim \chi_1^2 \quad (5.24)$$

An alternative version of the test is:

$$\frac{\hat{\beta} - \beta}{\sqrt{Var(\hat{\beta})}} \sim Normal(0, 1) \quad (5.25)$$

The t-value refers to the fact that we are using an approximation of  $N(0,1)$ , the t-distribution with  $n - 1$  degrees of freedom. For small sample sizes (say  $n < 18$ ), the use of the t-distribution rather than the normal has important consequences because the t-distribution for such small n has somewhat fatter tails than the normal—slightly more probability is concentrated in the tails of the t-distribution than in the normal. For larger sample sizes, the normal and t-distribution are essentially indistinguishable. We can visualize this difference between the t-distribution and normal distribution quite easily.

```
range <- seq(-4, 4, .01)

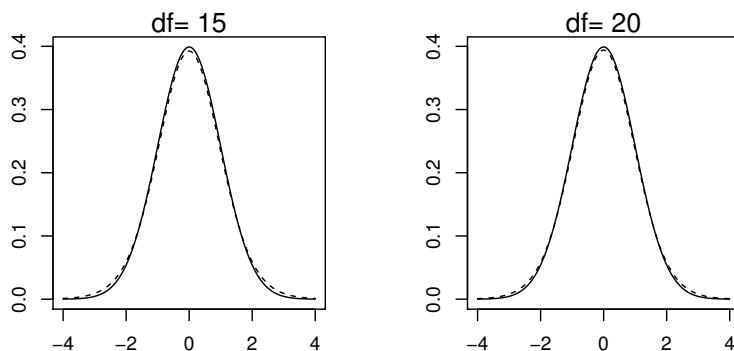
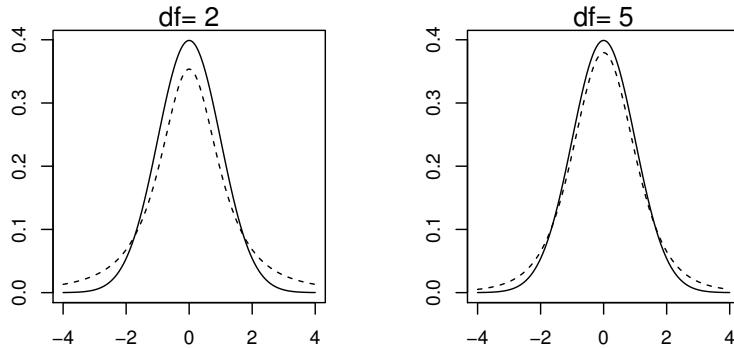
op<-par(mfrow=c(2,2),pty="s")

for(i in c(2,5,15,20)){
  plot(range,dnorm(range),type="l",lty=1,
       xlab="",ylab="",
       cex.axis=1)
```

```

  lines(range, dt(range,df=i),lty=2,lwd=1)
  mtext(paste("df=",i),cex=1.2)
}

```



#### 5.1.4 The p-value, and Type S and M errors

Note that R also prints out a “p-value”:

```

summary(m0)$coef

##             Estimate Std. Error   t value    Pr(>|t|)
## (Intercept) 210.7777  5.599685 37.64100 1.800690e-299
## x           129.4064  4.108989 31.49348 1.968918e-212

```

This is the **conditional** probability of getting an estimate as extreme or more extreme than the absolute value  $|\pm 210.78|$  for the estimate of  $\beta_0$ , assuming that the true distribution is  $N(0, \text{Var}(\hat{\beta}_0))$ , i.e., assuming that the true  $\beta_0 = 0$ .

We can compute it by hand using the CDF of the normal or t-distribution (there are some rounding errors piling up here, so the numbers don't match the lm output exactly):

```
2*pnorm(210.78,mean=0,sd=sqrt(31.36),
        lower.tail=FALSE)

## [1] 0

2*pt(210.78/sqrt(31.36),df=length(y)-1,
     lower.tail=FALSE)

## [1] 1.910062e-299
```

Because the conditional probability (the p-value) is essentially zero, we feel entitled to reject the null hypothesis that  $\beta_0 = 0$ . Rejecting this null hypothesis implies that we conclude that  $\beta_0 \neq 0$ .

One important fact to remember is that when we do a statistical test and obtain a p-value, if the null hypothesis is in fact true, then the distribution of the p-value under repeated sampling is Uniform(0,1). We prove this below.

The significance of this fact is that a low p-value from a single experiment doesn't necessarily tell us anything useful; if the null is true (and we can't be sure it isn't), then this p-value is coming from a uniform distribution, i.e., every value between 0 and 1 has the same likelihood. A low p-value isn't anything special in this situation.

*The distribution of the p-value under the null hypothesis is Uniform(0,1)*  
 Notice that when a random variable Z comes from a Uniform(0,1) distribution, then  $P(Z < z) = z$ . Consider some examples:

1. when  $z = 0$ , then  $P(Z < 0) = 0$ ;
2. when  $z = 0.25$ , then  $P(Z < 0.25) = 0.25$ ;
3. when  $z = 0.5$ , then  $P(Z < 0.5) = 0.5$ ;
4. when  $z = 0.75$ , then  $P(Z < 0.75) = 0.75$ ;
5. when  $z = 1$ , then  $P(Z < 1) = 1$ .

With this background, we prove the following proposition:

**Proposition 2** *If a random variable  $Z = F(T)$ , i.e., if Z is the CDF for a random variable T, then  $Z \sim \text{Uniform}(0,1)$ .*

Note here that the p-value is a random variable, call it Z. The p-value is the CDF of the random variable T that is a transformation of

the random variable  $\bar{X}$ :  $T = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ . This random variable has a CDF  $F(T)$ .

So, if we can prove the above proposition, we have shown that the p-value's distribution under the null hypothesis is *Uniform(0,1)*. The proof is actually really amazing; it's called the **probability integral transform**.

**Proof:**

Let  $Z = F(T)$ .

$$\begin{aligned}
 P(Z \leq z) &= P(F(T) \leq z) \\
 &= P(F^{-1}F(T) \leq F^{-1}(z)) \\
 &= P(T \leq F^{-1}(z)) \\
 &= F(F^{-1}(z)) \\
 &= z
 \end{aligned} \tag{5.26}$$

■

Please read our review article <sup>2</sup> for more discussion on p-values.

*Type I and II errors, Type S and M errors* In frequentist statistics, we can also compute Type I and Type II error rates. Type I error is the probability of incorrectly rejecting the null (when it's actually true); this is typically set at 0.05 by the researcher and is called the  $\alpha$  value. Type II error is defined as the probability of incorrectly "accepting" (more accurately, failing to reject) the null hypothesis when it's false. (1-Type II) error is called power, and is the probability of correctly rejecting the null.<sup>3</sup>

Another important point is that just computing the p-value is not particularly informative. If you have some way to determine an estimate of the true effect size for a particular phenomenon (say, through a meta-analysis or literature review or expert knowledge about the topic you are studying), then you can and should (I would say, must) also compute Type S and M errors; see the Gelman and Carlin article<sup>4</sup> for further discussion.

For example, if your true effect size is believed to be  $D=15$ , then we can compute (apart from statistical power) these error rates, which are defined as follows:

1. Type S error: the probability that the sign of the effect is incorrect, given that (a) the result is statistically significant, or (b) the result is statistically non-significant.
2. Type M error: the expectation of the ratio of the absolute magnitude of the effect to the hypothesized true effect size (conditional on whether the result is significant or not). Gelman and Carlin also

<sup>2</sup> Shravan Vasishth and Bruno Nicenboim. Statistical methods for linguistic research: Foundational ideas – Part I, 2016

<sup>3</sup> Note that all our definitions here are with respect to the null hypothesis—it is a mistake to think that Type II error is the probability of failing to accept the alternative hypothesis when it's true. We can only ever reject or not reject the null; our hypothesis test is always with reference to the null.

<sup>4</sup> Andrew Gelman and John Carlin. Beyond power calculations assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6):641–651, 2014

call this the exaggeration ratio, which is perhaps more descriptive than “Type M error”.

Suppose a particular study has standard error 46, and sample size 37. And suppose that our estimated true D=15. Then, we proceed as follows:

```
## probable effect size derived from past studies:
D<-15
## SE from the study of interest:
se<-46
stddev<-se*sqrt(37)
nsim<-10000
drep<-rep(NA,nsim)
for(i in 1:nsim){
drep[i]<-mean(rnorm(37,mean=D, sd=stddev))
}

##power: a depressingly low 0.056
pow<-mean(ifelse(abs(drep/se)>2,1,0))

## which cells in drep are significant at alpha=0.05?
signif<-which(abs(drep/se)>2)

## Type S error rate | signif: 19%
types_sig<-mean(drep[signif]<0)
## Type S error rate | non-signif: 37%
types_nonsig<-mean(drep[-signif]<0)

## Type M error rate | signif: 7
typem_sig<-mean(abs(drep[signif])/D)
## Type M error rate | not-signif: 2.3
typem_nonsig<-mean(abs(drep[-signif])/D)
```

So, you can see that the Type S error and the exaggeration ratio, conditional on a result being significant, are pretty high. The practical implication of this is that if most studies in psycholinguistics are low powered, then it doesn't matter much whether you got a significant result or not. You could be (and probably are) barking up the wrong tree. The main take-away point here is: run high powered studies, and replicate the results. There's really no statistical test out there that can match consistent replication.

### 5.1.5 Hypothesis tests and the sampling distribution of the mean

An important detail about the above Wald test or statistic is that it that the null hypothesis is expressed not over the data  $X_1, X_2, \dots, X_n$  generated by a random variable  $X$ , but over the **sampling distribution of the mean** of such data under repeated sampling. We discussed the sampling distribution in section 4.2, but I present the same idea in a different way below.

Suppose I gather independent and identically distributed data  $x_1, \dots, x_n$ , each of which is generated by a random variable  $X$ .

For each sample, suppose I compute the mean  $\bar{x}$ . Now, the estimator  $\bar{X}$  is also a random variable; it is just a linear combination of values generated by instances of the random variable  $X$ , which, we will assume, has some mean (expectation)  $\mu$  and some variance  $\sigma^2$ :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{n} X_1 + \dots + \frac{1}{n} X_n \quad (5.27)$$

So, its expectation is

$$\begin{aligned} E[\bar{X}] &= E\left[\frac{1}{n} X_1 + \dots + \frac{1}{n} X_n\right] \\ &= \frac{1}{n}(E[X] + \dots + E[X]) \\ &= \frac{1}{n}(\mu + \dots + \mu) \quad (5.28) \\ &= \frac{1}{n} n\mu \\ &= \mu \end{aligned}$$

And its variance is

$$\begin{aligned} Var(\bar{X}) &= Var\left(\frac{1}{n} X_1 + \dots + \frac{1}{n} X_n\right) \\ &= \frac{1}{n^2} Var(X_1 + \dots + X_n) \quad (5.29) \end{aligned}$$

Now,  $X_1, \dots, X_n$  are independent. We will use the fact that the variance of the sum of independent RVs is the sum of their variances. (If they were not independent, then the variance of the sum would have to take covariance between the  $X$ 's into account; more on this later). This gives us:

$$\begin{aligned}
\frac{1}{n^2} \text{Var}(X_1 + \dots + X_n) &= \frac{1}{n^2} (\text{Var}(X) + \dots + \text{Var}(X)) \\
&= \frac{1}{n^2} n \text{Var}(X) \\
&= \frac{1}{n} \text{Var}(X) \\
&= \frac{\sigma^2}{n}
\end{aligned} \tag{5.30}$$

We have derived the very important result that the expectation (i.e., the mean) and variance of the sampling distribution of the sample means are

$$E[\bar{X}] = \mu \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n} \tag{5.31}$$

The practical implication of this result is huge. From a *single* sample  $x_1, \dots, x_n$ , we can derive the distribution of hypothetical sample means under repeated sampling. That is, we can say something about what the plausible and implausible values of the sample mean are. This is the basis for all hypothesis testing and statistical inference in the frequentist framework we are studying.

Note that I was careful above to not stipulate that  $X$  is normally distributed. As discussed in section 4.2, one amazing fact is that, as long as  $X$  has an expectation and variance defined for it, the sampling distribution of the sample mean  $\bar{X}$  will have a normal distribution if sample size  $n$  is large enough. This statement is called the Central Limit Theorem (see page 267 of Freund's Mathematical Statistics<sup>5</sup> for a proof), which I will write compactly as:

<sup>5</sup> I. Miller and M. Miller. *John E. Freund's Mathematical Statistics with Applications*. Prentice Hall, 2004

### Theorem 1 Central Limit Theorem

Let  $f(X)$  be the pdf of a random variable  $X$ , and assume that the pdf has mean  $\mu$  and variance  $\sigma^2$ . Then:

$$\bar{X} \sim N(\mu, \sigma^2/n) \quad X \sim f(X), E[X] = \mu, \text{Var}(X) = \sigma^2 \quad n \text{ large} \tag{5.32}$$

In our linear model example above, the variance estimate of  $\beta$  is an estimate of  $\sigma^2/n$ . This is the square of the standard error ( $\sigma/\sqrt{n}$ ). Make sure you distinguish it from the standard deviation  $\sigma$ ; note that  $\sigma/\sqrt{n}$  is also a standard deviation, but it's the standard deviation of the sampling distribution of the sample mean.

#### 5.1.6 Hypothesis testing using the likelihood ratio

Suppose now that we have some data  $x_1, \dots, x_n$  from a random variable  $X$ <sup>6</sup> whose distribution depends on the parameter  $\theta$ . Suppose also

<sup>6</sup> I will hereafter assume that data are iid unless stated otherwise.

that we want to test a hypothesis  $H_0$  against  $H_1$ .

Define the **likelihood ratio test statistic** as follows. Here,  $L(\theta)$  refers to the likelihood given some values of  $\theta$ , and  $\ell(\theta)$  refers to the log likelihood.

$$\begin{aligned}\lambda &= 2 \times \log(L(\theta_1)/L(\theta_0)) \\ &= 2 \times \{\ell(\theta_1) - \ell(\theta_0)\} \\ \lambda &= 2 \times \{\ell(\theta_1) - \ell(\theta_0)\}\end{aligned}\tag{5.33}$$

where  $\theta_1$  and  $\theta_0$  are the estimates of  $\theta$  under the alternative and null hypotheses, respectively. The likelihood ratio test rejects  $H_0$  if  $\lambda$  is sufficiently large. As the sample size approaches infinity,

$$\lambda = \chi_r^2\tag{5.34}$$

where  $r$  is called degrees of freedom and is the difference in the number of parameters estimated under  $H_1$  and  $H_0$ . This is called Wilks' theorem. The proof of Wilks' theorem is fairly involved but you can find it on the internet if you are interested.

Note that sometimes you will see the form:

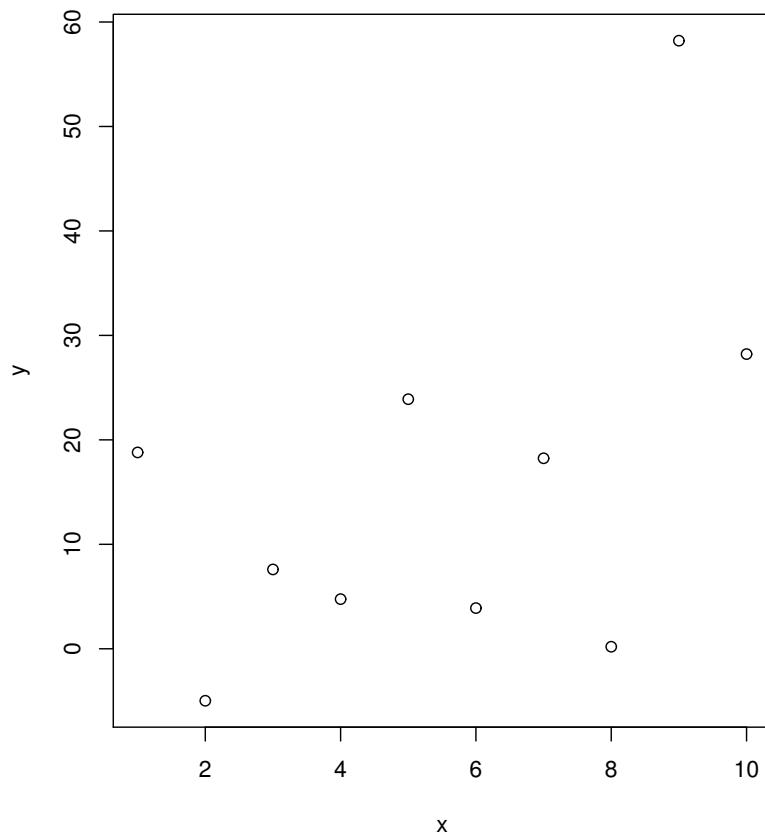
$$\lambda = -2\{\ell(\theta_0) - \ell(\theta_1)\}\tag{5.35}$$

It should be clear that both statements are saying the same thing; in the second case, we are just subtracting the alternative hypothesis log likelihood from the null hypothesis log likelihood.

A practical example will make the usage of this test clear. Let's just simulate a linear model:

```
x<-1:10
y<- 10 + 2*x+rnorm(10, sd=10)
```

```
plot(x,y)
```



```
## null hypothesis model:
m0<-lm(y~1)
## alternative hypothesis model:
m1<-lm(y~x)
```

```
lambda<- -2*(logLik(m0)-logLik(m1))
## observed value:
lambda[1]
## [1] 3.008948
## critical value:
qchisq(0.95,df=1)
## [1] 3.841459
# p-value:
pchisq(lambda[1],df=1,lower.tail=FALSE)
## [1] 0.08280602
```

Here, we fit the null hypothesis model which only has an intercept term  $\beta_0$ , and the alternative model that has  $\beta_1$  as well. Finally, we compare the  $\lambda$  with the critical chi-squared value for degrees of freedom 1. We also computed the probability of getting a  $\lambda$  as extreme as we got assuming that the null is true:

Note that in the likelihood test above, we are comparing one nested model against another: the null hypothesis model is nested inside the alternative hypothesis model. What this means is that the alternative hypothesis model contains all the parameters in the null hypothesis model plus some others.

Another way to test hypotheses is to use analysis of variance, or ANOVA.

### 5.1.7 Hypothesis testing using Analysis of variance (ANOVA)

We can compare two models, one nested inside another, as follows:

```
anova(m0,m1)

## Analysis of Variance Table
##
## Model 1: y ~ 1
## Model 2: y ~ x
##   Res.Df   RSS Df Sum of Sq    F Pr(>F)
## 1      9 3039.3
## 2      8 2249.6  1    789.75 2.8085 0.1323
```

The F-score you get here is actually the square of the t-value you get in the linear model summary:

```
sqrt(anova(m0,m1)$F[2])

## [1] 1.675869

summary(m1)$coefficients[2,3]

## [1] 1.675869
```

This is because  $t^2 = F$ . The proof is discussed on page 9 of the Dobson and Barnett book.

The ANOVA works as follows. First define the residual as:<sup>7</sup>

$$e = Y - X\hat{\beta} \quad (5.36)$$

The square of this is:

$$e^T e = (Y - X\hat{\beta})^T (Y - X\hat{\beta}) \quad (5.37)$$

<sup>7</sup> Note that I do not use  $e$  here, but  $e$ , to refer to the residual. This is because these are the estimates of  $e$  derived from the estimates  $\hat{\beta}$ .

Define the **deviance** as:

$$\begin{aligned}
 D &= \frac{1}{\sigma^2} (Y - X\hat{\beta})^T (Y - X\hat{\beta}) \\
 &= \frac{1}{\sigma^2} (Y^T - \hat{\beta}^T X^T)(Y - X\hat{\beta}) \\
 &= \frac{1}{\sigma^2} (Y^T Y - Y^T X\hat{\beta} - \hat{\beta}^T X^T Y + \hat{\beta}^T X^T X\hat{\beta}) \\
 &= \frac{1}{\sigma^2} (Y^T Y - \hat{\beta}^T X^T Y)
 \end{aligned} \tag{5.38}$$

Notice that  $-Y^T X\hat{\beta} + \hat{\beta}^T X^T Y = 0$ ; they are scalar  $1 \times 1$ .

Assume that we have data of size  $n$ . Now suppose we have a null hypothesis  $H_0 : \beta = \beta_0$  and an alternative hypothesis  $H_1 : \beta = \beta_1$ . Let the null hypothesis have  $q$  parameters, and the alternative  $p$ , where  $q < p < n$ . Let  $X_0$  be the design matrix for  $H_0$ , and  $X_1$  the design matrix for  $H_1$ . Compute the deviances  $D_0$  and  $D_1$  for each hypothesis, and compute  $\Delta D$ :

$$\begin{aligned}
 \Delta D &= D_0 - D_1 = \frac{1}{\sigma^2} [(Y^T Y - \hat{\beta}_0^T X_0^T Y) - (Y^T Y - \hat{\beta}_1^T X_1^T Y)] \\
 &= \frac{1}{\sigma^2} [\hat{\beta}_1^T X_1^T Y - \hat{\beta}_0^T X_0^T Y]
 \end{aligned} \tag{5.39}$$

It turns out that the F-statistic has the following distribution if the null hypothesis is true:

$$F = \frac{\Delta D / (p - q)}{D_1 / (n - p)} \sim F(p - q, n - p) \tag{5.40}$$

So, an extreme value of F is inconsistent with the null and we reject it.

The F-statistic is:

$$\begin{aligned}
 F &= \frac{\Delta D / (p - q)}{D_1 / (n - p)} \\
 &= \frac{\hat{\beta}_1^T X_1^T Y - \hat{\beta}_0^T X_0^T Y}{p - q} / \frac{Y^T Y - \hat{\beta}_1^T X_1^T Y}{n - p}
 \end{aligned} \tag{5.41}$$

Traditionally, the way the F-test is summarized is:

There is much more to say here about ANOVA, but this is the basic idea.

### 5.1.8 Multiple regression

You are already familiar with multiple regression from the Gelman and Hill book, so I will not discuss this in much detail, except to note that in multiple regression an important issue is **multicollinearity**.

Source of variance	df	Sum of squares	Mean square
Model with $\beta_0$	q	$\beta_0^T X_0^T Y$	
Improvement due to $\beta_1$	p-q	$\hat{\beta}_1 X_1^T Y - \hat{\beta}_0^T X_0^T Y$	$\frac{\hat{\beta}_1 X_1^T Y - \hat{\beta}_0^T X_0^T Y}{p-q}$
Residual	n-p	$Y^T Y - \hat{\beta}_1^T X_1^T Y$	$\frac{Y^T Y - \hat{\beta}_1^T X_1^T Y}{n-p}$
Total	n	$y^T y$	

Table 5.1: default

This occurs when multiple predictors are highly correlated. The consequence of this is that  $X^T X$  can be nearly singular and the estimation equation

$$X^T X \beta = X^T Y \quad (5.42)$$

is ill-conditioned: small changes in the data can cause large changes in  $\beta$  (signs will flip for example). Also, some of the elements of  $\sigma^2(X^T X)^{-1}$  will be large—standard errors and covariances can be large.

We can check for multicollinearity using the Variance Inflation Factor, VIF.

Suppose you have fitted a model with several predictors. The definition of  $VIF_j$  for a predictor j is:

$$VIF_j = \frac{1}{1 - R_j^2} \quad (5.43)$$

where  $R_j^2$  is called the coefficient of determination, and quantifies goodness of fit of the model. We define this next.

Recall that the residual sum of squares is

$$e^T e = (Y - X\hat{\beta})^T (Y - X\hat{\beta}) = \hat{S} \quad (5.44)$$

We can compare this sum of squares with the minimal model, which has  $E[Y] = \mu$ . In this model, the design matrix is an  $n \times 1$  design matrix, so  $X^T X = n$ :

```
X<-matrix(rep(1,10),ncol=1)
##
t(X) %*% X
##
[,1]
## [1,] 10
```

Note that I am using  $e$  here, not  $\epsilon$ .  $e$  is computed from the estimated  $\hat{\beta}$ .

and  $X^T Y = \sum y_i$ . And  $\hat{\beta} = \hat{\mu} = \bar{y}$ . So, for this simple model, the sum of squares  $\hat{S}_0$  is:

$$\hat{S}_0 = Y^T Y - n\bar{y}^2 \quad (5.45)$$

$R^2$  is defined as follows:

$$R^2 = \frac{\hat{S}_0 - \hat{S}}{\hat{S}_0} = \frac{\hat{\beta} X^T Y - n\bar{y}^2}{Y^T Y - n\bar{y}^2} \quad (5.46)$$

The interpretation of  $R^2$  is that it represents the proportion of variance explained by the model. Note that  $R^2$  always increases if predictors are added, so an adjustment is made, called adjusted  $R^2$ :

$$R_{Adj}^2 = 1 - \frac{((Y - X\hat{\beta})^T(Y - X\hat{\beta}))/ (n - p)}{(Y^T Y - n\bar{y}^2)/ (n - 1)} \quad (5.47)$$

Now for a practical example. Consider word length and syllable length as predictors in the Hindi data:

```
library(car)
vif(lm(TFT~syll_len+word_len,hindi10))

## syll_len word_len
## 3.425663 3.425663
```

Here is a somewhat worse situation:

```
m<-lm(TFT ~ word_complex + word_freq + type_freq+
       word_bifreq + type_freq+
       word_len + IC + SC,
       hindi10)
summary(m)

##
## Call:
## lm(formula = TFT ~ word_complex + word_freq + type_freq + word_bifreq +
##      type_freq + word_len + IC + SC, data = hindi10)
##
## Residuals:
##    Min      1Q Median      3Q     Max
## -524.8 -149.4  -66.2   74.8 3511.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.315e+02  8.047e+00  28.763 < 2e-16
## word_complex -3.235e+01  4.452e+00 -7.268 3.84e-13
## word_freq    -9.432e-04  9.789e-04 -0.964  0.3353
## type_freq    -9.898e-04  5.185e-04 -1.909  0.0563
## word_bifreq  -1.043e-02  5.575e-03 -1.871  0.0613
## word_len      3.540e+01  1.764e+00  20.061 < 2e-16
## IC            1.055e-01  5.042e-01  0.209  0.8342
```

```

## SC           1.920e+01  3.896e+00   4.927 8.45e-07
##
## Residual standard error: 248.6 on 14884 degrees of freedom
##   (4775 observations deleted due to missingness)
## Multiple R-squared:  0.05745, Adjusted R-squared:  0.057
## F-statistic: 129.6 on 7 and 14884 DF,  p-value: < 2.2e-16

round(vif(m),digits=3)

## word_complex   word_freq    type_freq  word_bifreq   word_len
##          1.874      4.425       4.455      1.089       2.437
##          IC          SC
##          1.202      1.238

```

If the predictors are uncorrelated, VIF will be near 1 in each case. Dobson et al mention that VIF of greater than 5 is cause for worry.

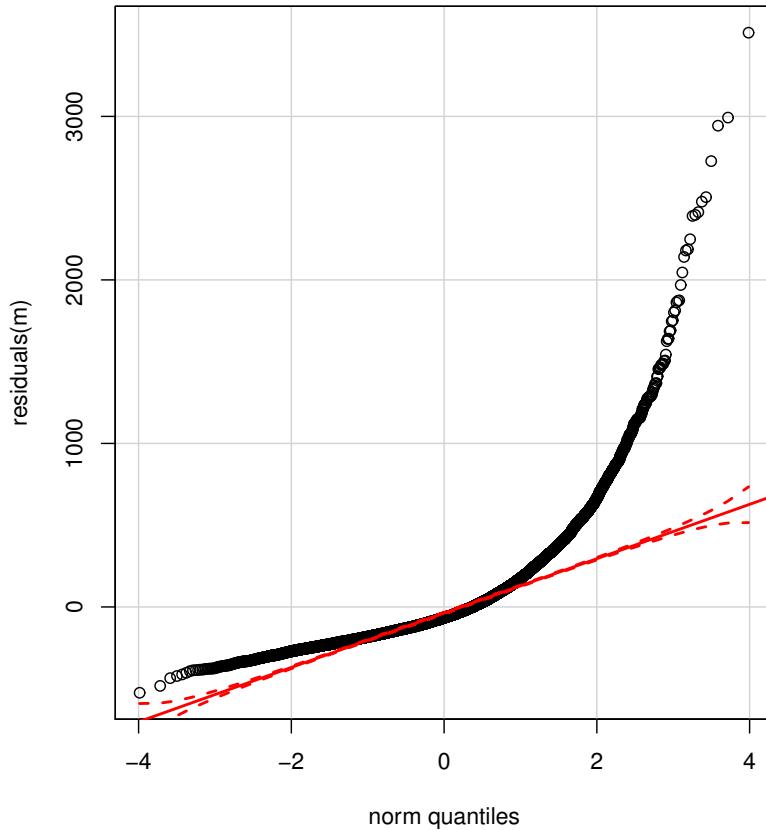
### 5.1.9 Checking model assumptions

In practical terms, the first thing you need to check is whether the residuals are normally distributed. This can be done by plotting the residuals against the quantiles of the normal distribution:

```

library(car)
qqPlot(residuals(m))

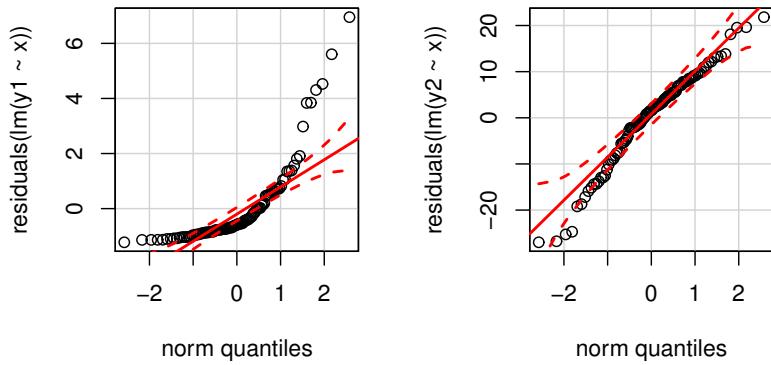
```



I have heard people say that there is no need to check for normality of residuals; indeed, Gelman and Hill state that it is the least important assumption in linear models. However, this is a highly misleading statement and should be disregarded when the goal is null hypothesis testing.

The normality assumption is necessary for hypothesis testing, but one other consequence of a violation of normality in linguistics is that it can reduce statistical power. We can test this with a simulation. Let's simulate data with non-normal residuals:

```
op<-par(mfrow=c(1,2),pty="s")
x<-1:100
y1<- 10 + 2*x+rchisq(100,df=1)
qqPlot(residuals(lm(y1~x)))
y2<- 10 + 2*x+rnorm(100,sd=10)
qqPlot(residuals(lm(y2~x)))
```



We know that  $H_0 : \beta_1 = 0$  is false: it's 0.01. So here is an example of how often the statistical test fails to detect this significant effect compared to the case when the residual is normal.

```
nsim<-1000
n<-100
x<-1:n
store_y1_results<-rep(NA,nsim)
store_y2_results<-rep(NA,nsim)
for(i in 1:nsim){
  e<-rchisq(n,df=1)
  e<-scale(e,scale=F)
  y1<- 10 + 0.01*x + e
  m1<-lm(y1~x)
  store_y1_results[i]<-summary(m1)$coefficients[2,4]
  y2<- 10 + 0.01*x + rnorm(n, sd=1.2)
  m2<-lm(y2~x)
  store_y2_results[i]<-summary(m2)$coefficients[2,4]
```

```

}

## power
y1_results<-table(store_y1_results<0.05)
y1_results[2]/sum(y1_results)

## TRUE
## 0.533

y2_results<-table(store_y2_results<0.05)
y2_results[2]/sum(y2_results)

## TRUE
## 0.654

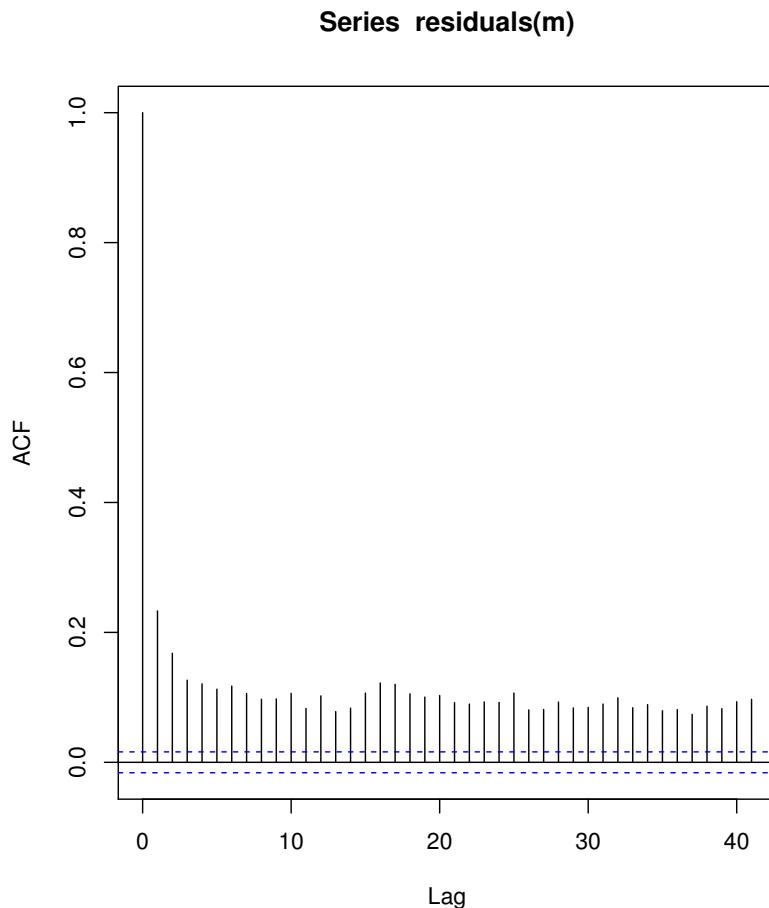
```

The above simulation is just a crude demonstration and can be improved on considerably to reflect reality (exercise).

How to test for normality of residuals? Komogorov-Smirnov and Shapiro-Wilk are formal tests of normality and are only useful for large samples; they not very powerful and not much better than diagnostic plots. These tests may be useful as follow-ups if non-normality is suspected.

Apart from normality, we should also check the independence assumption (the errors are assumed to be independent). Index-plots plot residuals against observation number; note that they are not useful for small samples. An alternative is to compute the correlation between  $e_i, e_{i+1}$  pairs of residuals. The auto-correlation function is not normally used in linear modeling (it's used more in time-series analyses), but can be used to check for this correlation:

```
acf(residuals(m))
```

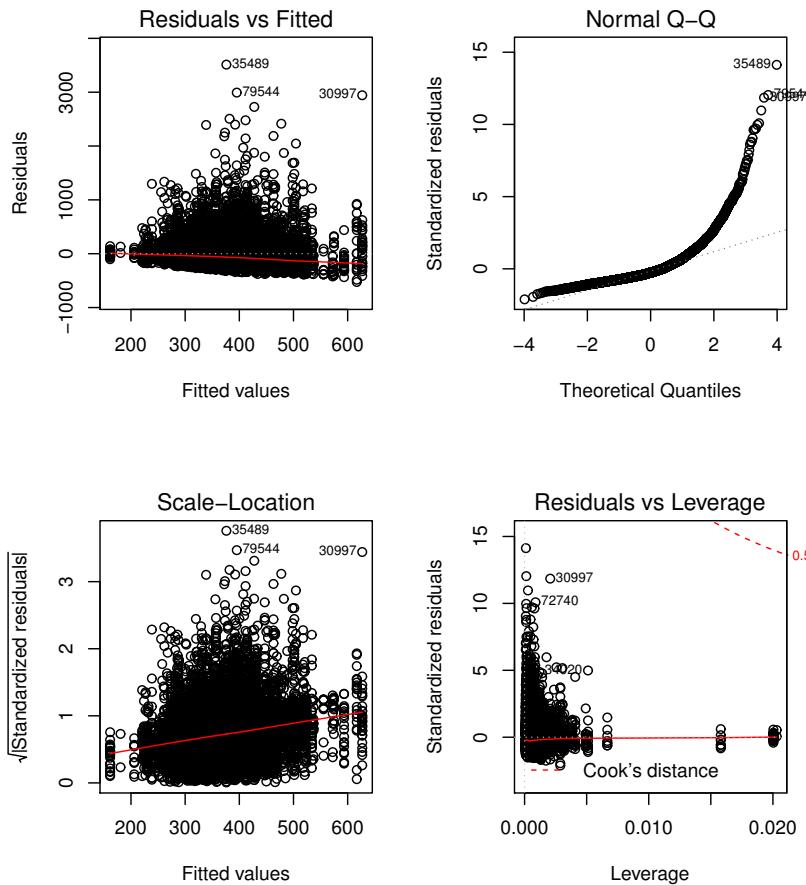


In our model (which is the multiple regression we did in connection with the collinearity issue), we have a serious violation of independence.

Finally, we should check for homoscedasticity (equality of variance). For checking this, plot residuals against fitted values. Fan out suggests violation. A quadratic trend in a plot of residuals against predictor  $x$  could suggest that a quadratic predictor term is needed; note that  $X^T e = 0$ , so we will never have a perfect straight line in such a plot.

R also provides a diagnostics plot, which is generated using the `model fit`:

```
op<-par(mfrow=c(2,2),pty="s")
plot(m)
```



I explain some relevant concepts next.

*Standardized deletion residuals (studres in R)* We can write

$$e = Y - X\hat{\beta} = Y - X(X^T X)^{-1}X^T Y = M\gamma \quad (5.48)$$

where

$$M = I_n - X(X^T X)^{-1}X^T \quad (5.49)$$

M is symmetric, idempotent  $n \times n$ .

Define:

$$\hat{\beta}_{-i} = (X_{-i}^T X_{-i})^{-1} X_{-i}^T Y_{-i} \quad (5.50)$$

where the  $-i$  refers to removing data point  $i$ . Standardized deletion residuals are

$$s_{-i} = \frac{e_i}{\hat{\sigma}_{-i} \sqrt{m_{ii}}} \quad (5.51)$$

where  $m_{ii}$  is the  $i$ -th diagonal element of  $M$ . We can compute  $s_{-i}$  from  $s_i$ :

$$s_{-i} = \frac{s_i \sqrt{n-p-1}}{\sqrt{n-p-s_i^2}} \sim t_{n-p-1} \quad (5.52)$$

If  $n$  is large,  $s_{-i} \approx s_i$ .

*Influence and leverage* (See `lm.influence$hat` in R)

A point can influence the parameter estimates without being an exceptional outlier. Influence does not depend on “outlyingness”. Potential to influence (e.g., by being an extreme  $x$  value) is called leverage; once the  $y$  value is also extreme, we have influence. I.e., it takes an extreme  $x$  and  $y$  value to be influential, and it takes only an extreme  $x$  value to have leverage.

Leverage more formally defined: recall that  $M = I_n - X(X^T X)^{-1} X^T$ . Define a hat matrix  $H = I - M = X(X^T X)^{-1} X^T$ . It's called a hat matrix because it puts a hat on  $y$ :  $\hat{y} = X\hat{\beta} = Hy$ . Since  $x_i^T$  is the  $i$ -th row of  $X$ , we have  $h_{ii} = x_i^T (X^T X)^{-1} x_i$ . The measure for leverage is:

$$h_{ii} = 1 - m_{ii} \quad (5.53)$$

Notice that  $h_{ii}$  is a scalar, so  $\text{trace}(h_{ii}) = h_{ii}$ . So (because for a square matrix  $A, B$ ,  $\text{tr}(AB) = \text{tr}(BA)$ ):

$$h_{ii} = \text{tr}(x_i^T (X^T X)^{-1} x_i) = \text{tr}(x_i^T x_i (X^T X)^{-1}) \quad (5.54)$$

Since  $X^T X = \sum_{i=1}^n x_i x_i^T$ ,  $h_{ii}$  represents the magnitude of  $x_i x_i^T$  relative to the sum of the values for all observations. Note that  $h_{ii}$  only depends on  $X$ .

Also note that

$$\sum_{i=1}^n h_{ii} = \text{tr}(X^T X (X^T X)^{-1}) = \text{tr}(I_p) = p \quad \text{mean}(h_{ii}) = p/n \quad (5.55)$$

$h_{ii}$  measures leverage because  $\text{Var}(e_i) = \sigma^2 m_{ii} = \sigma^2(1 - h_{ii})$  and  $\text{Var}(\hat{y}_i) = \sigma^2 h_{ii}$ . Therefore  $h_{ii}$  has to lie between 0 and 1. When it is close to one, the fitted value will be close to the actual value of  $y_i$ —signalling potential for leverage.

A cutoff one can use to identify high leverage points is  $h_{ii} > 2p/n$  or  $h_{ii} > 3p/n$ .

The leverage of a data point is directly related to how far away it is from the mean:

$$h_{ii} = n^{-1} + \frac{(x_i - \bar{x})^2}{S_{xx}} \quad (5.56)$$

*Cook's distance D: A measure of influence* Let  $s_i$  be the i-th standardized residual,  $\hat{\beta}_{-i}$  the estimate of the vector of parameters with the i-th row removed.

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{-i})^T (X^T X)^{-1} (\hat{\beta} - \hat{\beta}_{-i})}{p\hat{\sigma}^2} = \frac{s_i^2 h_{ii}}{p(1-h_{ii})} \quad (5.57)$$

A data point is influential if it is outlying as well as high leverage. Cutoff for Cook's distance is  $\frac{4}{n}$ .

### 5.1.10 Correcting for multiple testing

This is relevant if you are doing null hypothesis significance testing.

Suppose we are performing  $n$  tests and in each test we specify the probability of making a type I error to be  $\beta$  (note: don't confuse this as type II error). Then, if the tests are independent, the probability of at least one false positive claim in the  $n$  tests is given by

$$1 - (1 - \beta)^n = \alpha \Leftrightarrow \beta = 1 - (1 - \alpha)^{1/n} \quad (5.58)$$

This is called the Šidák correction, and has a stronger bound than the Bonferroni correction and so has greater statistical power.

The Bonferroni just divides  $\beta$  with the number of statistical tests done. So 10 tests would give a corrected  $\alpha$  of  $0.05/10=0.005$ .

### 5.1.11 Transformations: Box-Cox procedure

If the normality assumption is not satisfied, what can we do? One option is to relax the assumption that the errors are normally distributed; we will see this in the Bayesian part of the course.

The more conventional thing to do is to find a transformation of the random variable  $Y$  such that the errors are normally distributed. Let's assume that there exists a transformation  $f_\lambda(Y)$  such that

$$f_\lambda(y_i) = x_i^T \beta + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2) \quad (5.59)$$

The function  $f_\lambda$  is a family of transformations, so for any particular value of  $\lambda$ , we can define a transformation  $z_\lambda = f_\lambda(y)$  on our dependent variable. An example is the log transform. Another example is the reciprocal transform. A third example is the square root transform.

We use maximum likelihood estimation to estimate  $\lambda$ . Note that

$$L(\beta_\lambda, \sigma_\lambda^2, \lambda; y) \propto$$

$$\left(\frac{1}{\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2} \sum [f_\lambda(y_i) - x_i^T \beta]^2\right] \left[\prod_{i=1}^n f'_\lambda(y_i)\right] \quad (5.60)$$

For fixed  $\lambda$ , we first estimate  $\hat{\beta}$  and  $\hat{\sigma}^2$  using the usual MLE methods we learnt. So, we first choose  $\hat{\beta}$  to minimize the residual sum of

squares in the exponent. Call this  $S_\lambda$ . Maximization with respect to  $\sigma^2$  gives  $\hat{\sigma}_\lambda^2 = S_\lambda/n$ .

The Likelihood is going to be proportional to  $\hat{\sigma}$  times the Jacobian:

$$L(\hat{\beta}_\lambda, \hat{\sigma}_\lambda^2, \lambda; y) \propto S_\lambda^{-n/2} \prod f'_\lambda(y_i) \quad (5.61)$$

Next, we will take logs and then maximize with respect to  $\lambda$ :

$$\ell = c - \frac{n}{2} \log S_\lambda + \sum \log f'_\lambda(y_i) \quad (5.62)$$

The above is a general procedure, but an often used family of transformations is the power transformation, proposed in a famous paper by Box and Cox.<sup>8</sup> This family corrects non-normality and/or unequal variance.

If the response is positive, the transformation is

$$f_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log y & \lambda = 0 \end{cases} \quad (5.63)$$

We assume that  $f_\lambda(y) \sim N(x_i^T \beta, \sigma^2)$ . So we have to just estimate  $\lambda$  by MLE, along with  $\beta$ . Here is how to do it by hand:

Since  $f_\lambda = \frac{y^\lambda - 1}{\lambda}$ , it follows that  $f'_\lambda(y) = y^{\lambda-1}$ .

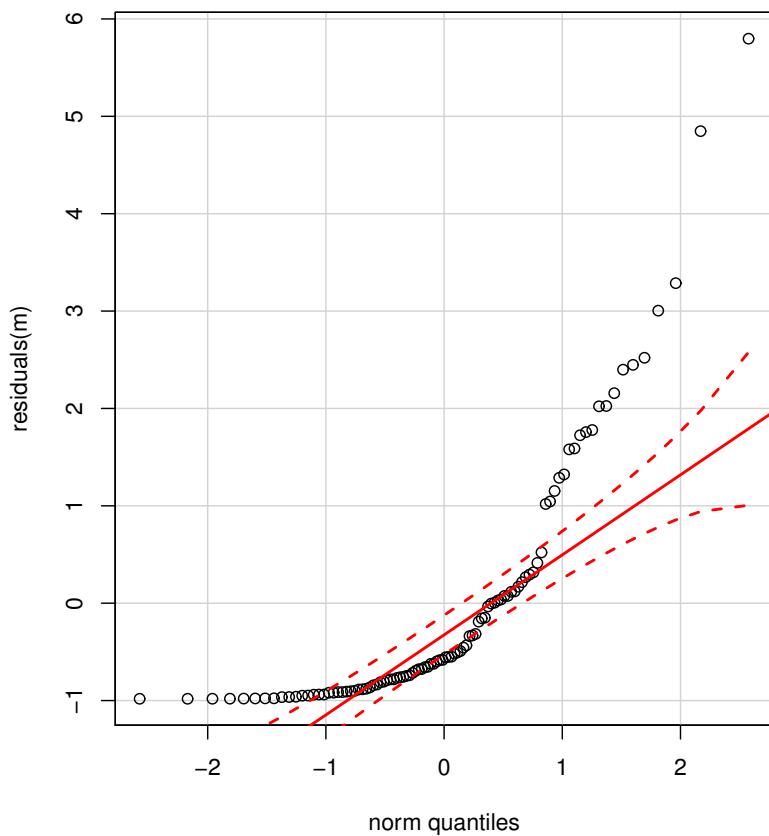
Now, for different  $\lambda$  you can figure out the log likelihoods by hand by solving this equation (remember that this is for a specific data-set and model, i.e., we are given  $y$  and can compute  $S_\lambda$ ):

$$\ell = c - \frac{n}{2} \log S_\lambda + (\lambda - 1) \sum_{\text{Residual sum of squares}} \log(y_i) \quad (5.64)$$

We illustrate this using R. If we have non-normal residuals, we can use the `boxcox` function to determine the relevant transform. Here, the function estimates that  $\lambda = 0$ , hence a log transform is suggested.

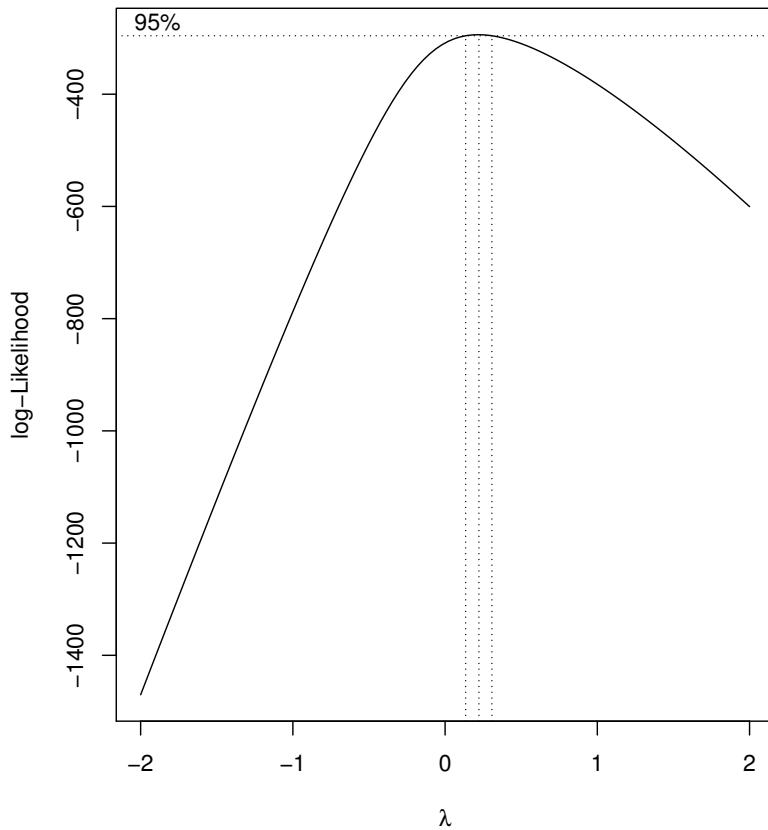
```
## generate some non-normally distributed data:
data<-rchisq(100,df=1)
m<-lm(data~1)
qqPlot(residuals(m))
```

<sup>8</sup> George E.P. Box and David R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252, 1964



```
library(MASS)
## suggests log:
boxcox(m)

m<- lm(log(data)~1)
```



## 5.2 Generalized Linear Models

### 5.2.1 Introduction: Logistic regression

We start with an example data-set that appears in the Dobson et al book: the Beetle dataset.

This data-set shows the number of beetles killed when they were exposed to different doses of some toxic chemical.

```
(beetle<-read.table("datacode/beetle.txt",header=TRUE))

##      dose number killed
## 1 1.6907      59       6
## 2 1.7242      60      13
## 3 1.7552      62      18
## 4 1.7842      56      28
## 5 1.8113      63      52
```

```
## 6 1.8369      59      53
## 7 1.8610      62      61
## 8 1.8839      60      60
```

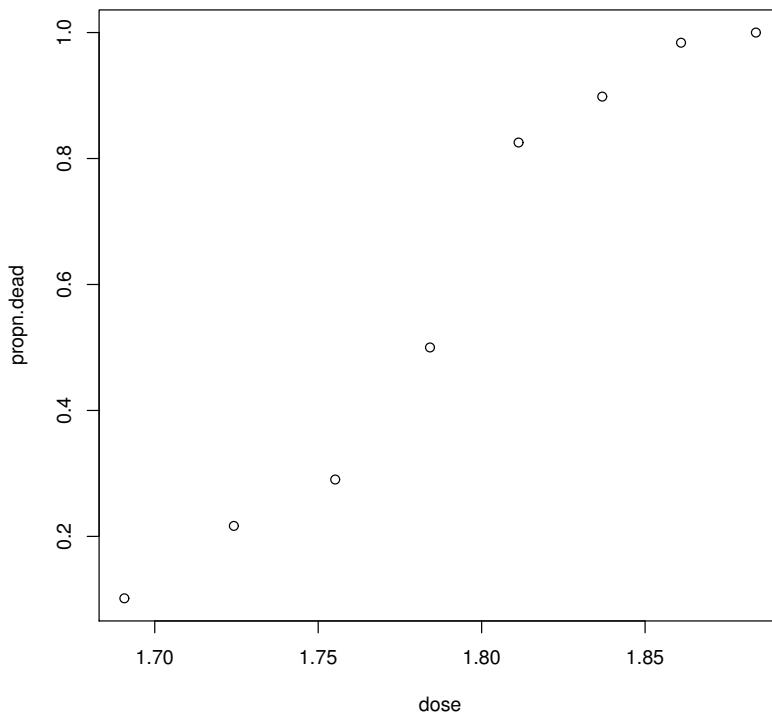
The research question is: does dose affect probability of killing insects? The first thing we probably want to do is calculate the proportions:

```
(beetle$propn.dead<-beetle$killed/beetle$number)

## [1] 0.1016949 0.2166667 0.2903226 0.5000000 0.8253968 0.8983051 0.9838710
## [8] 1.0000000
```

It's also reasonable to just plot the relationship between dose and proportion of deaths.

```
with(beetle,plot(dose,propn.dead))
```



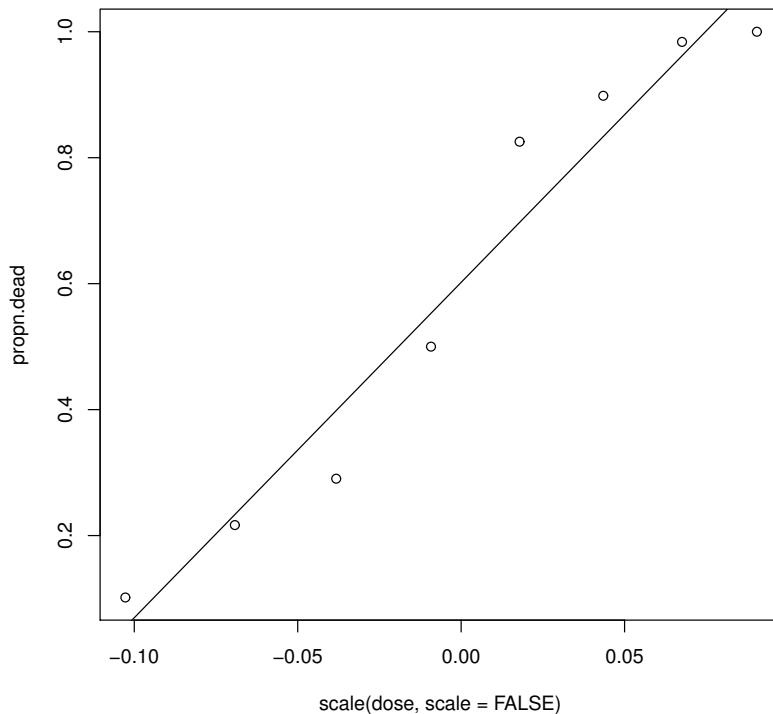
Notice that the y-axis is by definition bounded between 0 and 1.

We could easily fit a linear model to this data-set. We may want to center the predictor, for reasons discussed earlier:

```
fm<-lm(propn.dead~scale(dose,scale=FALSE),beetle)
summary(fm)

##
## Call:
## lm(formula = propn.dead ~ scale(dose, scale = FALSE), data = beetle)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.10816 -0.06063  0.00263  0.05119  0.12818 
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)               0.60203   0.03065 19.64 1.13e-06  
## scale(dose, scale = FALSE) 5.32494   0.48573 10.96 3.42e-05  
##
## Residual standard error: 0.08669 on 6 degrees of freedom
## Multiple R-squared:  0.9524, Adjusted R-squared:  0.9445 
## F-statistic: 120.2 on 1 and 6 DF,  p-value: 3.422e-05
```

```
with(beetle,plot(scale(dose,scale=FALSE),
                 propn.dead))
abline(coef(fm))
```



What's the interpretation of the coefficients?

Clearly the linear model is failing us here. This is the motivation for the generalized linear model.

Instead of using the linear model, we model log odds instead of proportions as a function of dose. Odds are defined as:

$$\frac{p}{1-p} \quad (5.65)$$

and taking the log will give us log odds.

We are going to model log odds (instead of probability) as a linear function of dose.

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 \text{dose} \quad (5.66)$$

The model above is called the logistic regression model.

Once we have estimated the  $\beta$  parameters, we can move back from the log odds space to probability space using simple algebra.

Given a model like

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 \text{dose} \quad (5.67)$$

If we exponentiate each side, we get:

$$\exp \log \frac{p}{1-p} = \frac{p}{1-p} = \exp(\beta_0 + \beta_1 \text{dose}) \quad (5.68)$$

So now we just solve for p, and get (check this):

$$p = \frac{\exp(\beta_0 + \beta_1 \text{dose})}{1 + \exp(\beta_0 + \beta_1 \text{dose})} \quad (5.69)$$

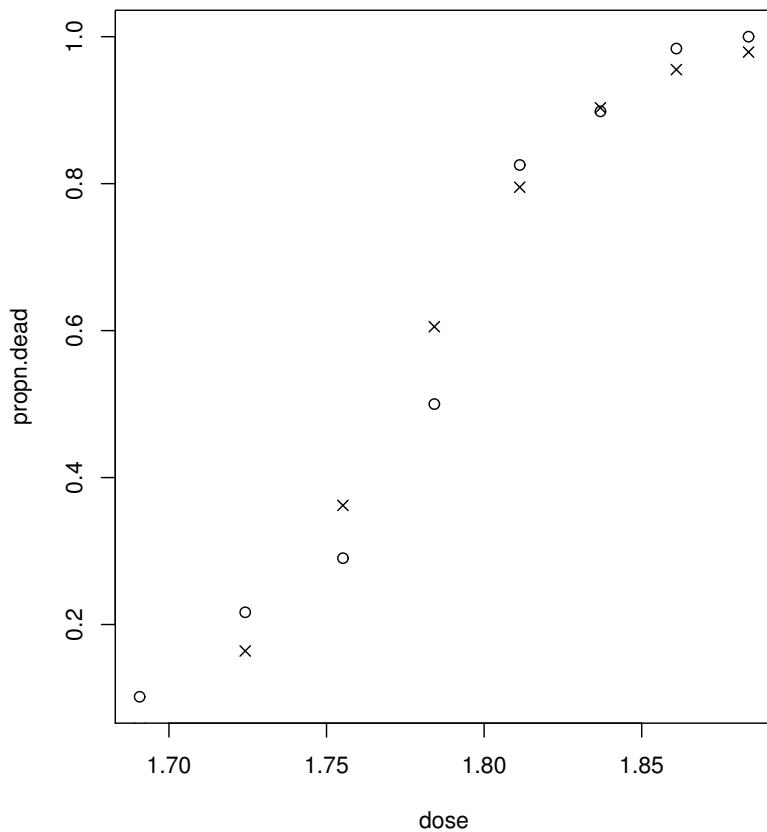
We fit the model in R as follows. Note that as long as I am willing to avoid interpreting the intercept and just interpret the estimate of  $\beta_1$ , there is no need to center the predictor here:

```
fm1<-glm(propn.dead~dose,
           binomial(logit),
           weights=number,
           data=beetle)
summary(fm1)

##
## Call:
## glm(formula = propn.dead ~ dose, family = binomial(logit), data = beetle,
##      weights = number)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.5941 -0.3944  0.8329  1.2592  1.5940
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -60.717     5.181  -11.72  <2e-16
## dose         34.270     2.912   11.77  <2e-16
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 284.202 on 7 degrees of freedom
## Residual deviance: 11.232 on 6 degrees of freedom
## AIC: 41.43
##
## Number of Fisher Scoring iterations: 4
```

We can also plot the observed proportions and the fitted values together; the fit looks pretty good.

```
plot(propn.dead~dose,beetle)
points(fm1$fitted~dose,beetle,pch=4)
```



We can now compute the log odds of death for concentration 1.7552 (for example):

```
## compute log odds of death for
## concentration 1.7552:
x<-as.matrix(c(1, 1.7552))
#log odds:
(log.odds<-t(x)%*%coef(fm1))

##           [,1]
## [1,] -0.5661788
```

We can also obtain the variance-covariance matrix of the fitted coefficients:

```
### compute CI for log odds:
## Get vcov matrix:
(vcovmat<-vcov(fm1))
```

```

##           (Intercept)      dose
## (Intercept) 26.83966 -15.082090
## dose        -15.08209   8.480525

##  $x^T VCOV x$  for dose 1.7552:
(var.log.odds<-t(x)%*%vcovmat%*%x)

##          [,1]
## [1,] 0.0216782

```

And using a normal approximation, based on the asymptotic properties discussed in section 4.2, we can compute the confidence interval for the probability of death given dose 1.7552:

```

##lower
log.odds-1.96*sqrt(var.log.odds)

##          [,1]
## [1,] -0.8547598

##upper
log.odds+1.96*sqrt(var.log.odds)

##          [,1]
## [1,] -0.2775979

```

Note that one should not try to predict outside the range of the design matrix. For example, in the beetle data, the dose ranges from 1.69 to 1.88. We should not try to compute probabilities for dose 2.5, say, since we have no knowledge about whether the relationship remains unchanged beyond the upper bound of our design matrix.

### 5.2.2 Multiple logistic regression: Example from Hindi data

In the Hindi data, we can compute skipping probability, the probability of skipping a word entirely (i.e., never fixating it). We first have to create a vector that has value 1 if the word has 0 ms total reading time, and 0 otherwise.

```

skip<-ifelse(hindi10$TFT==0,1,0)
hindi10$skip<-skip
fm_skip<-glm(skip ~ word_complex+SC,family=binomial(),hindi10)

```

The above example also illustrates the second way to set up the data for logistic (multiple) regression: the dependent variable can simply be a 1,0 value instead of proportions. So, in the beetle data, you could re-code the data to have 1s and 0s instead of proportions. Assuming that

you have recoded the column for status (dead or alive after exposure), the `glm` function call would be:

```
glm(dead~dose,family=binomial(),beetle)
```

Note that logistic regression assumes independence of each data point; this assumption is violated in the Hindi data.

### 5.2.3 Some theory for GLMs

We have considered linear models like

$$E[Y_i] = \mu_i = x_i^T \beta \quad y_i \sim N(\mu_i, \sigma^2) \quad (5.70)$$

GLMs allow us to stay within the linear modeling framework, even if the relationship between response and explanatory variable is not linear.

There is a wider class of distributions beyond the two we have seen (normal, binomial), that are called the **exponential family of distributions**; the normal and binomial fall within this family.

The likelihood function of the exponential family's distributions can be written in very general terms as follows:

$$f(y; \theta_i, \phi) = \exp \left[ \frac{y\theta_i - b(\theta_i)}{\phi/w} + c(y, \phi) \right] \quad (5.71)$$

*Example: The normal distribution* Consider the normal distribution. We can write it in the general form of equation 5.71.

$$\begin{aligned} f(y) &= \frac{1}{\sigma\sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \frac{(y-\mu)}{\sigma} \right)^2 \right] \\ &= \exp \left[ \log \frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2} \left( \frac{(y-\mu)}{\sigma} \right)^2 \right] \quad (5.72) \\ &= \exp \left[ -\frac{1}{2} \left( \frac{y^2 + \mu^2 - 2y\mu}{\sigma^2} \right) - \log \sigma\sqrt{2\pi} \right] \end{aligned}$$

A little bit of algebraic manipulation (exercise) will now give us:

$$\begin{aligned} &= \exp \left[ \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{y^2}{2\sigma^2} + \frac{\log \sigma\sqrt{2\pi}}{2} \right] \\ &= \exp \left[ \frac{y\mu - \mu^2/2}{\sigma^2} + c(y, \phi) \right] \quad \text{i.e., } c(y, \phi) = -\frac{y^2}{2\sigma^2} + \frac{\log \sigma\sqrt{2\pi}}{2} \\ &= \exp \left[ \frac{y\theta - b(\theta)}{\phi/w} + c(y, \phi) \right] \quad (5.73) \end{aligned}$$

Here,  $\theta = \mu$ ,  $\phi = \sigma^2$ ,  $w = 1$ , and we have  $b(\theta) = \mu^2/2$ ,  $c(y, \phi) = -\frac{y^2}{2\sigma^2} + \frac{\log \sigma \sqrt{2\pi}}{2}$ .

This general formulation gives us two useful results (not proved here, because that would take us too far afield; but I will try to add the full proof later if there is interest):

1. The first derivative of  $b(\theta) = \frac{\mu^2}{2}$ , is  $b'(\theta) = \mu$ . This is a general result for the exponential family:

$$E[y] = b'(\theta) = \mu$$

2. The variance of Y is  $Var(Y) = \frac{\phi}{w} b''(\theta)$ . So, here, we'd get

$$Var(Y) = \frac{\sigma^2}{1} 1 = \sigma^2$$

*Example 2: Binomial distribution* Let's look at another example of how we can write an exponential family distribution in this general form. Consider the binomial distribution, which we will start by writing as below. Here, n is the total number of trials, and y is the proportion of successes. For example,  $n=10$ ,  $y=7/10$ , gives us 7 successes out of 10. This is just another way to parameterize the binomial distribution, although it is not one that you have seen before.

$$ny \sim \text{Binomial} \left( n, \frac{\exp(\theta)}{1 + \exp(\theta)} \right) \quad \text{i.e., } p = \frac{\exp(\theta)}{1 + \exp(\theta)} \quad (5.74)$$

$$\begin{aligned} f(ny; \theta, \phi) &= \binom{n}{ny} p^{ny} (1-p)^{n-ny} \\ &= \exp \left[ \log \binom{n}{ny} + ny \log p + (n-ny) \log(1-p) \right] \\ &= \exp \left[ ny \log \frac{p}{1-p} + n \log(1-p) + c(y, \phi) \right] \quad \text{i.e., } c(y, \phi) = \log \binom{n}{ny} \end{aligned} \quad (5.75)$$

Since  $p = \frac{\exp(\theta)}{1+\exp(\theta)}$ , we can write

$$n \log(1-p) = n \log \frac{1}{1 + \exp(\theta)} = -n \log(1 + \exp(\theta)) \quad (5.76)$$

Also, let  $\theta = \log \frac{p}{1-p}$ .

Then, we can continue as follows:

$$\begin{aligned} f(ny; \theta, \phi) &= \exp \left[ ny \log \frac{p}{1-p} + n \log(1-p) + c(y, \phi) \right] \quad \text{i.e., } c(y, \phi) = \log \binom{n}{ny} \\ &= \exp [ny\theta - n \log(1 + \exp(\theta)) + c(y, \phi)] \\ &= \exp \left[ \frac{y\theta - b(\theta)}{\phi/n} + c(y, \phi) \right] \quad \text{i.e., } b(\theta) = n \log(1 + \exp(\theta)) \end{aligned} \quad (5.77)$$

### 5.2.4 The canonical link

For each data point  $Y_i$  from a distribution that's a member of the exponential family, the general form of the likelihood function is:

$$f(y; \theta_i, \phi) = \exp \left[ \frac{y\theta_i - b(\theta_i)}{\phi/w} + c(y, \phi) \right] \quad (5.78)$$

where  $E[Y_i] = \mu_i = h(x_i^T \beta)$ . Since we know that  $E[Y_i] = b'(\theta_i)$ , we can write

$$E(Y_i) = \mu_i = h(x_i^T \beta) = b'(\theta) \quad (5.79)$$

Now, if we want to get  $x_i^T \beta$ , we just take the inverse of the function  $h(\cdot)$ , call it  $g(\cdot)$ . This gives us something called the canonical link function:

$$x_i^T \beta = h^{-1}(b'(\theta)) = \underset{\text{canonical link}}{\overset{\uparrow}{gb'(\theta)}} \quad (5.80)$$

For different distributions in the exponential family, the canonical link functions are as follows:

Distribution	$h(x_i^T \beta) = \mu_i$	$g(\mu_i) = \theta_i$
Binomial logit link	$\frac{\exp[\theta_i]}{1+\exp[\theta_i]}$	$\log \frac{y}{1-y}$
Normal identity	$\theta$	$g = h$
Poisson log	$\exp[\theta]$	$\log[\mu]$
Gamma inverse	$-\frac{1}{\theta}$	$-\frac{1}{\mu_i}$
Cloglog cloglog	$1 - \exp[-\exp[\theta_i]]$	$\log(-\log(1 - \mu_i))$
Probit probit	$\Phi(\theta)$	$\Phi^{-1}(\theta)$ (qnorm)

The big thing about the canonical link is that it expresses  $\theta_i$  as a linear combination of the parameters:  $x_i^T \beta$ . You can decide which link to use by plotting  $g(\mu_i)$  against the predictor (in case we have only a single predictor x).

### 5.2.5 Estimation of parameters

In linear models, we know how to estimate the  $\beta$  parameters:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (5.81)$$

and the covariance matrix is  $\sigma^2(X^T X)^{-1}$ .

For reasons we won't get into in this course, in GLMs we use **iteratively reweighted least squares**. Here is how it works:

1. Specify an **initial vector of parameters**:  $b^{(m)} = (\beta_0, \dots, \beta_p)^T$ , where initially  $m = 1$ :

```
## eta=xbeta:
eta.i<- -60+35*beetle$dose
```

2. Specify a **weight matrix W** that depends on current parameter estimates:

If we define:

$$w_{ii} = \frac{n_i \exp[\eta_i]}{(1 + \exp[\eta_i])^2} \quad (5.82)$$

we can compute W:

```
n.i <- beetle$number
w.ii.fn<-function(n.i,eta.i){
  (n.i*exp(eta.i))/(1+exp(eta.i))^2
}
w.iis<-w.ii.fn(n.i,eta.i)
##weights matrix:
W<-diag(as.vector(w.iis))
```

3. Specify a **vector z** that depends on the current parameter estimates and response values:

$$z_i = \eta_i + \frac{y_i - \mu_i}{\mu_i(1 - \mu_i)} \quad \mu_i = \frac{\exp[\eta_i]}{1 + \exp[\eta_i]} \quad (5.83)$$

```
mu.i<-exp(eta.i)/(1+exp(eta.i))
z.i<-eta.i + ((beetle$propn.dead-mu.i))/(
  (mu.i*(1-mu.i)))
```

4. Compute new estimate of parameters:  $b^{(m+1)} = (X^T W X)^{-1} X^T W z$ :

```
##The design matrix:
col1<-c(rep(1,8))
X<-as.matrix(cbind(col1,beetle$dose))
## update coeffs:
eta.i<-solve(t(X)%%W%%X)%%
  t(X)%%W%%z.i
```

Repeat with updated coefficients; stop at convergence.

If you implement this approach (exercise), you will find that it takes 7 iterations to get convergence. R can do it in four iterations because it uses a different approach.

### 5.2.6 Deviance

We saw encountered deviance earlier (page 71) in connection with ANOVA.

The deviance is more generally defined as

$$D = 2[\ell(b_{max}; y) - \ell(b; y)] \quad (5.84)$$

where  $\ell(b_{max}; y)$  is the log likelihood of the saturated model (the model with the maximal number of parameters that can be fit), and  $\ell(b; y)$  is the log likelihood of the model with the parameters b. As we saw earlier, D has a chi-squared distribution.

*Deviance for the normal distribution* The deviance is

$$D = \frac{1}{\sigma^2} \sum (y_i - \bar{y})^2$$

[See p. 80 onwards in the Dobson et al book for proofs and more detail.]

*Deviance for the binomial distribution* Deviance is defined as  $D = \sum d_i$ , where:

$$d_i = -2 \times n_i [y_i \log(\frac{\hat{\mu}_i}{y_i}) + (1 - y_i) \log(\frac{1 - \hat{\mu}_i}{1 - y_i})] \quad (5.85)$$

The basic idea here is that if the model fit is good, Deviance will have a  $\chi^2$  distribution with  $N - p$  degrees of freedom. So that is what we will use for assessing model fit.

We will also use deviance for hypothesis testing. The difference in deviance (residual deviance) between two models also has a  $\chi^2$  distribution (this should remind you of ANOVA), with dfs being  $p - q$ , where  $q$  is the number of parameters in the first model, and  $p$  the number of parameters in the second.

I discuss hypothesis testing first, then evaluating goodness of fit using deviance.

### 5.2.7 Hypothesis testing: Residual deviance

Returning to our beetle data, let's say we fit our model:

```
glm1<-glm(propn.dead~dose,binomial(logit),
            weights=number,data=beetle)
```

The summary output shows us the number of iterations that led to the parameter estimates:

```
summary(glm1)

##
## Call:
## glm(formula = propn.dead ~ dose, family = binomial(logit), data = beetle,
##       weights = number)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q        Max
## -1.5941   -0.3944    0.8329   1.2592   1.5940
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -60.717     5.181  -11.72  <2e-16
## dose         34.270     2.912   11.77  <2e-16
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 284.202 on 7 degrees of freedom
## Residual deviance: 11.232 on 6 degrees of freedom
## AIC: 41.43
##
## Number of Fisher Scoring iterations: 4
```

But we also see something called **Null deviance** and **Residual deviance**. These are used to evaluate quality of model fit. Recall that we can compute the fitted values and compare them to the observed values:

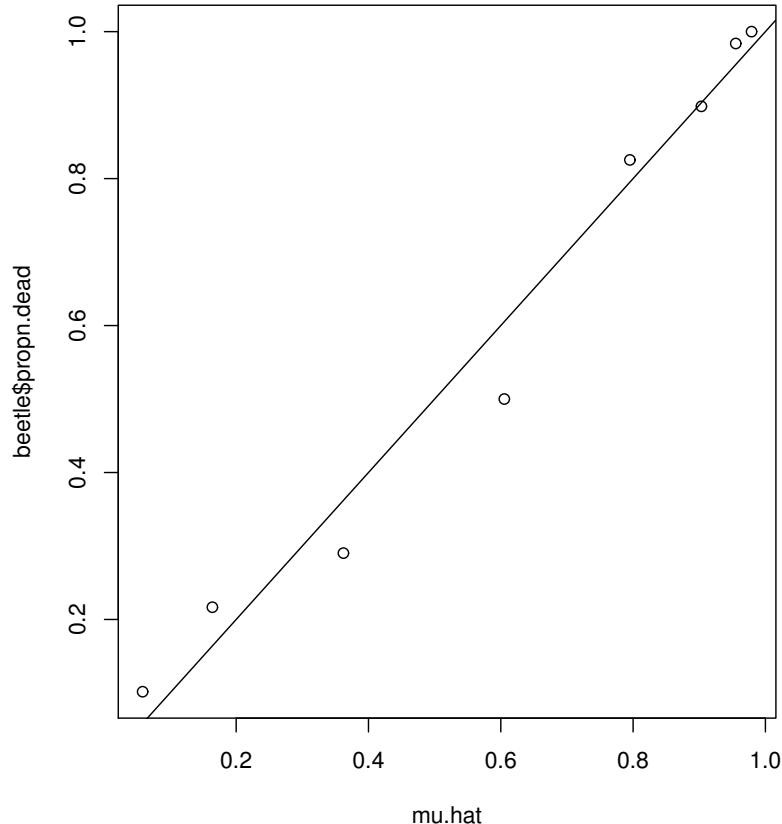
```
# beta.hat is (-60.71745 , 34.27033)
(eta.hat<- -60.71745 + 34.27033*beetle$dose)

## [1] -2.7766031 -1.6285470 -0.5661668  0.4276728  1.3563987  2.2337192
## [7]  3.0596341  3.8444247

(mu.hat<-exp(eta.hat)/(1+exp(eta.hat)))

## [1] 0.05860168 0.16402950 0.36212179 0.60531781 0.79517377 0.90323690
## [7] 0.95519664 0.97904960
```

```
# compare mu.hat with observed proportions
plot(mu.hat,beetle$propn.dead)
abline(0,1)
```

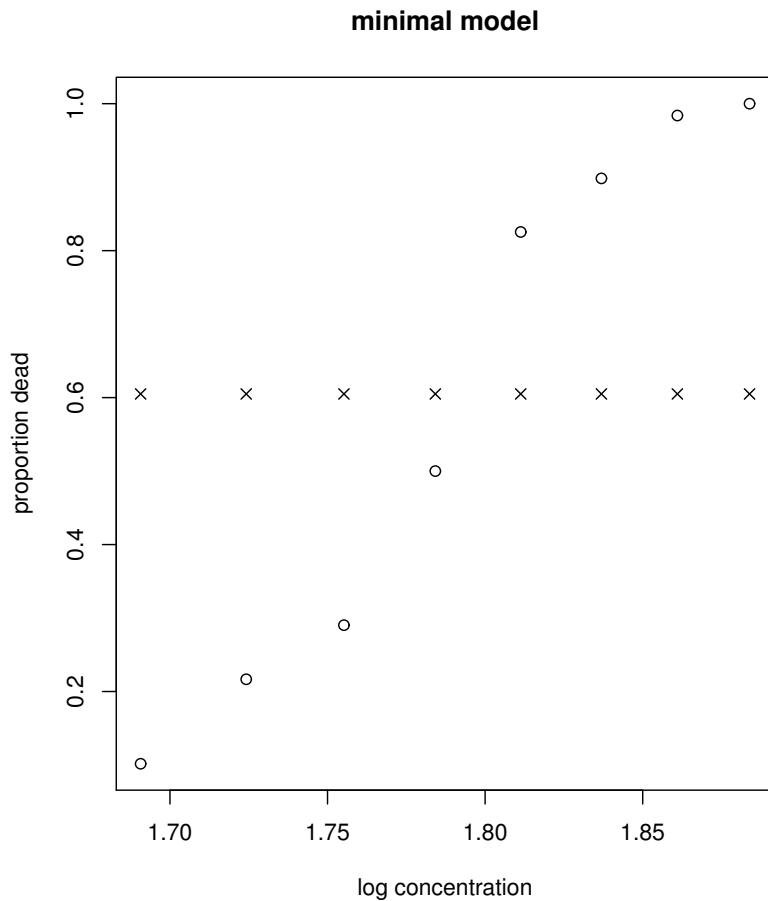


To evaluate whether dose has an effect, we will do something analogous to the model comparison methods we saw earlier. First, fit a model with only an intercept. Notice that the null deviance is 284 on 7 degrees of freedom.

```
null.glm<-glm(propn.dead~1,binomial(logit),
                 weights=number,data=beetle)
summary(null.glm)

##
## Call:
## glm(formula = propn.dead ~ 1, family = binomial(logit), data = beetle,
##      weights = number)
##
```

```
## Deviance Residuals:  
##      Min       1Q   Median       3Q      Max  
## -8.105  -5.294   1.099   5.615   7.766  
##  
## Coefficients:  
##              Estimate Std. Error z value Pr(>|z|)  
## (Intercept)  0.42630    0.09327  4.571 4.87e-06  
##  
## (Dispersion parameter for binomial family taken to be 1)  
##  
## Null deviance: 284.2 on 7 degrees of freedom  
## Residual deviance: 284.2 on 7 degrees of freedom  
## AIC: 312.4  
##  
## Number of Fisher Scoring iterations: 4  
  
plot(beetle$dose,beetle$propn.dead,xlab="log concentration",  
     ylab="proportion dead",main="minimal model")  
points(beetle$dose,null.glm$fitted,pch=4)
```



Add a term for dose. Now, the residual deviance is 11.2 on 6 dfs/

```
dose.glm<-glm(propn.dead~dose,binomial(logit),
  weights=number,data=beetle)
summary(dose.glm)

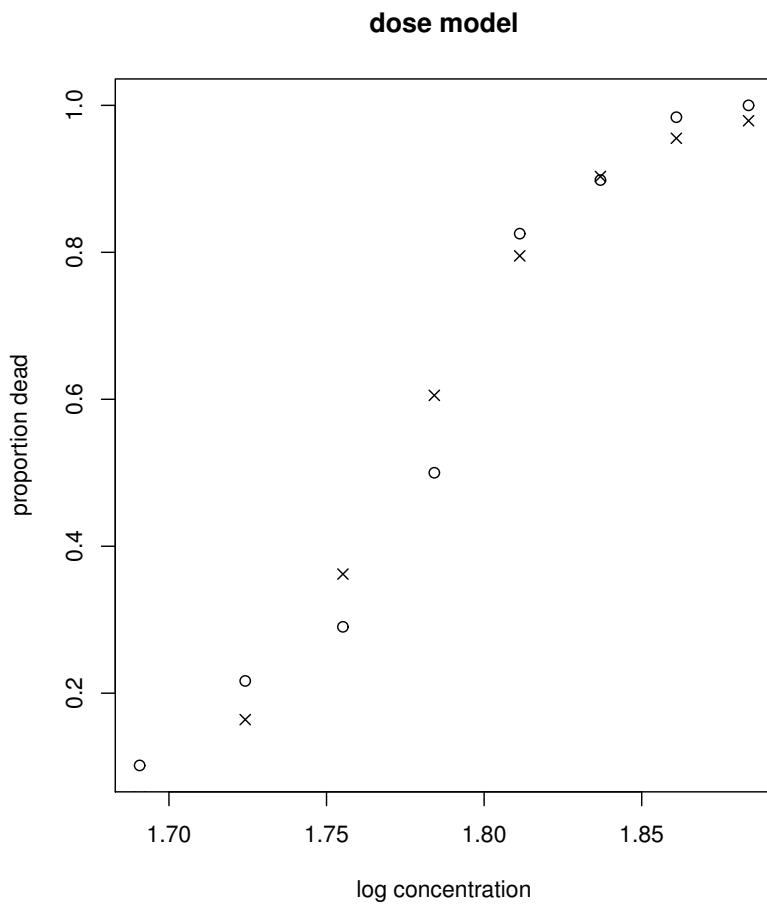
##
## Call:
## glm(formula = propn.dead ~ dose, family = binomial(logit), data = beetle,
##       weights = number)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.5941   -0.3944    0.8329    1.2592    1.5940
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -60.717     5.181  -11.72   <2e-16
```

```

## dose      34.270      2.912    11.77   <2e-16
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 284.202 on 7 degrees of freedom
## Residual deviance: 11.232 on 6 degrees of freedom
## AIC: 41.43
##
## Number of Fisher Scoring iterations: 4

plot(beetle$dose,beetle$propn.dead,xlab="log concentration",
      ylab="proportion dead",main="dose model")
points(beetle$dose,dose.glm$fitted,pch=4)

```



The change in deviance from the null model is  $284.2 - 11.2 = 273$  on 1 df. Since the critical  $\chi^2_1 = 3.84$ , we reject the null hypothesis that  $\beta_1 = 0$ .

You can do the model comparison using the anova function. Note

that no statistical test is calculated; you need to do that yourself.

```
anova(null.glm,dose.glm)

## Analysis of Deviance Table
##
## Model 1: propn.dead ~ 1
## Model 2: propn.dead ~ dose
##   Resid. Df Resid. Dev Df Deviance
## 1       7    284.202
## 2       6    11.232  1    272.97
```

Actually, you don't even need to define the null model; the `anova` function automatically compares the fitted model to the null model:

```
anova(dose.glm)

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: propn.dead
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev
## NULL             7    284.202
## dose  1    272.97       6    11.232
```

### 5.2.8 Assessing goodness of fit of a fitted model

The deviance for a given degrees of freedom  $v$  should have a  $\chi_v^2$  distribution for the model to be adequate. As an example, consider the null model above. The deviance is clearly much larger than the 95th percentile cutoff point of the chi-squared distribution with 7 dfs, so the model is not adequate.

```
deviance(null.glm)

## [1] 284.2024

## critical value:
qchisq(0.95,df=7)

## [1] 14.06714
```

Now consider the model with dose as predictor. The deviance is less than the 95th percentile, so the fit is adequate.

```
deviance(dose.glm)
## [1] 11.23223
qchisq(0.95,df=6)
## [1] 12.59159
```

### 5.2.9 Residuals in GLMs

In the binomial distribution, Deviance  $D = \sum d_i$ , where:

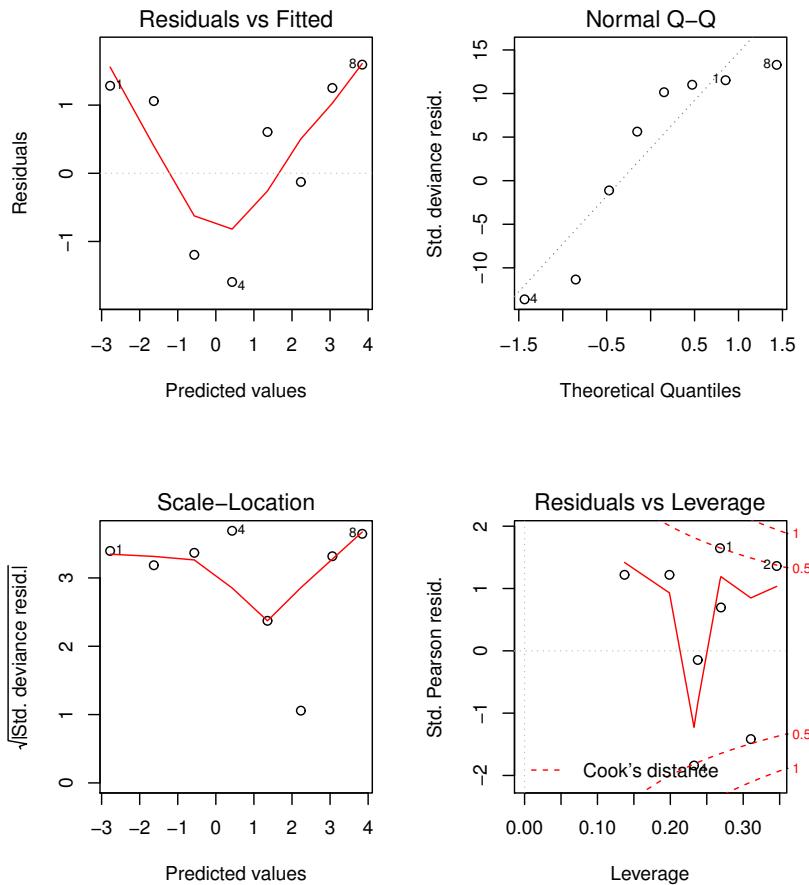
$$d_i = -2 \times n_i [y_i \log\left(\frac{\hat{\mu}_i}{y_i}\right) + (1 - y_i) \log\left(\frac{1 - \hat{\mu}_i}{1 - y_i}\right)] \quad (5.86)$$

The  $i$ -th deviance residual is defined as:

$$e_{D,i} = \text{sgn}(y_i - \hat{\mu}_i) \times \sqrt{d_i} \quad (5.87)$$

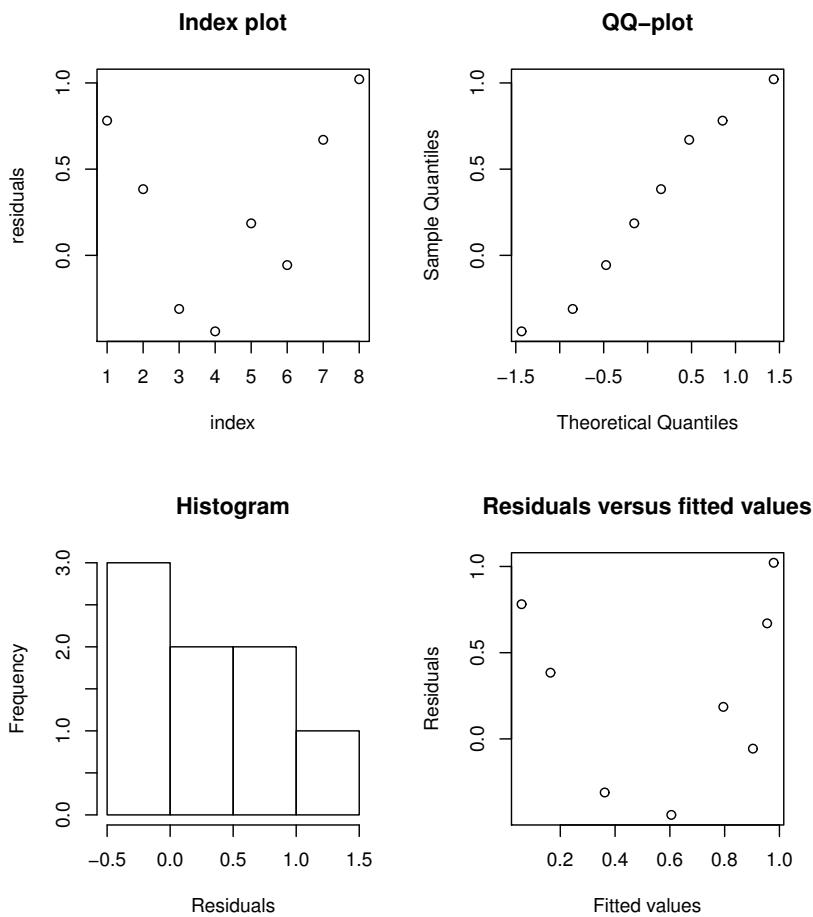
These can be used to check for model adequacy as discussed earlier in the context of linear models. One can just use the plot function inspect the residuals:

```
op<-par(mfrow=c(2,2),pty="s")
plot(dose.glm)
```



Alternatively, one can do this by hand:

```
op<- par(mfrow=c(2,2),pty="s")
plot(dose.glm$resid,
      xlab="index",ylab="residuals",main="Index plot")
qqnorm(dose.glm$resid,main="QQ-plot")
hist(dose.glm$resid,xlab="Residuals",main="Histogram")
plot(dose.glm$fit,dose.glm$resid,xlab="Fitted values",
      ylab="Residuals",
      main="Residuals versus fitted values")
```



# 6

## *Hierarchical linear models: Frequentist and Bayesian approaches*

### *6.1 Linear mixed models*

In linear modeling, we model the mean of a response  $Y_1, \dots, Y_n$  as a function of a vector of predictors  $x_1, \dots, x_n$ . We assume that the  $Y_i$  are conditionally independent given  $\mathbf{x}, \mathbf{f}_i$ . When  $Y$ 's are not marginally independent, we have  $\text{Cor}(Y_1, Y_2) \neq 0$ , or  $P(Y_2 | Y_1) \neq P(Y_2)$ .

Linear mixed models are useful for correlated data where  $\mathbf{Y} | \mathbf{X}, \mathbf{f}_i$  are not independently distributed.

#### *6.1.1 Informal presentation of LMMs*

Consider the following fake data set, taken from a textbook (Maxwell and Delaney, p. 497):

```
noisedeg<-read.table("datacode/noisedeg.txt")
```

This is a  $2 \times 2$  factorial design, where each subject sees a stimulus in a no noise and noise condition; the stimulus is either angled at 0 degrees or 8 degrees. The dependent variable is reaction time (in msec).

One important point here, which was also true in the Hindi data, is that different subjects have different effects of noise and deg. In the linear models we fit earlier we ignored this.

```
## returning to our noise data (noisedeg):
## here's an important fact about our data:
# different subjects have different means for no.noise and noise
# and different means for the three levels of deg

t(means.noise<-with(noisedeg,tapply(rt,list(subj,noise),mean)))
```

```

##          s1 s10 s2 s3 s4 s5 s6 s7 s8 s9
## no.noise 420 450 450 480 480 600 390 480 540 570
## noise    540 600 420 720 630 570 420 630 630 600

t(means.deg<-with(noisedeg,tapply(rt,list(subj,deg),mean)))

##      s1 s10 s2 s3 s4 s5 s6 s7 s8 s9
## 0 450 510 390 570 450 510 360 510 510 510
## 4 510 540 480 630 660 660 450 600 660 660

```

We can view the differential behavior of subjects in a graph (Figures 6.1 and 6.2).

Given these differences between subjects, you could fit a separate linear model for each subject, collect together the intercepts and slopes for each subject, and then check if the intercepts and slopes are significantly different from zero.

Try this for one subject (`s1`):

```

## fit a separate linear model for subject s1:
s1data<-subset(noisedeg,subj=="s1")
lm(rt~noise,s1data)

##
## Call:
## lm(formula = rt ~ noise, data = s1data)
##
## Coefficients:
## (Intercept)  noisenoise
##           420           120

```

Go back and look at the means for `s1` for noise and compare them to the coefficients above. Now we can do this for every one of our 10 subjects. I don't print this result out because it's consume a lot of pages.

```

## do the same for each subject using a for-loop
subjects<-paste("s",rep(1:10),sep="")
for(i in subjects){
  sdata<-subset(noisedeg,subj==i)
  lm(rt~noise,sdata)
}

```

There is a function in the package `lme4` that does the above for you: `lmList`.

```

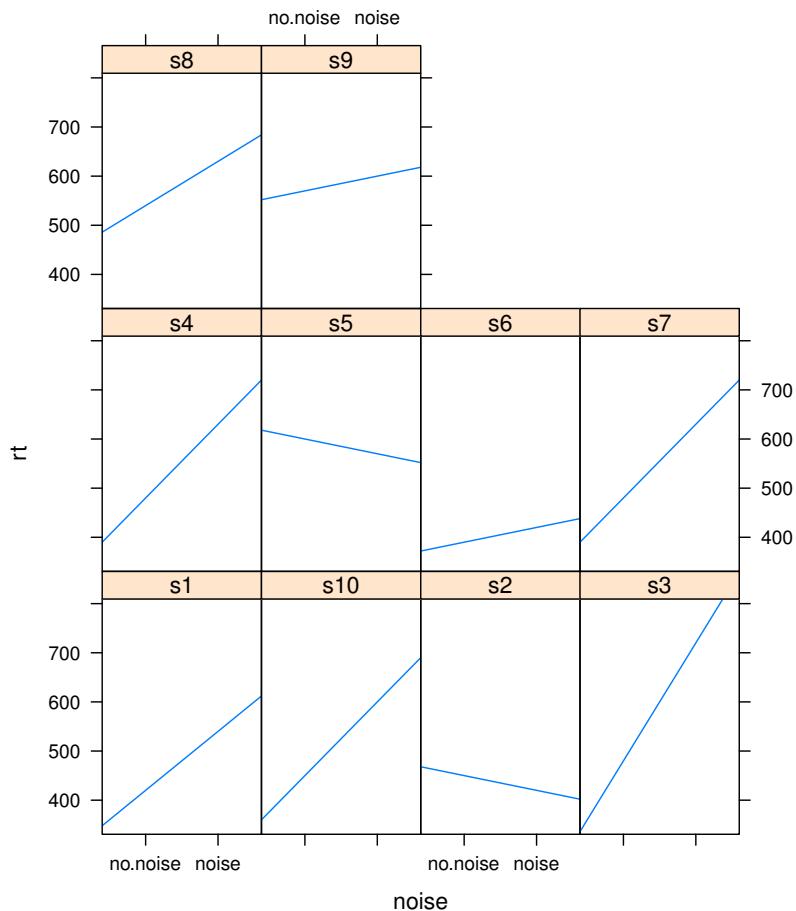
## We can visualize these differences graphically:

library(lattice)

## noise by subject (data points):
print(xyplot(rt~noise|subj,
    panel=function(x,y,...){panel.xyplot(x,y,type="r")},noisedeg))

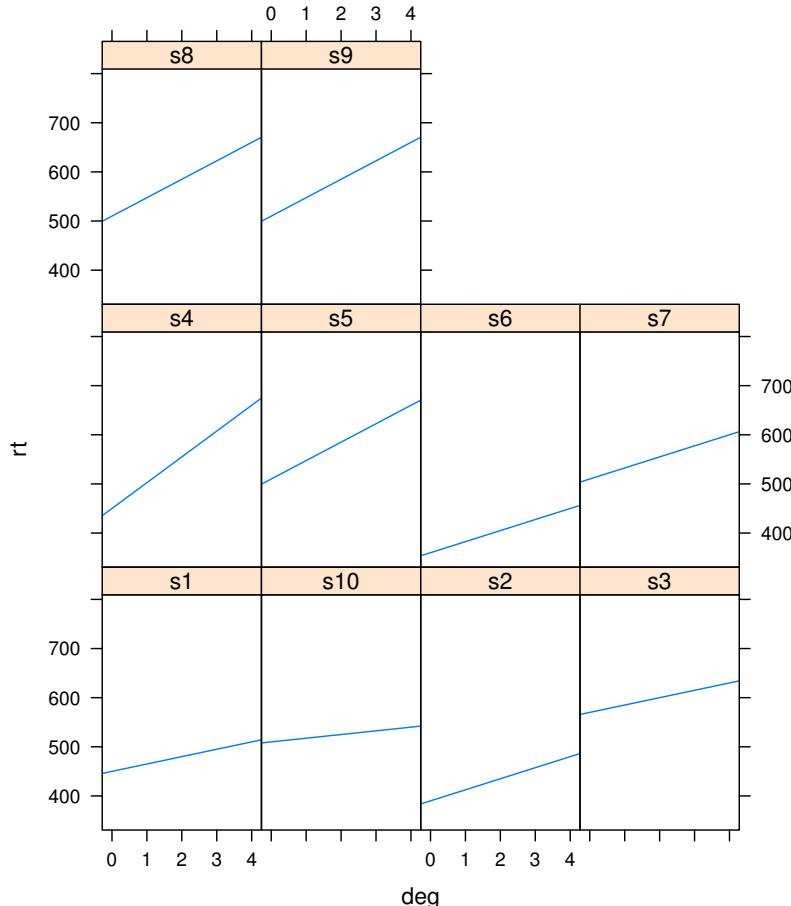
```

Figure 6.1: Noise effects by subject.



```
## same as above, but for deg:
print(xyplot(rt~deg|subj,
  panel=function(x,y,...){panel.xyplot(x,y,type="r")},noisedeg))
```

Figure 6.2: Degree effects by subject.



```
library(lme4)
lmlist.fml<-lmList(rt~noise|subj,noisedeg)
print(lmlist.fml$s1)

##
## Call:
## (function (formula, data, subset, weights, na.action, method = "qr",
##   model = TRUE, x = FALSE, y = FALSE, qr = TRUE, singular.ok = TRUE,
##   contrasts = NULL, offset, ... )
## {
##   ret.x <- x
##   ret.y <- y
```

```

##   cl <- match.call()
##   mf <- match.call(expand.dots = FALSE)
##   m <- match(c("formula", "data", "subset", "weights", "na.action",
##              "offset"), names(mf), 0L)
##   mf <- mf[c(1L, m)]
##   mf$drop.unused.levels <- TRUE
##   mf[[1L]] <- quote(stats::model.frame)
##   mf <- eval(mf, parent.frame())
##   if (method == "model.frame")
##       return(mf)
##   else if (method != "qr")
##       warning(gettextf("method = '%s' is not supported. Using 'qr'", 
##                      method), domain = NA)
##   mt <- attr(mf, "terms")
##   y <- model.response(mf, "numeric")
##   w <- as.vector(model.weights(mf))
##   if (!is.null(w) && !is.numeric(w))
##       stop("'weights' must be a numeric vector")
##   offset <- as.vector(model.offset(mf))
##   if (!is.null(offset)) {
##       if (length(offset) != NROW(y))
##           stop(gettextf("number of offsets is %d, should equal %d (number of observations)", 
##                         length(offset), NROW(y)), domain = NA)
##   }
##   if (is.empty.model(mt)) {
##       x <- NULL
##       z <- list(coefficients = if (is.matrix(y)) matrix(, 0,
##                                              3) else numeric(), residuals = y, fitted.values = 0 * 
##                                              y, weights = w, rank = 0L, df.residual = if (!is.null(w)) sum(w != 
##                                              0) else if (is.matrix(y)) nrow(y) else length(y))
##       if (!is.null(offset)) {
##           z$fitted.values <- offset
##           z$residuals <- y - offset
##       }
##   }
##   else {
##       x <- model.matrix(mt, mf, contrasts)
##       z <- if (is.null(w))
##             lm.fit(x, y, offset = offset, singular.ok = singular.ok,
##                     ...)
##             else lm.wfit(x, y, w, offset = offset, singular.ok = singular.ok,
##                          ...)
##   }

```

```

##      class(z) <- c(if (is.matrix(y)) "m1m", "lm")
##      z$na.action <- attr(mf, "na.action")
##      z$offset <- offset
##      z$contrasts <- attr(x, "contrasts")
##      z$xlevels <- .getXlevels(mt, mf)
##      z$call <- cl
##      z$terms <- mt
##      if (model)
##          z$model <- mf
##      if (ret.x)
##          z$x <- x
##      if (ret.y)
##          z$y <- y
##      if (!qr)
##          z$qqr <- NULL
##      z
## })(formula = rt ~ noise, data = structure(list(rt = c(420L, 420L,
## 480L, 600L), subj = structure(c(1L, 1L, 1L, 1L), .Label = c("s1",
## "s10", "s2", "s3", "s4", "s5", "s6", "s7", "s8", "s9"), class = "factor"),
## deg = c(0L, 4L, 0L, 4L), noise = structure(c(1L, 1L, 2L,
## 2L), .Label = c("no.noise", "noise"), class = "factor")), .Names = c("rt",
## "subj", "deg", "noise"), terms = rt ~ subj + deg + noise, row.names = c("1",
## "11", "31", "41"), class = "data.frame"), weights = NULL, offset = NULL)
##
## Coefficients:
## (Intercept) noisenoise
##           420          120

```

One can plot the individual lines for each subject, as well as the linear model mo's line (this shows how each subject deviates in intercept and slope from the model mo's intercept and slopes). See Figure 6.3.

To find out if there is an effect of noise, you can simply check whether the slopes of the individual subjects' fitted lines taken together are significantly different from zero:

```

t.test(coef(lmlist.fm1)[2])

##
## One Sample t-test
##
## data: coef(lmlist.fm1)[2]
## t = 3.2225, df = 9, p-value = 0.01045
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## 26.82139 153.17861

```

```

plot(as.numeric(noisedeg$noise)-1,
     noisedeg$rt,axes=F,
     xlab="noise",ylab="rt")
axis(1,at=c(0,1),
     labels=c("no.noise","noise"))
axis(2)

subjects<-paste("s",1:10,sep="")

for(i in subjects){
  abline(lmlist.fml1[[i]])
}

abline(lm(rt~noise,noisedeg),lwd=3,col="red")

```

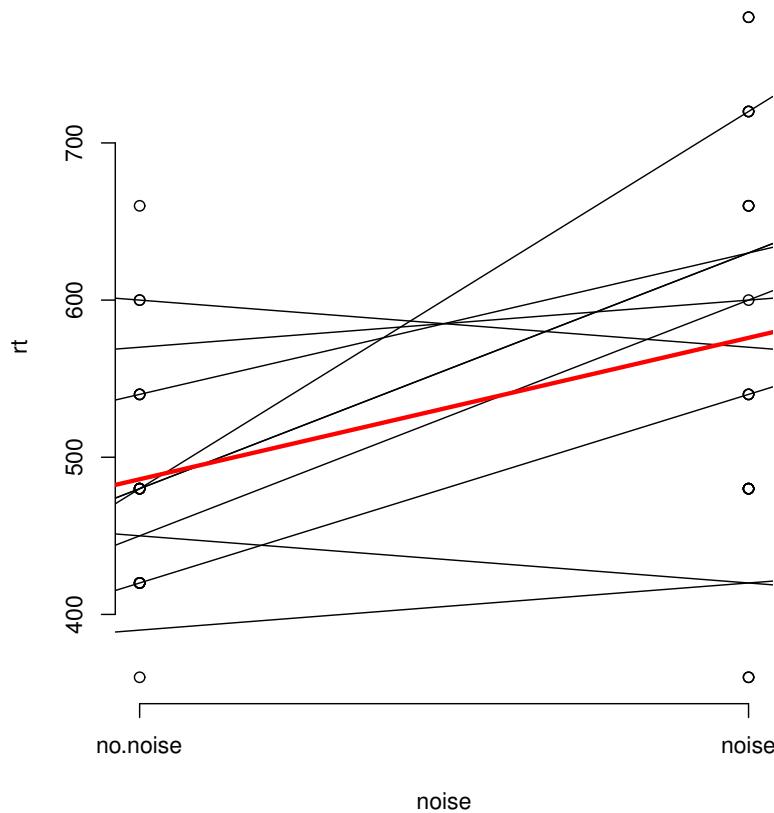


Figure 6.3: Each subject's intercept and slope line varies about the averager over all subjects (the intercept and slope from the lm function).

```
## sample estimates:
## mean of x
##      90
```

The above is called repeated measures regression (see ?? for details). We now transition to the next stage: the linear mixed model.

### 6.1.2 Linear mixed model

The **linear mixed model** does something related to the above by-subject fits, but with some crucial differences, as we see below. In the model below, the statement (1|subj) means that the variance associated with subject intercepts should be estimated, and from that variance the intercepts for each subject should be predicted (I explain in section 6.1.5 how this prediction is done).

```
## the following command fits a linear model,
## but in addition estimates between-subject variance:
summary(m0.lmer<-lmer(rt~noise+(1|subj),noisedeg))

## Linear mixed model fit by REML ['lmerMod']
## Formula: rt ~ noise + (1 | subj)
##   Data: noisedeg
##
## REML criterion at convergence: 466.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.7538 -0.5452 -0.1984  0.5529  2.0306
##
## Random effects:
##   Groups   Name        Variance Std.Dev.
##   subj     (Intercept) 2491      49.91
##   Residual           8876      94.21
##   Number of obs: 40, groups: subj, 10
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 486.00     26.32 18.463
## noise noise  90.00     29.79  3.021
##
## Correlation of Fixed Effects:
##          (Intr)
## noise noise -0.566
```

One thing to notice is that the coefficients of the fixed effects of the above model are identical to those in the linear model `mo` above. The predicted varying intercepts for each subject can be viewed by typing:

```
ranef(m0.lmer)

## $subj
##   (Intercept)
## s1    -26.972985
## s10   -3.173292
## s2    -50.772677
## s3     36.492862
## s4     12.693169
## s5     28.559631
## s6    -66.639139
## s7     12.693169
## s8     28.559631
## s9     28.559631
```

Or you can display them graphically as shown in Figure 6.4.

The model `mo.lmer` above prints out the following type of linear model;  $i$  indexes subjects,  $j$  indexes the noise factor (no noise or noise). Since we have two replicates of each noise factor level (one for degree 0 and one for degree 4), the index  $k$  indexes the degree level, although the degree factor is not fit in this example (in order to simplify the example).

$$Y_{ijk} = \hat{\beta}_0 + \hat{\beta}_{1j}x_{ijk} + b_i + \epsilon_{ijk} \quad (6.1)$$

It's just like our linear model except that there are different *predicted* (cf. the `lmlist` function above, where they are *estimated* for each subject) intercepts  $b_i$  for each subject. These are assumed by `lmer` to come from a normal distribution centered around 0; see <sup>1</sup> for more details. The ordinary linear model `mo` has one intercept  $\beta_0$  for all subjects, whereas the linear mixed model with varying intercepts `mo.lmer` has a different intercept ( $\beta_0 + b_i$ ) for each subject.

We can visualize these different intercepts for each subject as shown below in Figure 6.5.

Note that, unlike the figure associated with the `lmlist.fm1` model above, which also involves fitting separate models for each subject, the model `mo.lmer` assumes different intercepts for each subject **but the same slope**. We can have `lmer` fit different intercepts as well as different slopes for each subject:

<sup>1</sup> A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, Cambridge, UK, 2007

```
print(dotplot(ranef(m0.lmer, condVar=TRUE)))
## $subj
```

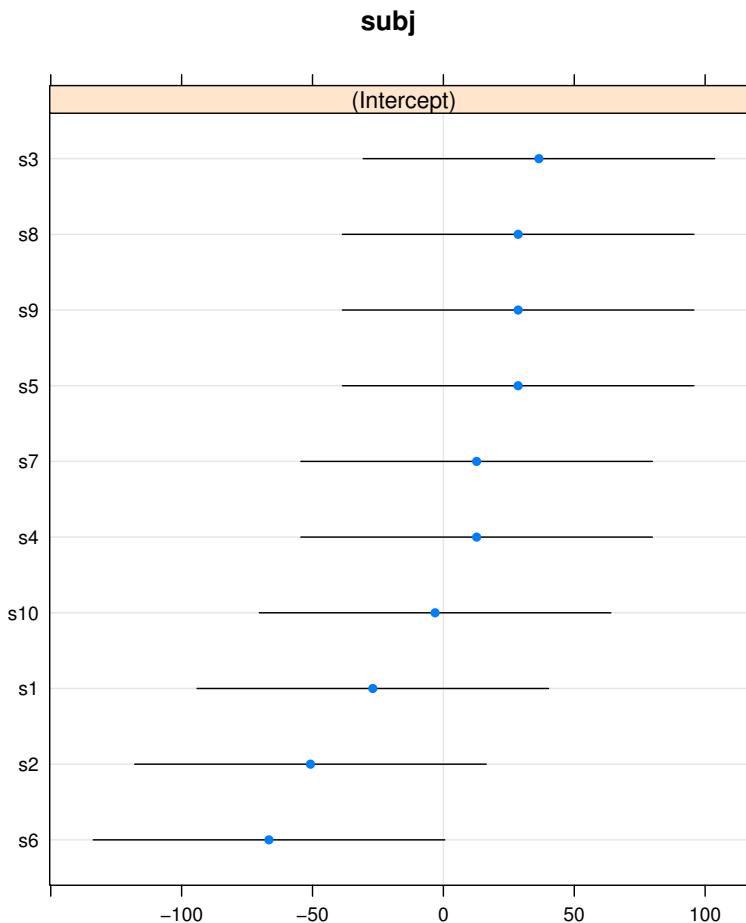


Figure 6.4: Plotting the varying intercepts.

```
summary(m1.lmer<-lmer(rt~noise+(1+noise|subj),noisedeg))

## Linear mixed model fit by REML ['lmerMod']
## Formula: rt ~ noise + (1 + noise | subj)
##   Data: noisedeg
##
## REML criterion at convergence: 465.1
##
## Scaled residuals:
##      Min     1Q Median     3Q    Max 
## -1.4441 -0.6729 -0.1930  0.6303  1.9593
##
```

```

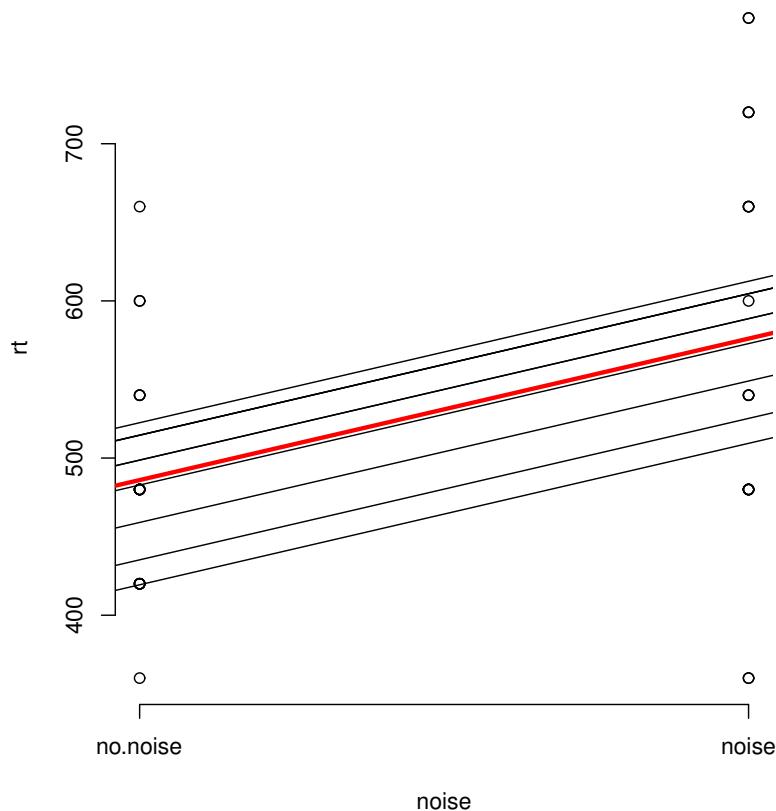
a<-fixef(m0.lmer)[1]
newa<-a+raneff(m0.lmer)$subj

ab<-data.frame(newa=newa,b=fixef(m0.lmer)[2])

plot(as.numeric(noisedeg$noise)-1,noisedeg$rt,xlab="noise",ylab="rt",axes=F)
axis(1,at=c(0,1),labels=c("no.noise","noise"))
axis(2)
for(i in 1:10){
  abline(a=ab[i,1],b=ab[i,2])
}
abline(lm(rt~noise,noisedeg),lwd=3,col="red")

```

Figure 6.5: Varying intercepts for subjects, with the linear model fit superimposed.



```

## Random effects:
## Groups   Name        Variance Std.Dev. Corr
## subj     (Intercept) 1093     33.05
##           noisenoise  1408     37.52    1.00
## Residual            8359     91.43
## Number of obs: 40, groups: subj, 10
##
## Fixed effects:
##                   Estimate Std. Error t value
## (Intercept)      486.00     22.96  21.17
## noisenoise       90.00     31.25   2.88
##
## Correlation of Fixed Effects:
##          (Intr)
## noisenoise -0.410

```

These fits for each subject are visualized below (the red line shows the model with a single intercept and slope, i.e., our old model *mo*):

```

(a<-fixef(m1.lmer)[1])

## (Intercept)
##      486

(b<-fixef(m1.lmer)[2])

## noisenoise
##      90

newa<-a+raneff(m1.lmer)$subj[1]
newb<-b+raneff(m1.lmer)$subj[2]
## make this into a data frame:
ab<-data.frame(newa=newa,b=newb)

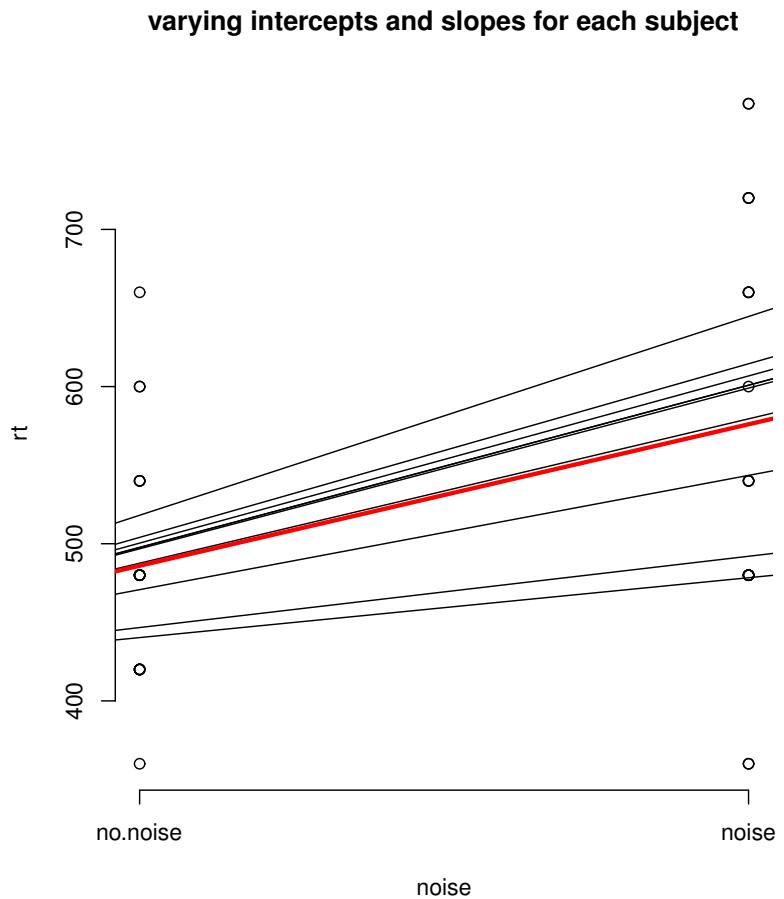
plot(as.numeric(noisedeg$noise)-1,noisedeg$rt,xlab="noise",ylab="rt",axes=F,
main="varying intercepts and slopes for each subject")
axis(1,at=c(0,1),labels=c("no.noise","noise"))

axis(2)

for(i in 1:10){
  abline(a=ab[i,1],b=ab[i,2])
}

abline(lm(rt~noise,noisedeg),lwd=3,col="red")

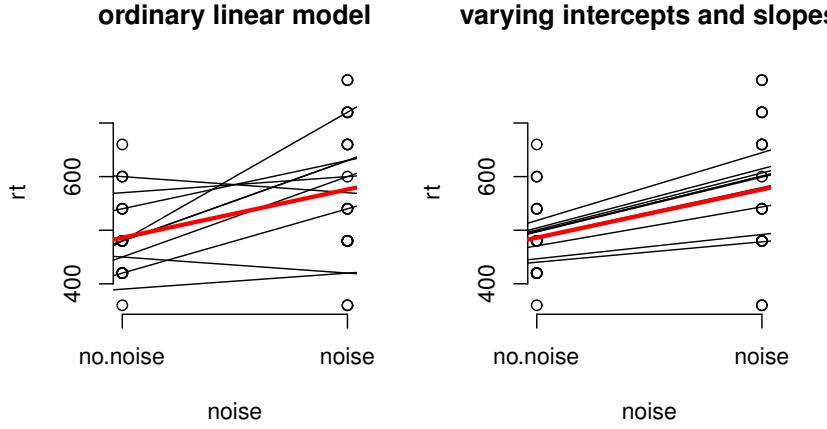
```



Compare this model with the `lmlist.fm1` model we fitted earlier; see Figure 6.6.

The above graphic shows some crucial differences between the `lmlist` (repeated measures) model and the `lmer` model. Note that the fitted line for each subject in the `lmer` model is much closer to the `mo` model's fitted (red) line. This is because `lmlist` uses each subject's data separately (resulting in possibly wildly different models, depending on the variability between subjects), whereas `lmer` "borrows strength from the mean" and pushes (or "shrinks") the estimated intercepts and slopes of each subject closer to the mean intercepts and slopes (the model `mo`'s intercepts and slopes). Because it shrinks the coefficients towards the means, this is called shrinkage. This is particularly useful when several data points are missing in a particular condition for a particular subject: in an ordinary linear model, estimating coefficients using `lmList` would lead to very poor estimates for that subject; by contrast, `lmer` assumes that the estimates for such a subject are not reliable and therefore shrinks that subject's estimate to the mean val-

Figure 6.6: Comparing the lmList and lmer estimates for each subject.



ues. Gelman and Hill provide an example of this (their Laq qui parle example).

### 6.1.3 Some basic types of linear mixed model and their variance components

**Varying intercepts model.** The model for a categorical predictor is:

$$Y_{ijk} = \beta_j + b_i + \epsilon_{ijk} \quad (6.2)$$

$i = 1, \dots, 10$  is subject id,  $j = 1, 2$  is the factor level,  $k$  is the number of replicates (here 1).  $b_i \sim N(0, \sigma_b^2)$ ,  $\epsilon_{ijk} \sim N(0, \sigma^2)$ .

For a continuous predictor:

$$Y_{ijk} = \beta_0 + \beta_1 t_{ijk} + b_{ij} + \epsilon_{ijk} \quad (6.3)$$

**Varying intercepts and slopes (with correlation).** The model for a categorical predictor is:

$$Y_{ij} = \beta_1 + b_{1i} + (\beta_2 + b_{2i})x_{ij} + \epsilon_{ij} \quad i = 1, \dots, M, j = 1, \dots, n_i \quad (6.4)$$

with  $b_{1i} \sim N(0, \sigma_1^2)$ ,  $b_{2i} \sim N(0, \sigma_2^2)$ , and  $\epsilon_{ij} \sim N(0, \sigma^2)$ .

Another way to write such models is:

$$Y_{ijk} = \beta_j + b_{ij} + \epsilon_{ijk} \quad (6.5)$$

$b_{ij} \sim N(0, \sigma_b)$ . The variance  $\sigma_b$  must be a  $2 \times 2$  matrix:

$$\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \quad (6.6)$$

In an lmer model, the output shows the variance covariance matrix of the random effects. Recall the degree noise data:

```
m<-lmer(rt~noise + (1+noise|subj),noisedeg)
summary(m)

## Linear mixed model fit by REML ['lmerMod']
## Formula: rt ~ noise + (1 + noise | subj)
##   Data: noisedeg
##
## REML criterion at convergence: 465.1
##
## Scaled residuals:
##      Min    1Q Median    3Q   Max
## -1.4441 -0.6729 -0.1930  0.6303  1.9593
##
## Random effects:
##   Groups   Name        Variance Std.Dev. Corr
##   subj     (Intercept) 1093     33.05
##          noisenoise  1408     37.52    1.00
##   Residual       8359     91.43
## Number of obs: 40, groups:  subj, 10
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 486.00     22.96  21.17
## noisenoise   90.00     31.25   2.88
##
## Correlation of Fixed Effects:
##          (Intr)
## noisenoise -0.410
```

Note also that the correlation estimate depends on the parameterization of the fixed effect noise:

```

contrasts(noisedeg$noise)

##          noise
## no.noise    0
## noise      1

## set to sum contrasts:
contrasts(noisedeg$noise)<-contr.sum(2)
contrasts(noisedeg$noise)

##          [,1]
## no.noise   1
## noise     -1

m<-lmer(rt~noise + (1+noise|subj),noisedeg)
summary(m)

## Linear mixed model fit by REML ['lmerMod']
## Formula: rt ~ noise + (1 + noise | subj)
##   Data: noisedeg
##
## REML criterion at convergence: 466.5
##
## Scaled residuals:
##    Min     1Q  Median     3Q    Max
## -1.4441 -0.6729 -0.1930  0.6303  1.9593
##
## Random effects:
## Groups   Name        Variance Std.Dev. Corr
## subj     (Intercept) 2684.7   51.81
##          noisel       351.9   18.76   -1.00
## Residual           8359.0   91.43
## Number of obs: 40, groups: subj, 10
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 531.00     21.85  24.30
## noisel      -45.00     15.63  -2.88
##
## Correlation of Fixed Effects:
##          (Intr)
## noisel -0.285

```

So the correlation here is estimated as `1` or `-1` depending on the parameterization. This estimate of a perfect correlation is actually a

**failure** to estimate the correlation, a consequence of overfitting the model (asking the lmer function to estimate parameters even though you don't have enough data to estimate them). The sensible thing to do here is to fit a model without such a correlation being assumed. There is an important technicality here, that you have to define the contrast as a vector of -1's and 1's (for sum contrast coding), instead of using the noise column in the data frame. The way to tell lmer not to estimate the correlation is to use two vertical bars in the random effects specification.

```
c1<-ifelse(noisedeg$noise=="noise",-1,1)
m<-lmer(rt~c1 + (c1||subj),noisedeg)
summary(m)

## Linear mixed model fit by REML [ 'lmerMod' ]
## Formula: rt ~ c1 + ((1 | subj) + (0 + c1 | subj))
##   Data: noisedeg
##
## REML criterion at convergence: 467.5
##
## Scaled residuals:
##      Min     1Q Median     3Q    Max
## -1.7538 -0.5452 -0.1984  0.5529  2.0306
##
## Random effects:
##   Groups   Name        Variance Std.Dev.
##   subj     (Intercept) 2.491e+03 4.991e+01
##   subj.1   c1          9.521e-13 9.757e-07
##   Residual           8.876e+03 9.421e+01
## Number of obs: 40, groups:  subj, 10
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  531.0      21.7  24.467
## c1         -45.0      14.9  -3.021
##
## Correlation of Fixed Effects:
##   (Intr)  c1
## c1  0.000
```

There is much more to be said here; see <sup>2</sup> for more.

<sup>2</sup> Douglas Bates, Reinhold Kliegl, Shrawan Vasishth, and Harald Baayen. Parsimonious mixed models. ArXiv e-print; submitted to *Journal of Memory and Language*, 2015

### 6.1.4 Parameter estimation

There are two basic procedures, likelihood based and REML (restricted or residual ML).

**The Likelihood based model fitting procedure.** Here are some relevant facts we need to know in order to work out how parameter estimation is done in LMMs using likelihoods.

1. If we have two continuous random variables  $Y$  and  $Z$ , with density functions  $f_Y(y)$  and  $f_Z(z)$  and joint density  $f_{Y,Z}(y,z)$ , then

$$f_Y(y) = \int f_{Y,Z}(y,z) dz. \quad (6.7)$$

2. The conditional density of  $Y | Z$  is defined as

$$f_{Y|Z}(y | z) = \frac{f_{Y,Z}(y,z)}{f_Z(z)} \quad (6.8)$$

so we can write

$$f_{Y,Z}(y,z) = f_{Y|Z}(y | z) \times f_Z(z). \quad (6.9)$$

3. Combining equations 6.7 and 6.9, we have

$$f_Y(y) = \int f_{Y|Z}(y | z) \times f_Z(z) dz \quad (6.10)$$

Equation 6.10, where we condition on a second random variable  $Z$  (note that  $Z$  could be a “non-observable”, which is what the varying intercepts and slopes are—they are not observed as data), can be helpful in deriving  $f_Y(y)$ , if the two densities on the RHS are easy to write down, and the integral can be solved.

Returning to parameter estimation in LMMs, the model is:

$$Y_i = X_i\beta + Z_i\beta_i + \epsilon_i, \quad i = 1, \dots, M \quad (6.11)$$

where  $b_i \sim N(0, \Psi)$ ,  $\epsilon_i \sim N(0, \sigma^2 I)$ . Let  $\theta$  be the parameters that determine  $\Psi$ .

$$\begin{aligned} L(\beta, \theta, \sigma^2 | y) &= p(y : \beta, \theta, \sigma^2) \\ &= \prod_i^M p(y_i : \beta, \theta, \sigma^2) \\ &= \prod_i^M \int p(y_i | b_i, \beta, \sigma^2) p(b_i : \theta, \sigma^2) db_i \end{aligned} \quad (6.12)$$

we want the density of the observations ( $y_i$ ) given the parameters  $\beta, \theta$  and  $\sigma^2$  only. In this case, using equation 6.10 above, with  $Y = y_i$  and

$Z = b_i$  is helpful for deriving the density for  $y_i$ , because  $f(y_i | b_i)$  (or  $p(y_i | b_i, \beta, \sigma^2)$ ) has a simple form, and so we can get a closed form expression for the integral.

**REML estimation (REstricted/REsidual ML).** To estimate variance parameters, first fit fixed effects using least squares, and then focus attention on residuals. The **residuals'** distribution depends on  $\sigma^2$  and variance parameters  $\theta$  of random effects.

A **likelihood** for these parameters is formed based on the residuals alone. Maximization of this **marginal likelihood** gives estimates of  $\sigma^2$  and the other variance-covariance parameters which are less biased than the full maximum likelihood estimates.

Once the REML variance-covariance estimates are obtained the **fixed effects are re-estimated by maximum likelihood assuming the random effects parameters are known.**

Alternatively, we can define a restricted likelihood:

$$L_R(\theta, \sigma^2 | y) = \int L(\beta, \theta, \sigma^2 | y) d\beta \quad (6.13)$$

and maximize this to obtain estimates of these parameters.

Unlike full (max.) likelihood, restricted likelihood is not invariant to parameterization, so we cannot compare models with different fixed effects.

### 6.1.5 Computing the BLUPs

In a varying intercepts model, the intercept  $b_i$  for subjects  $i = 1, \dots, I$  (assume that each subject has  $n_i$  data points) is a latent variable, it cannot be observed. So how is it computed? We can predict these BLUPs using a method called **empirical Bayes estimation**, and generate something called posterior means. This method combines two kinds of information: (a) the data from group  $i$ , and (b) the fact that the unobserved  $b_i$  is a random variable with mean  $0$  and variance  $\sigma_b^2$ .

The posterior means  $\hat{b}_i$  are given by

$$\hat{b}_i = E[\hat{b}_i | y, \beta] = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2/n_i} (\bar{y}_{\cdot i} - \bar{y}) \quad (6.14)$$

You have never seen an expression like  $\bar{y}_{\cdot i}$ , but it just means the group means (that subject's mean score over the  $n_i$  data points). The term  $\frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2/n_i}$  is called a shrinkage factor; it can be at most 1. It will approach 1 when we have  $n_i$  approaching infinity, and will be small if  $n_i$  is a small number. The implication of this shrinkage factor is that if we have only a few data points for a particular subject, then its value will be shrunk towards the grand mean—the example we saw earlier.

### 6.1.6 Correlation of fixed effects

For an ordinary linear model, the covariance matrix (from which we can get the correlation matrix) of  $\hat{\beta}$  is:

$$\sigma^2 \times (X^T X)^{-1}. \quad (6.15)$$

For a mixed effects model, the standard deviations (standard errors) and correlations for the fixed effects estimators are listed at the end of the lmer output. Here is an example from a data-set that measures some dependent variable “wear” as a function of three material types for 10 subjects:

```
BHHshoes<-read.table("datacode/BHHshoes.txt")
lm.full<-lmer(wear~material-1+
  (1|Subject),
  data = BHHshoes)
```

Correlation of Fixed Effects:

```
matrlA
materialB 0.988
```

Doing this by hand:

$$\hat{\beta}_1 = (Y_{1,1} + Y_{2,1} + \dots + Y_{10,1})/10 \quad (6.16)$$

$$\hat{\beta}_2 = (Y_{1,2} + Y_{2,2} + \dots + Y_{10,2})/10 \quad (6.17)$$

```
b1.vals<-subset(BHHshoes,
  material=="A")$wear
b2.vals<-subset(BHHshoes,
  material=="B")$wear

vcovmatrix<-var(cbind(b1.vals,b2.vals))

## get covariance from off-diagonal:
covar<-vcovmatrix[1,2]
sds<-sqrt(diag(vcovmatrix))
## correlation of fixed effects:
covar/(sds[1]*sds[2])

## b1.vals
## 0.9882255

#cf:
covar/((0.786*sqrt(10))^2)
```

```
## [1] 0.9875248
```

## 6.2 Bayesian data analysis: Some introductory ideas

Recall Bayes rule:

**Theorem 2 Bayes' Rule.** Let  $B_1, B_2, \dots, B_n$  be mutually exclusive and exhaustive and let  $A$  be an event with  $\mathbb{P}(A) > 0$ . Then

$$\mathbb{P}(B_k|A) = \frac{\mathbb{P}(B_k)\mathbb{P}(A|B_k)}{\sum_{i=1}^n \mathbb{P}(B_i)\mathbb{P}(A|B_i)}, \quad k = 1, 2, \dots, n. \quad (6.18)$$

When A and B are observable events, we can state the rule as follows:

$$p(A | B) = \frac{p(B | A)p(A)}{p(B)} \quad (6.19)$$

Note that  $p(\cdot)$  is the probability of an event.

When looking at probability distributions, we will encounter the rule in the following form.

$$f(\theta | \text{data}) = \frac{f(\text{data} | \theta)f(\theta)}{f(y)} \quad (6.20)$$

Here,  $f(\cdot)$  is a probability density, not the probability of a single event.  $f(y)$  is called a “normalizing constant”, which makes the left-hand side a probability distribution.

$$f(y) = \int f(x, \theta) d\theta = \int \underset{\substack{\uparrow \\ \text{likelihood}}}{f(y | \theta)} \underset{\substack{\uparrow \\ \text{prior}}}{f(\theta)} d\theta \quad (6.21)$$

If  $\theta$  is a discrete random variable taking one value from the set  $\{\theta_1, \dots, \theta_n\}$ , then

$$f(y) = \sum_{i=1}^n f(y | \theta_i) P(\theta = \theta_i) \quad (6.22)$$

Without the normalizing constant, we have the relationship:

$$f(\theta | \text{data}) \propto f(\text{data} | \theta)f(\theta) \quad (6.23)$$

Note that the likelihood  $L(\theta; \text{data})$  (our data is fixed) is proportional to  $f(\text{data} | \theta)$ , and that's why we can refer to  $f(\text{data} | \theta)$  as the likelihood in the following manner:

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior} \quad (6.24)$$

Our central goal is going to be to derive the posterior distribution and then summarize its properties (mean, median, 95% credible interval, etc.). Usually, we don't need the normalizing constant to understand the properties of the posterior distribution. That's why Bayes theorem is often stated in terms of the proportionality shown above.

Incidentally, this is supposed to be the moment of great divide between frequentists and Bayesians: the latter assign a probability distribution to the parameter, the former treat the parameter as a point value.

Two examples will clarify how we can use Bayes' rule to obtain the posterior.

"To a Bayesian, the best information one can ever have about  $\theta$  is to know the posterior density." p. 31 of Christensen et al's book.

### 6.2.1 Example 1: Proportions

This is a contrived example, just meant to provide us with an entry point into Bayesian data analysis. Suppose that an aphasic patient answered 46 out of 100 questions correctly in a particular task. The research question is, what is the probability that their average response is greater than 0.5, i.e., above chance.

The likelihood function will tell us  $P(\text{data} \mid \theta)$ :

```
dbinom(46, 100, 0.5)
## [1] 0.0579584
```

Note that

$$P(\text{data} \mid \theta) \propto \theta^{46}(1 - \theta)^{54} \quad (6.25)$$

So, to get the posterior, we just need to work out a prior distribution  $f(\theta)$ .

$$f(\theta \mid \text{data}) \propto f(\text{data} \mid \theta)f(\theta) \quad (6.26)$$

$\uparrow$   
*prior*

For the prior, we need a distribution that can represent our uncertainty about the parameter  $p$ . The beta distribution (a generalization of the continuous uniform distribution) is commonly used as prior for proportions.

The pdf is<sup>3</sup>

$$f(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

where

$$B(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1} dx$$

We say that the Beta distribution is conjugate to the binomial density; i.e., the two densities have similar functional forms.

<sup>3</sup> Incidentally, there is a connection between the beta and the gamma:

$$B(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1} dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

which allows us to rewrite the beta PDF as

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1. \quad (6.27)$$

Here,  $x$  refers to the probability  $p$ .

In R, we write  $X \sim \text{beta}(\text{shape1} = \alpha, \text{shape2} = \beta)$ . The associated R function is `dbeta(x, shape1, shape2)`.

The mean and variance are

$$E[X] = \frac{a}{a+b} \text{ and } \text{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}. \quad (6.28)$$

The Beta distribution's parameters  $a$  and  $b$  can be interpreted as (our beliefs about) prior successes and failures, and are called **hyper-parameters**. Once we choose values for  $a$  and  $b$ , we can plot the beta pdf. in Figure 6.7, I show the Beta pdf for three sets of values of  $a,b$ .

As the figure shows, as the  $a,b$  values are increased, the shape begins to resemble the normal distribution.

If we don't have much prior information, we could use  $a=b=2$ ; this gives us a uniform prior; this is called an uninformative prior or non-informative prior (although having no prior knowledge is, strictly speaking, not uninformative). If we have a lot of prior knowledge and/or a strong belief that  $p$  has a particular value, we can use a larger  $a,b$  to reflect our greater certainty about the parameter. Notice that the larger our parameters  $a$  and  $b$ , the smaller the spread of the distribution; this makes sense because a larger sample size (a greater number of successes  $a$ , and a greater number of failures  $b$ ) will lead to more precise estimates.

The central point is that the Beta distribution can be used to define the prior distribution of  $p$ .

Just for the sake of argument, let's take four different beta priors, each reflecting increasing certainty.

1. Beta( $a=2, b=2$ )
2. Beta( $a=3, b=3$ )
3. Beta( $a=6, b=6$ )
4. Beta( $a=21, b=21$ )

Each (except perhaps the first) reflects a belief that  $p=0.5$ , with varying degrees of (un)certainty. Now we just need to plug in the likelihood and the prior:

$$f(\theta | \text{data}) \propto f(\text{data} | \theta) f(\theta) \quad (6.29)$$

The four corresponding posterior distributions would be:

$$f(\theta | \text{data}) \propto [p^{46}(1-p)^{54}][p^{2-1}(1-p)^{2-1}] = p^{47}(1-p)^{55} \quad (6.30)$$

$$f(\theta | \text{data}) \propto [p^{46}(1-p)^{54}][p^{3-1}(1-p)^{3-1}] = p^{48}(1-p)^{56} \quad (6.31)$$

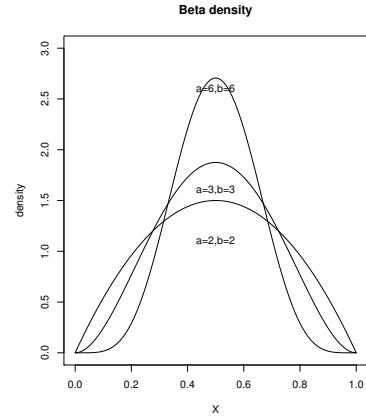


Figure 6.7: Examples of the beta distribution with different parameter values.

$$f(\theta | \text{data}) \propto [p^{46}(1-p)^{54}][p^{6-1}(1-p)^{6-1}] = p^{51}(1-p)^{59} \quad (6.32)$$

$$f(\theta | \text{data}) \propto [p^{46}(1-p)^{54}][p^{21-1}(1-p)^{21-1}] = p^{66}(1-p)^{74} \quad (6.33)$$

We can now visualize each of these triplets of priors, likelihoods and posteriors. Note that I use the Beta to model the likelihood because this allows me to visualize all three (prior, lik., posterior) in the same plot. The likelihood function is as shown in Figure 6.8.

We can represent the likelihood in terms of the beta as well:

As an exercise, you should try to plot the priors, likelihoods, and posterior distributions in the four cases above. I do the first case for you below, but I don't plot the result.

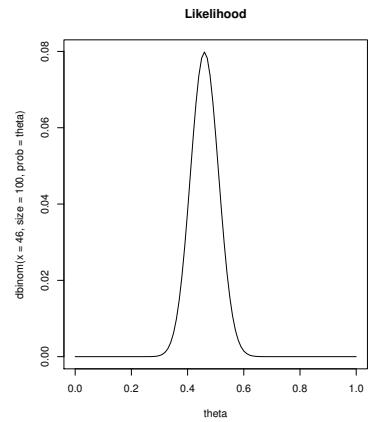
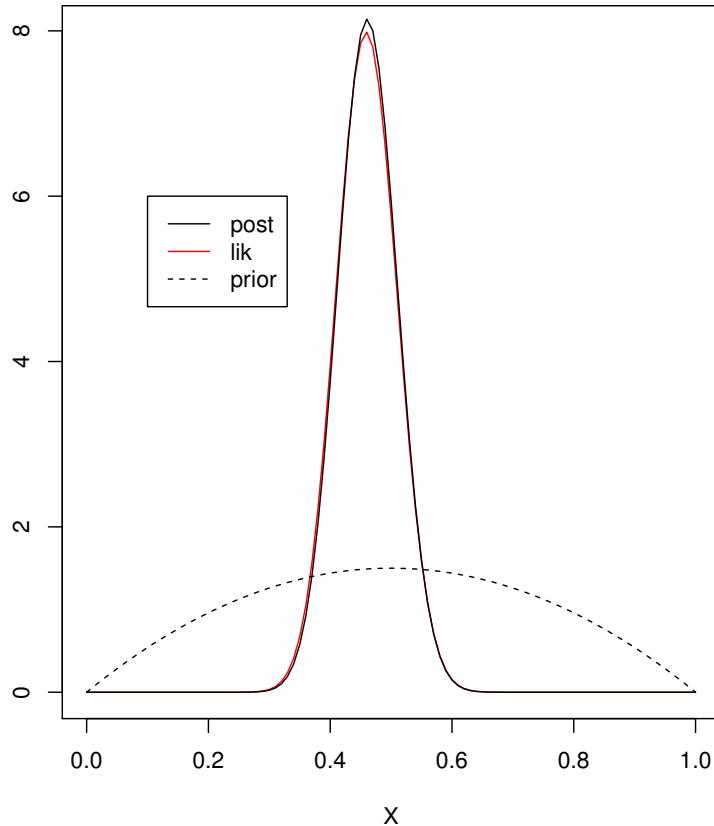


Figure 6.8: Binomial likelihood function.

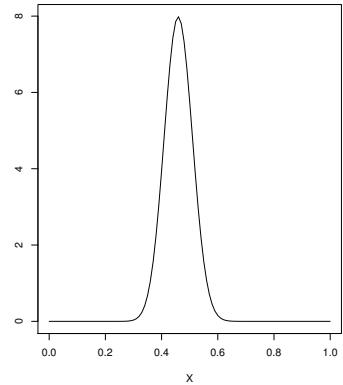


Figure 6.9: Using the beta distribution to represent a binomial.

### 6.2.2 Example 2: Proportions

This example is taken from a Copenhagen course on Bayesian data analysis that I attended in 2012.

For a single Bernoulli trial, the **likelihood** for possible parameters value  $\theta_j$  (which we, for simplicity, fix at four values, 0.2, 0.4, 0.6, 0.8.) is:<sup>4</sup>

$$p(y | \theta_j) = \theta_j^y (1 - \theta_j)^{1-y} \quad (6.34)$$

<sup>4</sup> See section 2.1.1 for the definition of the binomial distribution. The Bernoulli distribution is the binomial with n=1.

```
y<-1
n<-1

thetas<-seq(0.2,0.8,by=0.2)

likelihoods<-rep(NA,4)
for(i in 1:length(thetas)){
  likelihoods[i]<-dbinom(y,n,thetas[i])
}
```

Note that these do not sum to 1:

```
sum(likelihoods)

## [1] 2
```

Question: How can we make them sum to 1? I.e., how can we make the likelihood a proper probability mass function? Think about this.

Let the **prior** distribution of the parameters be  $p(\theta_j)$ ; let this be a uniform distribution over the possible values of  $\theta_j$ .

```
(priors<-rep(0.25,4))

## [1] 0.25 0.25 0.25 0.25
```

The prior is a proper probability distribution.

For any outcome  $y$  (which could be 1 or 0—this is a single trial), the posterior probability of success (a 1) is related to the prior and likelihood by the relation:

$$p(\theta_j | y) \propto p(\theta_j) \theta_j^y (1 - \theta_j)^{1-y} \quad (6.35)$$

To get the posterior to be a proper probability distribution, you have to make sure that the RHS sums to 1.

```

liks.times.priors<-likelihoods * priors

## normalizing constant:
sum.lik.priors<-sum(liks.times.priors)

posteriori<- liks.times.priors/sum.lik.priors

```

Note that the posterior distribution sums to 1, because we **normalized** it.

Now suppose our sample size was 20, and the number of successes 15. What does the posterior distribution look like now?

```

n<-20
y<-15

priors<-rep(0.25,4)

likelihoods<- rep(NA,4)
for(i in 1:length(thetas)){
  likelihoods[i]<-dbinom(y,n,thetas[i])
}

liks.priors<-likelihoods * priors

sum.lik.priors<-sum(liks.priors)

(posteriori<- liks.priors/sum.lik.priors)

## [1] 6.645594e-07 5.167614e-03 2.979907e-01 6.968411e-01

```

Now suppose that we had a non-zero prior probability to extreme values (0,1) to  $\theta$ . The prior is now defined over six values, not four, so the probability distribution on the priors changes accordingly to 1/6 for each value of  $\theta$ .

Given the above situation of  $n=20$ ,  $y=15$ , what will change in the posterior distribution compared to what we just computed:

```

posteriori

## [1] 6.645594e-07 5.167614e-03 2.979907e-01 6.968411e-01

```

Let's find out what the posteriors will look like:

```

thetas<-seq(0,1,by=0.2)
priors<-rep(1/6,6)

```

```

y<-15
n<-20

likelihoods<-rep(NA,6)
for(i in 1:length(thetas)){
  likelihoods[i]<-dbinom(y,n,thetas[i])
}

liks.priors<-likelihoods * priors

sum.lik.priors<-sum(liks.priors)

(posteriors<- liks.priors/sum.lik.priors)

## [1] 0.000000e+00 6.645594e-07 5.167614e-03 2.979907e-01 6.968411e-01
## [6] 0.000000e+00

```

How would the posteriors change if we had only one trial (a Bernoulli trial)? Let's find out:

```

thetas<-seq(0,1,by=0.2)
priors<-rep(1/6,6)

y<-1
n<-1

j<-6 ## no. of thetas
likelihoods<-rep(NA,6)
for(i in 1:length(thetas)){
  likelihoods[i]<-dbinom(y,n,thetas[i])
}

liks.priors<-likelihoods * priors

sum.lik.priors<-sum(liks.priors)

posterior<- liks.priors/sum.lik.priors

```

We have been using discrete prior distributions so far. We might want to use a continuous prior distribution if we have prior knowledge that, say, the true value lies between 0.2 and 0.6, with mean 0.4. If our prior should have mean 0.4 and sd 0.1, we can figure out what the corresponding parameters of the beta distribution should be. (Look up the mean and variance of the beta distribution, and solve for the parameters.) You should be able to prove analytically that  $a=9.2$ ,  $b=13.8$ .

Let's plot the prior (Figure 6.10). Then plot the likelihood (Figure 6.11). Recall that the likelihood is a function of the parameters  $\theta$ .

This likelihood can equally well be presented as a Beta distribution because the Beta distribution's parameters  $a$  and  $b$  can be interpreted as prior successes and failures. See Figure 6.12.

Since we multiply the Beta distribution representing the prior and the beta distribution representing the likelihood:

$$\text{Beta}(9.2, 13.8) * \text{Beta}(15, 5) = \text{Beta}(a = 9.2 + 15, b = 13.8 + 5) \quad (6.36)$$

We can also plot all three (prior, likelihood, posterior) in one figure. That's an exercise for you.

As a further exercise, you may want to write a generic function that, for a given set of values for the prior (given mean's and sd's), and given the data (number of successes and failures), plots the appropriate posterior (alongside the priors and likelihood).

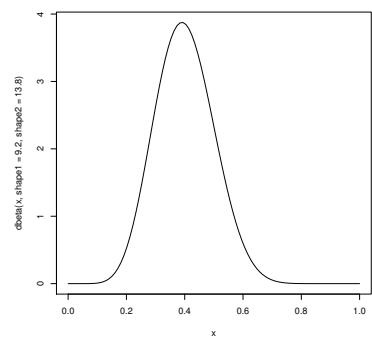


Figure 6.10: The prior in terms of the beta distribution.

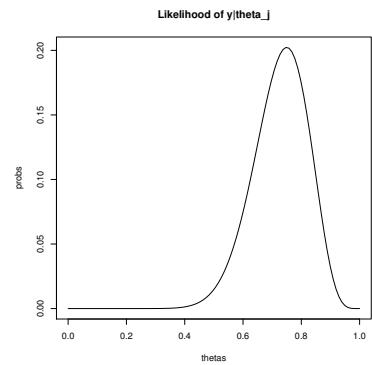


Figure 6.11: Likelihood.

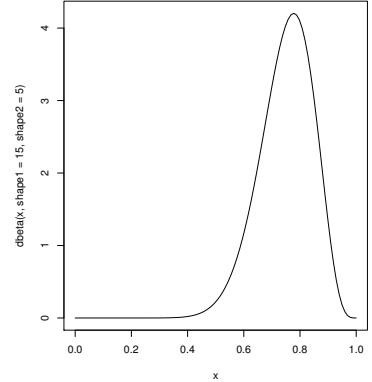


Figure 6.12: Likelihood in terms of the beta.

### 6.2.3 Exercise: The proportion of female births in France

Here is another exercise to help you solidify your understanding of the above concepts. It is taken from the Copenhagen course of 2012 that I mentioned earlier. I quote it verbatim.

"The French mathematician Pierre-Simon Laplace (1749-1827) was the first person to show definitively that the proportion of female births in the French population was less than 0.5, in the late 18th century, using a Bayesian analysis based on a uniform prior distribution<sup>5</sup>). Suppose you were doing a similar analysis but you had more definite prior beliefs about the ratio of male to female births. In particular, if  $\theta$  represents the proportion of female births in a given population, you are willing to place a Beta(100,100) prior distribution on  $\theta$ .

5

1. Show that this means you are more than 95% sure that  $\theta$  is between 0.4 and 0.6, although you are ambivalent as to whether it is greater or less than 0.5.
2. Now you observe that out of a random sample of 1,000 births, 511 are boys. What is your posterior probability that  $\theta > 0.5$ ?"

Here is yet another exercise, taken from Lunn et al.<sup>6</sup>, their example 3.1.1. Suppose that 1 in 1000 people in a population are (or is) expected to get HIV. Suppose a test is administered on a suspected HIV case, where the test has a true positive rate (the proportion of positives that are actually HIV positive) of 95% and true negative rate (the proportion of negatives that are actually HIV negative) 98%. Use Bayes theorem to find out the probability that a patient testing positive actually has HIV.

<sup>6</sup> David Lunn, Chris Jackson, David J Spiegelhalter, Nicky Best, and Andrew Thomas. *The BUGS book: A practical introduction to Bayesian analysis*, volume 98. CRC Press, 2012

Next, we turn to a key idea in Bayesian data analysis: **the posterior is a compromise between the prior and the likelihood**. Let's look at some specific examples.

### 6.2.4 The posterior is the weighted mean of the prior mean and the MLE

Suppose we are modeling the number of times that a speaker says the word "the" per day.

The number of times  $x$  that the word is uttered in one day can be modeled by a Poisson distribution:

$$f(x | \theta) = \frac{\exp(-\theta)\theta^x}{x!} \quad (6.37)$$

where the rate  $\theta$  is unknown, and the numbers of utterances of the target word on each day are independent given  $\theta$ .

Let's say we are told that the prior mean of  $\theta$  is 100 and prior variance for  $\theta$  is 225. This expresses a prior belief, based for example on prior knowledge (existing data), expert opinion or the like.

We can fit a Gamma density prior for  $\theta$  based on the above information.

It is known from standard statistical theory that for a Gamma density with parameters  $a, b$ , the mean is  $\frac{a}{b}$  and the variance is  $\frac{a}{b^2}$ . Since we are given values for the mean and variance, we can solve for  $a, b$ , which gives us the gamma density.

If  $\frac{a}{b} = 100$  and  $\frac{a}{b^2} = 225$ , it follows that  $a = 100 \times b = 225 \times b^2$  or  $100 = 225 \times b$ , i.e.,  $b = \frac{100}{225}$ .

This means that  $a = \frac{100 \times 100}{225} = \frac{10000}{225}$ . Therefore, the Gamma distribution for the prior is as shown below (also see Fig 6.13):

$$\theta \sim \text{Gamma}\left(\frac{10000}{225}, \frac{100}{225}\right) \quad (6.38)$$

A distribution for a prior is **conjugate** if, multiplied by the likelihood, it yields a posterior that has the distribution of the same family as the prior.

It turns out that the Gamma distribution is a conjugate prior for the Poisson distribution. That's why we chose a prior expressed in terms of the Gamma.

For the Gamma distribution to be a conjugate prior for the Poisson, the posterior needs to have the general form of a Gamma distribution.

Given that

$$\text{Posterior} \propto \text{Prior Likelihood} \quad (6.39)$$

and given that the likelihood is:

$$\begin{aligned} L(\mathbf{x} | \theta) &= \prod_{i=1}^n \frac{\exp(-\theta)\theta^{x_i}}{x_i!} \\ &= \frac{\exp(-n\theta)\theta^{\sum_i^n x_i}}{\prod_{i=1}^n x_i!} \end{aligned} \quad (6.40)$$

we can compute the posterior as follows:

$$\text{Posterior} = \left[ \frac{\exp(-n\theta)\theta^{\sum_i^n x_i}}{\prod_{i=1}^n x_i!} \right] \left[ \frac{b^a \theta^{a-1} \exp(-b\theta)}{\Gamma(a)} \right] \quad (6.41)$$

Disregarding the terms  $x!, \Gamma(a), b^a$ , which do not involve  $\theta$ , we have

$$\begin{aligned} \text{Posterior} &\propto \exp(-n\theta)\theta^{\sum_i^n x_i} \theta^{a-1} \exp(-b\theta) \\ &= \theta^{a-1 + \sum_i^n x_i} \exp(-\theta(b+n)) \end{aligned} \quad (6.42)$$

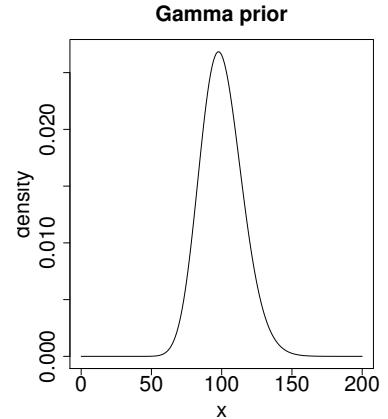


Figure 6.13: The gamma prior for  $\theta$ .

We can figure out the parameters of the posterior distribution, and show that it will be a Gamma distribution. Note that the Gamma distribution in general is  $\text{Gamma}(a, b) \propto \theta^{a-1} \exp(-\theta b)$ . So it's enough to state the above as a Gamma distribution with some parameters  $a^*$ ,  $b^*$ .

If we equate  $a^* - 1 = a - 1 + \sum_i^n x_i$  and  $b^* = b + n$ , we can rewrite the above as:

$$\theta^{a^*-1} \exp(-\theta b^*) \quad (6.43)$$

This means that  $a^* = a + \sum_i^n x_i$  and  $b^* = b + n$ . We can find a constant  $k$  such that the above is a proper probability density function, i.e.:

$$\int_{-\infty}^{\infty} k\theta^{a^*-1} \exp(-\theta b^*) = 1 \quad (6.44)$$

Thus, the posterior has the form of a Gamma distribution with parameters  $a^* = a + \sum_i^n x_i$ ,  $b^* = b + n$ . Hence the Gamma distribution is a conjugate prior for the Poisson.

### 6.2.5 Example: The Poisson-Gamma conjugate case

Returning to our specific example, if we are given that the number of "the" utterances is 115, 97, 79, 131, we can derive the posterior distribution as follows.

The prior is  $\text{Gamma}(a=10000/225, b=100/225)$ . The data are as given; this means that  $\sum_i^n x_i = 422$  and sample size  $n = 4$ . It follows that the posterior is

$$\begin{aligned} \text{Gamma}(a^* = a + \sum_i^n x_i, b^* = b + n) &= \text{Gamma}(10000/225 + 422, 4 + 100/225) \\ &= \text{Gamma}(466.44, 4.44) \end{aligned} \quad (6.45)$$

The mean and variance of this distribution can be computed using the fact that the mean is  $\frac{a^*}{b^*} = 466.44/4.44 = 104.95$  and the variance is  $\frac{a^*}{b^{*2}} = 466.44/4.44^2 = 23.61$ .

We can do this in R as follows.

```
## load data:
data<-c(115, 97, 79, 131)

a.star<-function(a,data){
  return(a+sum(data))
}
```

```

b.star<-function(b,n){
  return(b+n)
}

new.a<-a.star(10000/225,data)
new.b<-b.star(100/225,length(data))

## post. mean
post.mean<-new.a/new.b
## post. var:
post.var<-new.a/(new.b^2)

new.data<-c(200)

new.a.2<-a.star(new.a,new.data)
new.b.2<-b.star(new.b,length(new.data))

## new mean
new.post.mean<-new.a.2/new.b.2
## new var:
new.post.var<-new.a.2/(new.b.2^2)

```

### 6.2.6 Using JAGS for the Poisson-Gamma conjugate example

This is also a good place to introduce an important programming language for Bayesian data analysis. One such language is JAGS <sup>7</sup>. I present below the JAGS version of the closed-form solution we computed above.

```

## specify data:
dat<-list(y=c(115,97,79,131))

## model specification:
cat("
model
{
  for(i in 1:4){
    y[i] ~ dpois(theta)
  }
  ##prior
  ## gamma params derived from given info:
  theta ~ dgamma(10000/225,100/225)
}"

```

<sup>7</sup> Martyn Plummer. JAGS version 3.3.0 manual. International Agency for Research on Cancer. Lyon, France, 2012

```

  file="datacode/poissonexample.jag" )

## specify variables to track
## the posterior distribution of:
track.variables<-c("theta")

## load rjags library:
library(rjags,quietly=T)

## Linked to JAGS 4.1.0
## Loaded modules: basemod,bugs

## define model:
pois.mod <- jags.model(
  data = dat,
  file = "datacode/poissonexample.jag",
  n.chains = 4,
  n.adapt =2000 ,quiet=T)

## run model:
pois.res <- coda.samples( pois.mod,
  var = track.variables,
  n.iter = 50000,
  thin = 50 )

```

We can also visualize the results as shown in Figure 6.14.

According to the JAGS model, the mean is 105.01, and the variance is 23.38. This matches up well with the above analytically derived results. I got these results by typing:

```

print(summary(pois.res))

##
## Iterations = 50:50000
## Thinning interval = 50
## Number of chains = 4
## Sample size per chain = 1000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean          SD      Naive SE Time-series SE
## 105.01238   4.83541   0.07645   0.07166
##
## 2. Quantiles for each variable:

```

```
## summarize and plot:
plot(pois.res)
```

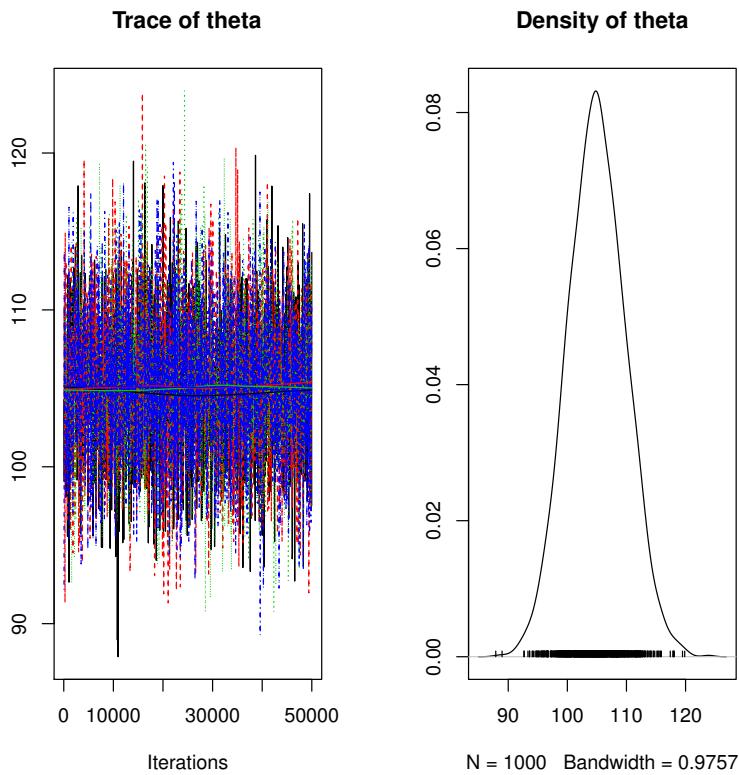


Figure 6.14: Plot showing the results of the JAGS model fit.

```
##  
##   2.5%    25%    50%    75%  97.5%  
##  95.67 101.70 104.95 108.27 114.60
```

The plot in Figure 6.14 shows, on the left hand side, the sampling from the posterior distribution done by JAGS. The word “sampling” here means that given some probability distribution function, we can sample from it using some standard computational statistical methods (e.g., Gibbs sampling, Metropolis-Hastings sampling, Hamiltonian Monte Carlo). Recall that we know how to sample from the normal distribution or the like, using the `rnorm` function; the JAGS software makes it possible to sample from a probability density function that we know up to proportionality and may not have some built-in R function for. In this course, we don’t have time to discuss these methods, but see the Dobson et al book for discussion, and also see my lecture notes on BDA (this course is taught in winter semesters). If you see “fat hairy caterpillars” in the left-hand side plot, this means

that JAGS is starting to sample from the target distribution and has converged as regards sampling from the posterior distribution.

The right-hand side plot in Figure 6.14 shows the posterior distribution of the parameter of interest.

We can plot the prior, likelihood, and posterior associated with the above model fit in a single figure (see Figure 6.15).

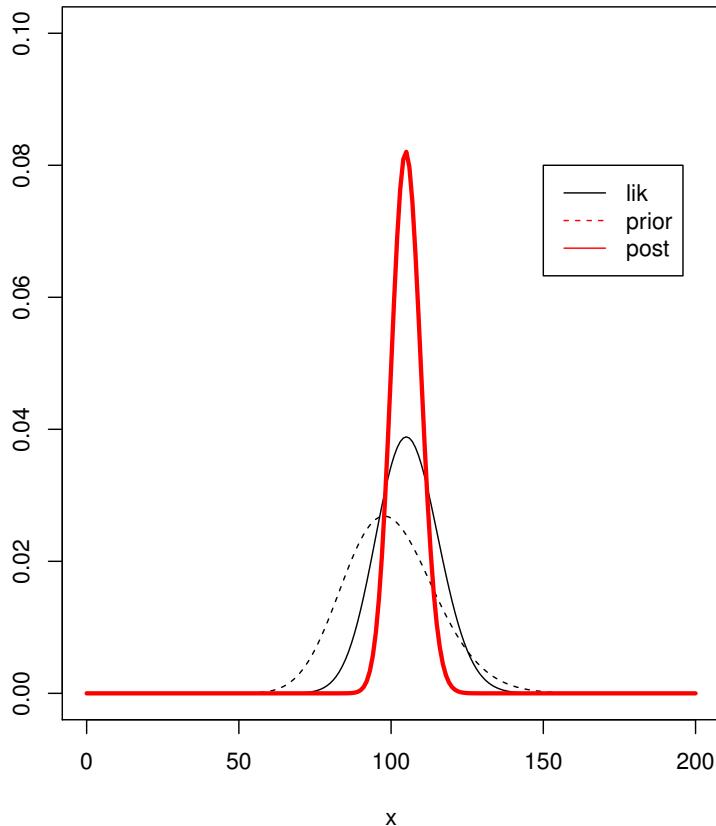


Figure 6.15: The broken line is the prior (Gamma), the solid black line is the likelihood, and the thick red line is the posterior.

### 6.2.7 The posterior in the Poisson-Gamma case as a weighted sum

In this example too, we can express the posterior mean as a weighted sum of the prior mean and the maximum likelihood estimate of  $\theta$ .

The posterior mean is:

$$\frac{a^*}{b^*} = \frac{a + \sum x_i}{n + b} \quad (6.46)$$

This can be rewritten as

$$\frac{a^*}{b^*} = \frac{a + n\bar{x}}{n + b} \quad (6.47)$$

Dividing both the numerator and denominator by  $b$ :

$$\frac{a^*}{b^*} = \frac{(a + n\bar{x})/b}{(n + b)/b} = \frac{a/b + n\bar{x}/b}{1 + n/b} \quad (6.48)$$

Since  $a/b$  is the mean  $m$  of the prior, we can rewrite this as:

$$\frac{a/b + n\bar{x}/b}{1 + n/b} = \frac{m + \frac{n}{b}\bar{x}}{1 + \frac{n}{b}} \quad (6.49)$$

We can rewrite this as:

$$\frac{m + \frac{n}{b}\bar{x}}{1 + \frac{n}{b}} = \frac{m \times 1}{1 + \frac{n}{b}} + \frac{\frac{n}{b}\bar{x}}{1 + \frac{n}{b}} \quad (6.50)$$

This is a weighted average: setting  $w_1 = 1$  and  $w_2 = \frac{n}{b}$ , we can write the above as:

$$m \frac{w_1}{w_1 + w_2} + \bar{x} \frac{w_2}{w_1 + w_2} \quad (6.51)$$

A  $n$  approaches infinity, the weight on the prior mean  $m$  will tend towards 0, making the posterior mean approach the maximum likelihood estimate of the sample.

This brings us to the key point: **as sample size increases, the likelihood will dominate in determining the posterior mean.**

Regarding variance, since the variance of the posterior is:

$$\frac{a^*}{b^{*2}} = \frac{(a + n\bar{x})}{(n + b)^2} \quad (6.52)$$

as  $n$  approaches infinity, the posterior variance will approach zero, which makes sense: more data will reduce variance (uncertainty).

### 6.2.8 Exercise: Using the posterior as a prior for new data

Here is an exercise you can do to understand the concepts above a bit better. If we have additional data from two weeks, with a count of 200,  $\sum x_i = 422 + 200 = 622$  and  $n = 6$  (not 5, because it is two weeks' data).<sup>8</sup> The task is to find the new posterior distribution given this new data. Try doing this yourself first before reading on.

**Solution:** The posterior would be a Gamma distribution with parameters:  $a^{**} = a + \sum_i^6 x_i = 10000/225 + 622 = 666.44$  and  $b^{**} = b + 6 = 100/225 + 6 = 6.44$ . In other words, the mean of the posterior is 103.41 and the variance is 16.05.

We can verify this using JAGS.

<sup>8</sup> One can also demonstrate this by multiplying the likelihood of the five data points; for the two weeks' measurement, the likelihood would be proportional to  $\exp(-2\theta)(2\theta)^x$ . Given that  $n = 4$  for the original data, this likelihood for the new data has the effect that the likelihood ends up being proportional to  $\exp(-\theta(n+2))(\theta)^{\sum x}$ .

```

dat2<-list(y=c(115,97,79,131,200))

## model specification:
cat("
model
{
for(i in 1:4){
  y[i] ~ dpois(theta)
}
y[5] ~ dpois(2*theta)

##prior
## gamma params derived from given info:
theta ~ dgamma(10000/225,100/225)
}",
      file="datacode/poisexample2.jag" )

## specify variables to track
## the posterior distribution of:
track.variables<-c("theta")

## define model:
poisex2.mod <- jags.model(
  data = dat2,
  file = "datacode/poisexample2.jag",
  n.chains = 4,
  n.adapt =2000 ,quiet=T)

## run model:
poisex2.res <- coda.samples( poisex2.mod,
                           var = track.variables,
                           n.iter = 100000,
                           thin = 50 )

```

JAGS returns the mean as 103.4, and the variance as 16.19. This is quite close to the analytically computed values.

Again, I got this summary by typing:

```

print(summary(poisex2.res))

##
## Iterations = 50:1e+05
## Thinning interval = 50
## Number of chains = 4

```

```

## Sample size per chain = 2000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean        SD     Naive SE Time-series SE
## 103.39724   4.02379   0.04499   0.04498
##
## 2. Quantiles for each variable:
##
##    2.5%    25%    50%    75%   97.5%
## 95.51 100.68 103.39 106.10 111.42

```

We can also use the posterior distribution from the old data as our prior for the new data. The posterior distribution given the old data is  $\text{Gamma}(a = 466.44, b = 4.44)$ . The new data is a count of 200 over two weeks, therefore the likelihood is proportional to:

$$\exp(-2\theta)(2\theta)^{200}$$

Multiplying the prior (the above posterior) with the likelihood, we get:

$$\begin{aligned}
 [\theta^{466.44-1} \exp(-4.44\theta)][\exp(-2\theta)(2\theta)^{200}] &= \theta^{666.44-1} \exp(-4.44\theta - 2\theta) \\
 &= \theta^{666.44-1} \exp(-6.44\theta)
 \end{aligned} \tag{6.53}$$

In other words, the posterior is  $\text{Gamma}(666.44, 6.44)$ .

This is identical to the posterior we obtained above by combining the data. This situation will always be true when we have conjugate priors: **the result will be the same regardless of whether we take all the data together or successively fit new data, using the posterior of the previous fit as our prior.** This is because the multiplication of the terms will always give the same result regardless of how they are rearranged.

In the above example, we can again express the posterior mean as a weighted sum of the prior mean and the maximum likelihood estimate of  $\theta$ .

The posterior mean is:

$$\frac{a^*}{b^*} = \frac{a + \sum x_i}{n + b} \tag{6.54}$$

This can be rewritten as

$$\frac{a^*}{b^*} = \frac{a + n\bar{x}}{n + b} \tag{6.55}$$

Dividing both the numerator and denominator by b:

$$\frac{a^*}{b^*} = \frac{(a + n\bar{x})/b}{(n + b)/b} = \frac{a/b + n\bar{x}/b}{1 + n/b} \quad (6.56)$$

Since  $a/b$  is the mean  $m$  of the prior, we can rewrite this as:

$$\frac{a/b + n\bar{x}/b}{1 + n/b} = \frac{m + \frac{n}{b}\bar{x}}{1 + \frac{n}{b}} \quad (6.57)$$

We can rewrite this as:

$$\frac{m + \frac{n}{b}\bar{x}}{1 + \frac{n}{b}} = \frac{m \times 1}{1 + \frac{n}{b}} + \frac{\frac{n}{b}\bar{x}}{1 + \frac{n}{b}} \quad (6.58)$$

This is a weighted average: setting  $w_1 = 1$  and  $w_2 = \frac{n}{b}$ , we can write the above as:

$$m \frac{w_1}{w_1 + w_2} + \bar{x} \frac{w_2}{w_1 + w_2} \quad (6.59)$$

As  $n$  approaches infinity, the weight on the prior mean  $m$  will tend towards 0, making the posterior mean approach the maximum likelihood estimate of the sample.

This makes sense: as sample size increases, the likelihood will dominate in determining the posterior mean.

Regarding variance, since the variance of the posterior is:

$$\frac{a^*}{b^{*2}} = \frac{(a + n\bar{x})}{(n + b)^2} \quad (6.60)$$

as  $n$  approaches infinity, the posterior variance will approach zero, which makes sense: more data will reduce variance (uncertainty).

This ends our introduction to Bayesian data analysis. We turn next to our main topic: linear modeling in a Bayesian framework.

### 6.3 Fitting Linear Models and Linear Mixed Models in a Bayesian setting

I will add notes here later, but for next week, please read the ArXiv preprint by Sorensen and Vasishth<sup>9</sup>.

<sup>9</sup> Tanner Sorensen, Sven Hohenstein, and Shravan Vasishth. Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. ArXiv e-print, 2015



## 7

## Bibliography

- [1] Douglas Bates, Reinhold Kliegl, Shravan Vasishth, and Harald Baayen. Parsimonious mixed models. ArXiv e-print; submitted to *Journal of Memory and Language*, 2015.
- [2] George E.P. Box and David R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252, 1964.
- [3] Annette J Dobson and Adrian Barnett. *An introduction to generalized linear models*. CRC press, 2011.
- [4] A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, Cambridge, UK, 2007.
- [5] Andrew Gelman and John Carlin. Beyond power calculations assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6):641–651, 2014.
- [6] C.M. Grinstead and J.L. Snell. *Introduction to probability*. American Mathematical Society, 1997.
- [7] G. Jay Kerns. *Introduction to Probability and Statistics Using R*. 2010.
- [8] André I Khuri. *Advanced calculus with applications in statistics*, volume 486. Wiley, 2003.
- [9] David Lunn, Chris Jackson, David J Spiegelhalter, Nicky Best, and Andrew Thomas. *The BUGS book: A practical introduction to Bayesian analysis*, volume 98. CRC Press, 2012.
- [10] I. Miller and M. Miller. *John E. Freund's Mathematical Statistics with Applications*. Prentice Hall, 2004.
- [11] Martyn Plummer. JAGS version 3.3.0 manual. *International Agency for Research on Cancer. Lyon, France*, 2012.

- [12] Sheldon Ross. *A first course in probability*. Pearson Education, 2002.
- [13] Tanner Sorensen, Sven Hohenstein, and Shravan Vasishth. Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. ArXiv e-print, 2015.
- [14] Shravan Vasishth and Bruno Nicenboim. Statistical methods for linguistic research: Foundational ideas – Part I. 2016.