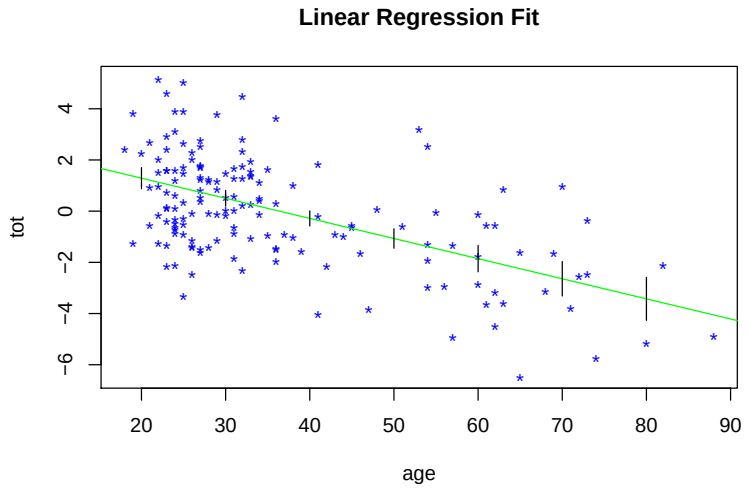
Assignment #2

Lindsay Brock, Gustavo Esparza, and Brian Schetzsle

September 12, 2019

Computer Age Statistical Inference, by Efron and Hastie

Figure 1.1



Above is the replication of Figure 1.1. The green line is a linear regression fit and the black bars represent ± 2 standard errors at each increment of *age*.

Figure 1.2

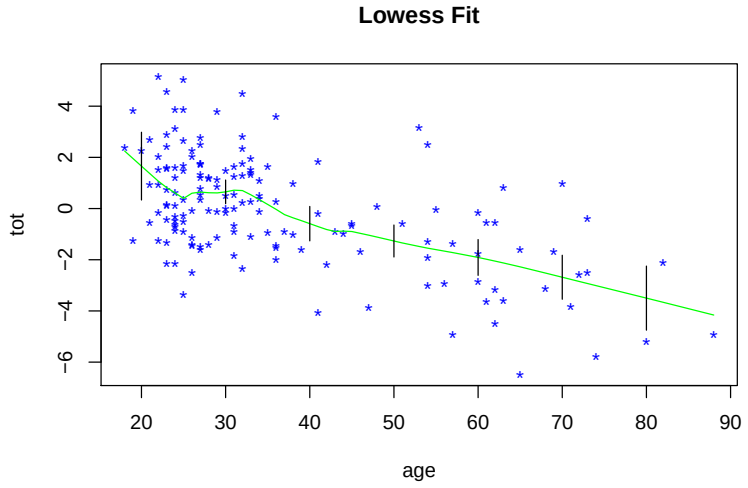
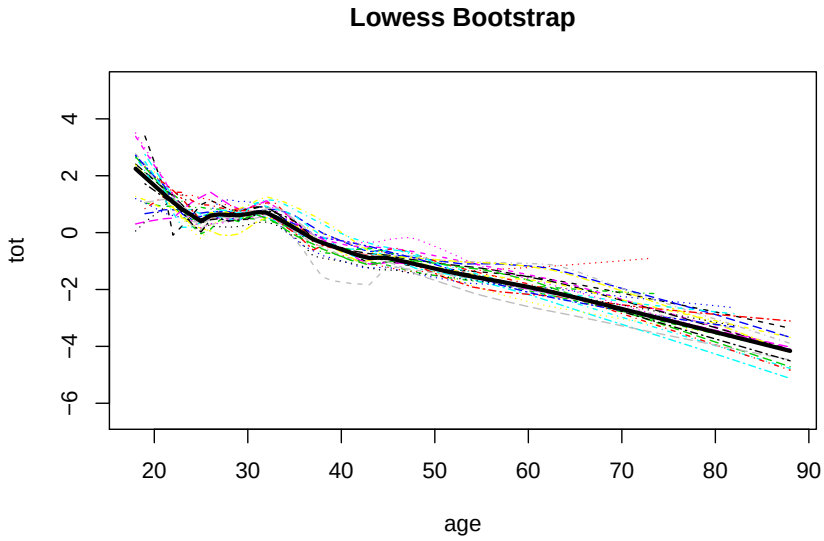


Figure 1.2 depicts the local polynomial lowess fit. Each black bar represents ± 2 bootstrap standard deviations.

Figure 1.3



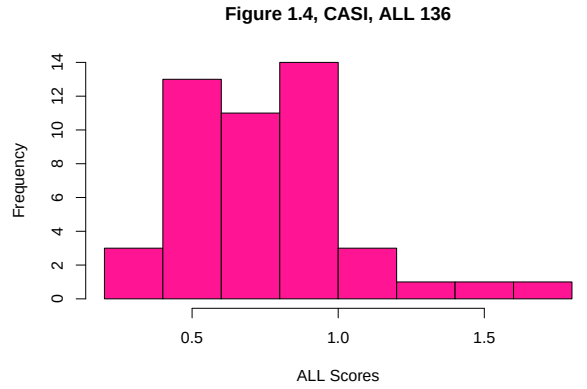
Above is the graphical representation of 25 bootstrapped lowess models.

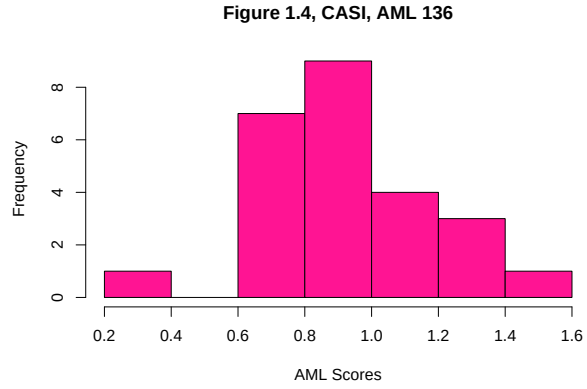
Table 1.1

The table below summarizes the information from the plots above. The first two lines are the linear regression estimates and standard errors from Figure 1.1 and the last two lines are the lowess estimates and their bootstrapped errors from Figure 1.2

	20	30	40	50	60	70	80
Linear Regression	1.29	0.50	-0.28	-1.07	-1.86	-2.64	-3.43
Std Error	0.21	0.15	0.15	0.19	0.26	0.34	0.42
Lowess	1.66	0.65	-0.59	-1.27	-1.91	-2.68	-3.50
Bootstrap Std Error	0.66	0.23	0.34	0.31	0.35	0.43	0.63

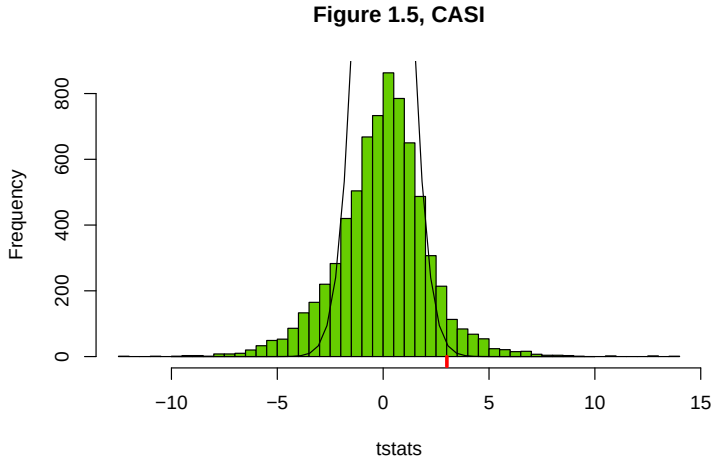
Figure 1.4 (Extra Optional)





Above are the plots of the ALL and AML scores for gene 36 from the leukemia data set.

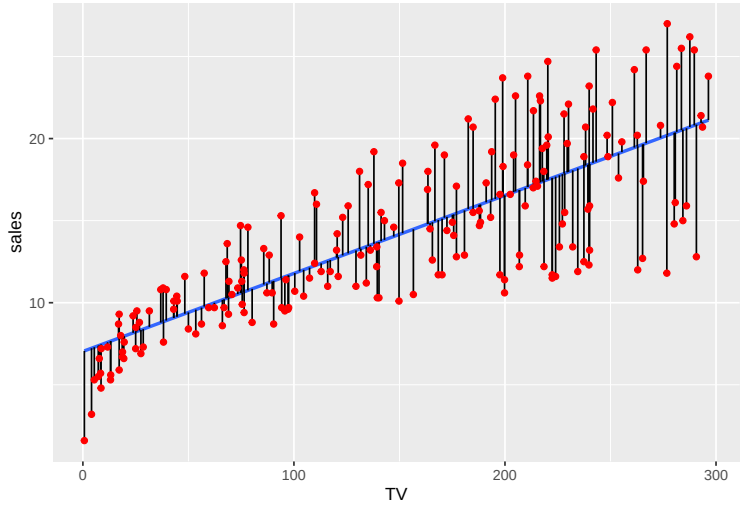
Figure 1.5 (Optional)



Above are the two-sample t-statistics for *genes*, in the leukemia data set. The black line is the theoretical null density for the t-statistic.

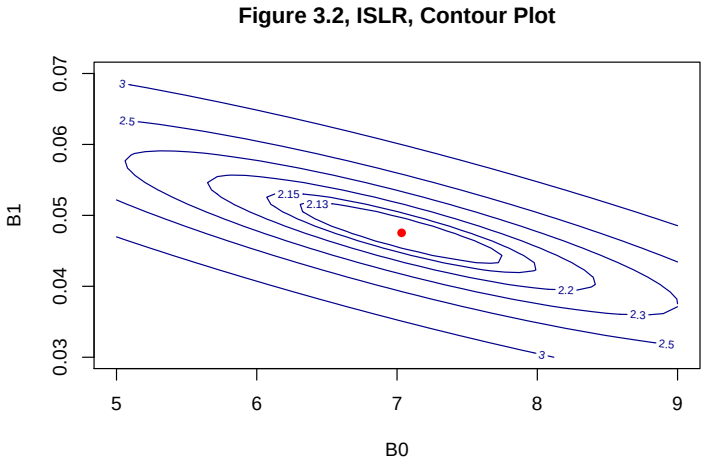
An Introduction to Statistical Learning, by James et al.

Figure 3.1



Above is a plot of the linear model with error bars added to each data point.

Figure 3.2

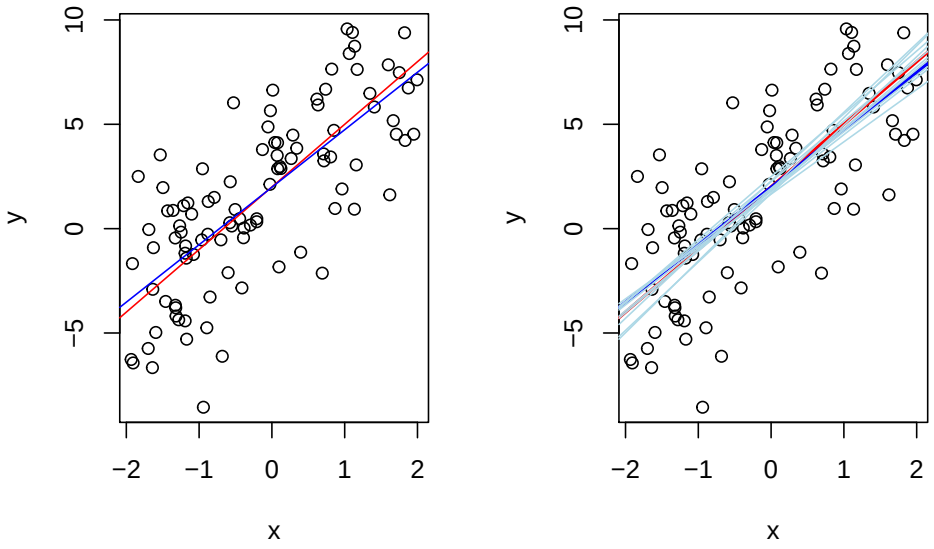


Above is a contour plot of the RSS on the data where the red dot is the least squares estimate of $\hat{\beta}_0$ and $\hat{\beta}_1$.

Figure 3.3

For our final figure, we will generate 100 random points and apply them to the linear formula $y = 2+3x+\epsilon$, where $\epsilon \sim n(0, 3)$. These are the true y values, which we will use to create a linear regression model $\hat{y} = \beta_0+\beta_1 x$. We will display our regression line alongside the line $y = 2 + 3x$ in order to compare their fits.

Finally, we will create ten new regression lines by recreating our original data points and fitting corresponding models. This plot is displayed to the right:



Bayesian Chapter

Section 4.1: Introduction

This chapter begins with a basic overview of the differences between Frequentist Statistics and Bayesian Statistics. Both methods are used for the general purpose of answering scientific questions, but their underlying difference relates to how much **domain knowledge** is incorporated in their respective analysis. Here, domain knowledge refers to prior information that is typically based on previous empirical data.

In a Frequentist framework, there is an attempt to minimize the usage of prior information when performing analysis. This does not imply, however, that no prior information is used whatsoever. As is explained, the inclusion of a set of variables in any statistical model is in itself an example of prior knowledge. Since we tend to choose variables and classes of models based on an assumption of how well they will fit our data, that is an example of using empirical data for analysis.

Bayesian methods, on the other hand, are based on incorporating prior knowledge in the process of making inference. Bayesian Statisticians utilize prior information with the expectation of other scientists having a larger perspective into the specifications and assumptions of provided models. Bayesian analysts also believe incorporating prior information with real life data can often lead to a more accurate inference.

Section 4.2: Bayesian Inference

This section is primarily interested in explaining the general principles and structures of Bayesian Inference. As previously stated, Bayesian Inference is essentially about making conclusions about unknown quantities in terms of probabilities, using observed data and prior knowledge. In this context, *prior* refers to the extent of our knowledge and uncertainty regarding the unobserved data. This prior information is then utilized via probability models.

Given our combination of observed data and prior information, Bayesian inference begins by defining a joint probability for prior data and the assumed distribution from which the data was generated. Once the joint distribution has been defined, we are then able to update our knowledge about the unobserved data by using our observed data. Here, the updated data is referred to as *posterior* data.

It is now apparent that the core basis of Bayesian Statistics relies on probabilities. More specifically, Bayesian practice refers to probability as a measure of uncertainty. Thus, Bayesian probability models are constructed for random variables that change in value and fixed parameters for which we are uncertain of their exact values. Going along this path of methodology, we can note that Bayesians tend to assign probabilities to events that are not repeatable. This is, once again, an other example of the difference between Frequentists and Bayesians. Returning to the probabilities assigned to our data (y) and the true parameters (θ), Bayesians begin by specifying the extent of knowledge and uncertainty about the potential values of θ using prior probabilities. This probability is denoted as $P(\theta)$.

A useful example of prior probabilities in real life data is a given study of body temperatures. If the assumed average body temperature is 98.6, then we can define the prior distribution of the actual body temperature (θ) as $\theta \sim N(98.6, \tau^2)$, Here, τ^2 refers to the amount of certainty we have regarding the actual body temperature being close to 98.6. If we believe that the true body temperature will be very close to 98.6, then we can reduce the size of τ. General advisory for priors is to keep an open mind and avoid restricting distributions (small variance).

This section concludes by referring to prior probabilities that are constructed using the scientific history of the data, as opposed to the empirical data that is observed.

Section 4.3: Exchangeability Instead of iid Assumption

This section will examine another difference between Frequentist and Bayesian Statistics: the assumptions we make about observed data. In Frequentist work, the general assumption is that the data is independent and identically distributed (iid). This is a very convenient assumption to make when constructed models and

making inference, but it may be too strong of an assumption. As an alternative, Bayesians use an assumption of symmetry in the data.

When considering a joint distribution for a set of observations, $y = (y_1, \dots y_n)$, we may assume that the indices are uninformative. If a given set of observations proves to have uninformative indices (invariant) then the sequence $(y_1, \dots, y_n)$ is referred to as *exchangeable*. An excellent example for illustrating exchangeability is a series of three coin tosses. If we are concerned with the number of heads that appear, then there is virtually no difference between the probabilities $P(1, 0, 0)$, $P(0, 1, 0)$ and $P(0, 0, 1)$. The appearance of event 1 does not have an effect on our probability; these coin tosses are *exchangeable*. Thus, we have seen that an exchangeable observation set has a joint probability distribution that remains the same for all possible permutations of them.

When discussing exchangeability, it is important to discuss observation sets that are *not* exchangeable. For example, consider a set of observations being defined by students with unique ID numbers. If these numbers are distributed randomly to students, then we need not worry about their permutations having different impacts on the joint distribution. However, if there is a grouping to how ID numbers are distributed (first digit determines class level or Major), then we can only assume exchangeability within the groups and not the entire set of individuals. This is referred to as *partial exchangeability*.

We have elaborated quite extensively on the definition and representation of exchangeability, but now we will focus on what the assumption allows us to derive for the probability distribution. Here are two primary derivations from the assumption of exchangeability:

1. The conditional distribution of y given θ is

$$P(y|\theta) = P(y_1, y_2, \dots, y_n|\theta) = \prod_{i=1}^{n} p(y_i|\theta)$$

2. There exists a prior probability distribution $P(\theta)$ over the parameters of the model such that we can find the unconditional(marginal) distribution of observations as follows:

$$P(y) = P(y_1, y_2, \dots, y_n) = \int_\Omega \prod_{i=1}^{n} P(y_i|\theta)P(\theta)d\theta$$

These two equations imply that there is a specific distribution for which can extract the exchangeable observations. From these two formulas, we can also see that we have set a path for obtaining the model for the observed data ($P(y|\theta)$) and the Prior model for the parameters of the model ($P(\theta)$). For these two distributions, we require a conclusion to be made about the unobserved θ conditionally on the observed y. This is referred to as the *prior distribution*, $P(\theta|y)$.

Section 4.4: Bayesian Inference Utilizes Bayes Theorem

The previous three sections have introduced the underlying aspects for Bayesian analysis, and now we will focus on performing actual analysis in the context of scientific studies. When closing section 4.3, we were left with the goal of obtaining the *prior distribution*, $P(\theta|y)$. In the previous sections, it was also mentioned that *Bayes' Theorem* was used to update our data. For events A and B, we can provide a basic definition of Bayes' Theorem:

$$P(B|A) = \frac{P(B)P(A|B)}{P(A)}$$

Now, we can use our Bayesian definition to provide a formula for obtaining the prior distribution, $P(\theta|y)$:

$$P(\theta|y) = \frac{P(\theta)P(y|\theta)}{P(y)}$$

Considering only the functions that depend on θ, we can simplify our prior distribution via the following form:

$$P(\theta|y) \propto P(\theta)P(y|\theta)$$

This last equation for the posterior distribution is the essence of Bayesian analysis.It is used for many utilities: expressing updated knowledge about θ, making future predictions, and making conclusions for hypothesis tests.

Starting with our next section, we will focus on specific models (Binomial,Exponential,Poisson,etc.) in the context of Bayesian analysis.

Section 4.5: Binomial Model

In order to introduce the Binomial model, let us begin by exploring a single event with two outcomes.Whatever the event may be, flipping a coin or any other win/loss situation, we can define the event space as $x \in 0, 1$. If 0 denotes a loss and 1 denotes a win with probability θ, then x is assumed to follow a Bernoulli distribution defined as:

$$P(x|\theta) = \theta^x(1-\theta)^{1-x}$$

Now considering n such observations from our event defined by $x_1, x_2, \ldots, x_n$, we can define $y = \sum_{i=1}^{n} x_i$ to be the total number of successes in our trials. In this case, y is defined as having a *Binomial* distribution with parameters n and θ:

$$p(y|n,\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

If $P(\theta)$ is again defined as the prior distribution and $P(y)$ is the marginal distribution found by integration of the joint distribution, then the desired posterior distribution is defined as:

$$P(\theta|y) = \frac{\binom{n}{y}\theta^y(1-\theta)^{n-y}P(\theta)}{\int_0^1 \binom{n}{y}\theta^y(1-\theta)^{n-y}P(\theta)d\theta}$$

At this point, it is important to note that the distribution of θ is still unknown. In Bayesian statistics, we will proceed to make an uninformed assumption about θ. In this particular case, we will propose that θ is uniformly distributed in $[0, 1]$. This assumption is saying that the probability of a success is equally likely to be any of the possible probabilities between 0 and 1 (we have not proposed any bias for θ). In turn, the distribution of θ is simply a constant of 1. Now, using this assumption, we have the following simplified marginal distribution:

$$P(y) = \int_0^1 \binom{n}{y}\theta^y(1-\theta)^{n-y} \times 1 d\theta = \frac{1}{n+1}$$

This evaluation was made by using the pdf of the *beta* distribution. Now that we have resolved any issues of undefined distributions, we can finally define our Posterior distribution as follows:

$$P(\theta|y) = \frac{(n+1)!}{y!(n-y)!}\theta^y(1-y)^{n-y} \sim Beta(y+1, n-y+1)$$

Now that we have a closed form for the posterior distribution of our binomial model, we can turn our attention to making a future prediction for y_{n+1}, denotes as $\tilde{y}$. This is done via the *posterior prediction distribution*:

$$P(\tilde{y}|y) = \int_\theta P(\tilde{y}|\theta)P(\theta|y)d\theta$$

In our binomial model, the posterior prediction distribution is simply the beta distribution and the integration over the probability space is its expected value:

$$P(\tilde{y} = 1|y) = E(Beta(y+1), n-y+1) = \frac{y+1}{n+2}$$

Thus, we found a posterior equation for the probability of a success of trial $n+1$. The text provides an application of this predictive distribution involving an election with a similar binomial distribution.

A final note for this section involves the choice of a prior distribution $\sim unif[0,1]$. We can note that this is equivalent to the $beta(1,1)$ distribution. Thus, our prior and posterior distributions are both Beta distributions. This occurrence is called *Conjugacy* and the prior is referred to as the "conjugate" prior.

Section 4.6: Exponential Family and Conjugate Priors

$$P(y_i|\theta) = h(y_i)g(\theta)exp(\phi(\theta)^T s(y_i))$$

This section begins with a discussion of a large class of distributions called the *exponential family* which follow the form above where $\phi(\theta)$ is the "natural parameter". Their joint distribution given θ and independent ys is below. The exponential family includes distributions like the Gaussian, Poisson, and Binomial and can be identified when its mathematical form can be separated into parts that are only dependent on the data, only dependent on the parameters, and utilize data and parameters together. An important property of the part that is dependent on the data alone is *sufficiency*. $t(y) = \sum_i s(y_i)$ is a sufficient statistic for θ as θ depends on the data y only through t.

$$P(y|\theta) = \left[\prod_{i=1}^{n} h(y_i)\right] g(\theta)^n exp(\phi(\theta)^T \sum_{i=1}^{n} s(y_i))$$

The section goes on to discuss conjugate priors and describes its meaning with an example. If the prior follows the form:

$$P(\theta) \propto g(\theta)^\eta exp(\phi(\theta)^T v)$$

And the posterior follows a similar form:

$$P(\theta|y) \propto g(\theta)^{\eta+n} exp(\phi(\theta)^T (v + t(y)))$$

Then $P(\theta)$ would be considered a conjugate prior as the forms are the same. This idea is reinforced with an example of the binomial model with the conjugate prior $P(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$ which is actually the Beta(α, β) distribution. The posterior distribution also happens to be Beta but with parameters $\alpha + y$ and $\beta + n - y$.

The next example provides R code to plot the prior and posterior. An election will take place and a survey was given to 10 people with 3 of them voting for candidate A who won in a previous election with 55% of the votes. A Beta(5.5, 4.5) is used to account for uncertainty and it is advised to plot the prior distribution or sample from it to ensure it accurately represents the prior knowledge (percentage of votes won in the previous election), shown in the code below.

```
#Code to Plot Prior
theta = seq(0, 1, 0.01)
prior = dbeta(theta, 5.5, 4.5)
plot(theta, prior, type = 'l')

#Code to Plot Posterior
theta = seq(0, 1, 0.01)
post = dbeta(theta, 8.5, 11.5)
plot(theta, post, type = 'l')
```

A continuation of this example goes on to show how to update knowledge based on new information while not ignoring previous data or starting from scratch. Adding 20 more people to the survey, it is shown that 12/20 will vote for candidate A. Using the previous posterior as the new prior a new posterior is calculated that now includes the additional information, Beta(20.5, 19.5). The posterior expectation and MLE become closer in value as the amount of data increases.

Section 4.7: Poission Model

The next model discussed from the *exponential family* is the Poisson which is commonly used for count data. Finding the Likelihood:

$$\begin{aligned} P(y|\theta) &= \prod_{i=1}^{n} \frac{\theta^{y_i} exp(-\theta)}{y_i!} \\ &\propto exp(-n\theta) exp(log\theta \sum y_i) \end{aligned}$$

And the conjugate prior:

$$\begin{aligned} P(\theta) &\propto (exp(-\theta))^{\eta} exp(vlog\theta) \\ &\propto exp(-\eta\theta)\theta^{v} \\ &\propto exp(-\beta\theta)\theta^{\alpha-1} \end{aligned}$$

Using the above Gamma(α, β) as the prior, the posterior is:

$$\theta|y \sim Gamma(\alpha + \sum_{i=1}^{n} y_i, \beta + n)$$

To reinforce the previous information, the section continues with another example. The goal is to predict the number of goals Beckham would score in the remaining games, after already scoring one goal in the first two games. y_i is the number of goals he scores and is modeled using Poisson with parameter θ. The maximum likelihood is $\hat{\theta} = 0.5$ and Gamma(α, β) is used as the prior for θ. Using his history in Real Madrid we can select an appropriate Gamma mean of 0.14, but if that information was not available another prior would need to be selected to reflect this lack of information. The code is as follows.

```
theta = seq(0, 1.5, 0.01)
prior = dgamma(theta, shape = 1.4, rate = 10)
plot(theta, prior, type = 'l')
```

As Gamma is a conjugate prior for the rate parameter of the Poisson model, the posterior also has a Gamma distribution but with the parameters $(1.4 + 1, 10 + 2)$. The expected number of goals is found to be $2.4/12 = 0.2$. The code is as follows.

```
theta = seq(0, 1.5, 0.01)
post = dgamma(theta, shape = 2.4, rate = 12)
plot(theta, post, type = 'l')
```

In the above example there is such a small amount of information that the posterior ends up being more similar to the prior than the likelihood. If the amount of data were to increase, the influence of the prior on the posterior decreases and the effect of likelihood increases.

Section 4.8: Univariate Normal Model

Section 4.8 concerns the Bayesian treatment of the univariate normal distribution. With a basic example of a single observation and known variance, we find that the general form of the conjugate prior is

$$P(\mu) \propto \exp(a\mu^2 + b\mu)$$

which can be parameterized as

$$P(\mu) \propto \exp(-\frac{1}{2\tau_0^2}(\mu - \mu_0)^2)$$

which is a $N(\mu_0, \tau_0^2)$ distribution and thus the posterior distribution takes the form

$$P(\mu|\sigma, y) \propto \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2 - \frac{1}{2\tau_0^2}(\mu-\mu_0)^2\right)$$

Completing the square in the posterior reveals that it is actually a $N(\mu_1, \tau_1^2)$ with

$$\mu_1 = \frac{\frac{\mu_0}{\tau_0^2} + \frac{y}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}} \qquad \text{and} \qquad \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

Generalizing to n observations, we get the same form but with

$$\mu_n = \frac{\frac{\mu_0}{\tau_0^2} + \frac{\bar{y}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \qquad \text{and} \qquad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

The section proceeds with an example of obtaining a posterior distribution of the height of students which is assumed normal with known variance. If the mean is fixed and the variance is unknown, then an inverse gamma distribution is used as the conjugate prior on the distribution of σ^2. The example with student heights is repeated, this time finding the posterior distribution of σ^2 given three measurements.

When both the mean and the variance are unknown, there is no conjugate prior. The normal distribution for the mean and the inverse-gamma for the variance are known as conditionally conjugate priors because they are only conjugate when conditioned on the other parameter. It is still possible to sample from the posterior distribution when the mean and variance are unknown, though. This can be done using Gibbs sampling, which iteratively conditions on each parameter.

Section 4.9: Multinomial Model

Section 4.9 concerns the multinomial distribution. The conjugate prior is the Dirichlet distribution, which has the form

$$P(\theta|\alpha) \propto \prod_{j=1}^{J} \theta_j^{\alpha_j - 1}$$

for J categories. The posterior, then, is also Dirichlet and has the form

$$P(\theta|\alpha, y) \propto \prod_{j=1}^{J} \theta_j^{y_j + \alpha_j - 1}$$

The multinomial distribution is an extension of the binomial distribution just as the Dirichlet distribution is an extension of the binomial distribution's conjugate prior, the beta distribution. The section concludes with an example of how to obtain a posterior for θs in an election setting.

Section 4.10: Multivariate Normal Model

Section 4.10 concerns the Multivariate Normal distribution. If μ is known, then the conjugate prior for the covariance matrix Σ is the inverse Wishart distribution Inv-Wishart(ν_n, Λ_n). The posterior distribution of Σ is thus also an inverse Wishart Inv-Wishart(ν_n, Λ_n) with $\nu_n = \nu_0 + n$ and $\Lambda_n = \Lambda_0 + \sum_i (Y_i - \mu)(Y_i - \mu)^T$. Similar to the univariate case, when both μ and Σ are unknown, the posterior can still be sampled from using Gibbs sampling.

Section 4.11: Further Reading on Bayesian Statistics

Elementary texts with examples:

- Bolstad and Curran (2016)
- Hoff (2009)
- Congdon (2003)
- Lee (2012)

Texts with more advanced perspecitives, theory, and applications:

- Gelman et al. (2013)
- Robert (2007)
- Kadane (2011)
- Carlin and Louis (2008)
- Box and Tiao (1992)

Texts for a foundation and theory through mathematics:

- Schervish (2011)

Texts for a philisophical foundation and evolution:

- Savage (1954)
- Jeffreys (1961)
- deFinetti (1974)

APPENDIX

Computer Age Statistical Inference, by Efron and Hastie

```
kidney = read.delim("kidney.txt", header = T, sep = "")
attach(kidney)

#Setting Up
#for Fig 1.1
model_kid = lm(tot~age, data = kidney)
predictions = predict.lm(model_kid,
                         newdata = data.frame(age = seq(20, 80, length.out = 7)),
                         se.fit = T)
estimations = predictions$fit
se = predictions$se.fit
upper = estimations + 2*se; lower = estimations - 2*se
points = seq(20, 80, length.out = 7)

#for Fig 1.2
low = lowess(age, tot, 1/3)
low_pred = approx(low$x, low$y, xout = seq(20, 80, length.out = 7), ties = mean)$y

#for Fig 1.3
low_x = matrix(0, 157, 250)
low_y = low_x
boot_pred = matrix(0,250,7)
for(i in 1:250){
  index = sample(1:157, 157, replace = T)
  new = kidney[index,]
  low_x[,i] = lowess(new[,1], new[,2], 1/3)$x
  low_y[,i] = lowess(new[,1], new[,2], 1/3)$y
  boot_pred[i,] = approx(low_x[,i], low_y[,i],
                         xout = seq(20, 80, length.out = 7), ties = mean)$y
}

#for Fig 1.2
boot_pred = na.omit(boot_pred)
se_boot = apply(boot_pred, 2, sd)
boot_lower = low_pred - 2*se_boot
boot_upper = low_pred + 2*se_boot
```

Figure 1.1

```
plot(age, tot, pch = "*", col = 'blue', main = "Linear Regression Fit")
abline(model_kid, col = 'green')
segments(points, lower, points, upper)
```

Figure 1.2

```
plot(age, tot, pch = "*", col = 'blue', main = "Lowess Fit")
points(low, type = 'l', col = 'green')
segments(points, boot_lower, points, boot_upper)
```

Figure 1.3

```
plot(age, tot, pch = "*", col = 'white', main = "Lowess Bootstrap")
color = c(1:25); line = rep(2:6, 5)
for(i in 1:25){
  points(low_x[,i], low_y[,i], type = 'l', lty = line[i], col = color[i])
}
points(low, type = 'l', col = 'black', lwd = 3)
```

Table 1.1

```
library(kableExtra)

table = round(rbind(estimations,se,low_pred,se_boot),2)
colnames(table) = c("20","30","40","50","60","70","80")
rownames(table) =c("Linear Regression","Std Error","Lowess","Bootstrap Std Error")

table %>%
kable() %>%
kable_styling()
```

Figure 1.4 (Extra Optional)

```
#Now try to recreate figure 1.5
temp = read.csv("leukemia_big.csv", header=FALSE, sep=",")

category = data.frame("category"=sapply(temp[1,],as.character))

p = dim(temp)[1]-1
n = dim(temp)[2]
data = matrix(NA,nrow=n,ncol=p)
for(i in 1:n){
  data[i,] = as.numeric(as.character(temp[2:(p+1),i]))
}
rm(temp)

ALL = data[category=="ALL",]
AML = data[category=="AML",]
n.all = dim(ALL)[1]
n.aml = dim(AML)[1]

hist(ALL[,136], col="deeppink", main="Figure 1.4, CASI, ALL 136",
     xlab = c("ALL Scores"))
hist(AML[,136], col="deeppink", main="Figure 1.4, CASI, AML 136",
     xlab = c("AML Scores"))
```

Figure 1.5 (Optional)

```
tstats = rep(NA,p)
for(i in 1:p){
  mean.aml = mean(AML[,i])
  mean.all = mean(ALL[,i])
```

```
  s2.aml = var(AML[,i])
  s2.all = var(ALL[,i])
  tstats[i] = (mean.aml-mean.all)/sqrt(((n.aml-1)*s2.aml + (n.all-1)*s2.all)/
                                        (n.aml+n.all-2))/sqrt(1/n.aml+1/n.all)
}

#dev.off()
hist(tstats, breaks=90, col="chartreuse3", probability=FALSE,
     main="Figure 1.5, CASI")
xseq = seq(-10,10,length.out=50)
lines(xseq, dt(xseq, 70)*p)
lines(c(3.01,3.01),c(-30,0),col="red",lwd=3)
```

An Introduction to Statistical Learning, by James et al.

```
advertising = read.csv("advertising.csv")
sales = advertising$sales
TV = advertising$TV

model = lm(sales~TV)
predicted = predict(model)
```

Figure 3.1

```
library(ggplot2)
ggplot(advertising, aes(x = TV, y = sales)) +
  geom_smooth(method='lm',se=F) +
  geom_segment(aes(xend = TV, yend = predicted)) +
  geom_point(color="red")
```

Figure 3.2

```
B0 = seq(5,9,length.out=40)
B1 = seq(0.03,0.07,length.out=40)

z_fun = function(B0,B1) {
  return(sum((sales - B0 - TV*B1)^2)/1000)
}

RSS = matrix(NA,ncol=40,nrow=40)
for(i in 1:40){
  for(j in 1:40){
    RSS[i,j] = z_fun(B0[i],B1[j])
  }
}

b = model$coefficients[1]
m = model$coefficients[2]
contour(B0,B1,RSS,nlevels=6,levels=c(2.13,2.15,2.2,2.3,2.5,3),xlab="B0",
        ylab="B1",col="darkblue", main="Figure 3.2, ISLR, Contour Plot")
points(b,m,pch=16,col="red")
```

Figure 3.3

```
x = runif(100,-2,2)
e = rnorm(100,0,3)
y = 2 + 3*x +e
model = lm(y~x)

par(mfrow=c(1,2))
plot(x,y)
abline(2,3,col="red")
abline(model,col="blue")

plot(x,y)
abline(2,3,col="red",lwd=2)
abline(model,col="blue",lwd=2)

for(i in 1:10){
x = runif(100,-2,2)
e = rnorm(100,0,3)
y = 2 + 3*x +e
abline(lm(y~x),col="light blue")
}
```