

## 2

# Introduction to Data Analysis

In this chapter we provide a taste of the types of problems and data that Bayesian methods can address. We start off with simple probability models including normal, binomial, and Poisson data and work up to somewhat more elaborate models including regression and nonlinear time series models and models that incorporate random effects. In some of these examples we have prior information that we incorporate into the model, in others we use vague priors that approximate the non-Bayesian approach.

### 2.1 Brass alloy zinc content: Normal data

A corrosion resistant brass alloy, widely used in plumbing fixtures, is composed mainly of copper, tin, lead, zinc, nickel, and iron in decreasing amounts. The addition of zinc to the alloy produces a “jump” in metal strength when added within a tolerance range between 4% and 6%. As zinc has a lower melting point than copper, a certain amount of zinc is dissipated by the end of the heating process and it is common practice to measure zinc content through spectrometry readings before pouring the metal into molds. If the amount of zinc is less than, say, 4.4% a zinc addition is made to the alloy to correct the percentage.

Twelve alloy samples were tested using spectrometry by the St. Paul Brass and Aluminum Foundry in St. Paul, Minnesota in June of 2003. Let  $Y_i$  be the zinc percentage of sample  $i$ , where  $i = 1, \dots, 12$ . We assume the normal model  $Y_i | \mu, \sigma^2 \stackrel{iid}{\sim} N(\mu, \sigma^2)$ . The Vice-President of Operations,

Kay Stinson, estimates with 95% certainty that the mean percentage before pouring should be between 4.5% and 5%, and centered at 4.75%. We therefore require  $E(\mu) = 4.75$  and  $P(4.5 < \mu < 5) = 0.95$ . The specification  $\mu \sim N(4.75, 0.0163)$  satisfies these requirements. We have no good information on the standard deviation  $\sigma$ , however, so we specify a sensibly vague prior  $\sigma^{-2} \sim \Gamma(0.001, 0.001)$  (see Sec. 4.6). The sample zinc percentages were 4.20, 4.36, 4.11, 3.96, 5.63, 4.50, 5.64, 4.38, 4.45, 3.67, 5.26, and 4.66. A histogram and normal quantile plot (Sec. 2.4, Christensen, 1996) show no serious deviations from normality.

qqq

$\sigma = \int \Phi\{(4.4 - \mu)/\sigma\} p(\mu, \sigma | Y_1, \dots, Y_{12}) d\mu d\sigma$  For the parameters  $\mu$  and  $\sigma$  we obtain the estimated posterior median and equal-tailed 95% credible intervals from computer simulations of  $\hat{\mu} = 4.69\%$  (4.49%, 4.90%) and  $\hat{\sigma} = 0.64\%$  (0.44%, 1.03%). That is, for example,  $Pr(\mu \leq 4.69 | Y_1, \dots, Y_{12}) = 0.5$  and  $Pr(4.49 < \mu < 4.90 | Y_1, \dots, Y_{12}) = 0.95$ . It is of additional interest to find the probability that zinc will be added to the metal during melting. Let  $Y_{13} | \mu, \sigma \sim N(\mu, \sigma^2)$  be independent of  $Y_1, \dots, Y_{12}$  given  $\mu$  and  $\sigma$ . Then we are interested in the parameter  $Pr(Y_{13} \leq 4.4)$  and wish to obtain  $Pr(Y_{13} \leq 4.4 | Y_1, \dots, Y_{12})$ . We estimate that the probability of adding zinc is 0.33 with a 95% credible interval of (0.20, 0.45).

## 2.2 Probability of a defective: Binomial data

The Par-Aide Corporation in Lino Lakes, Minnesota, makes ball washers for golf courses. St. Paul Brass and Aluminum Foundry makes a part called a “push rod eye,” an integral component of the golf ball washer. Out of 2,430 push rod eye’s poured over two days in May, 2003, only 2,211 actually shipped. It is of interest to estimate the probability of pouring a defective part. The Vice President of Operations thinks that for this particular part a plausible range for the proportion of scrap is 5% to 15%.

We assume that the number scrapped  $Y$  is binomial with parameter  $\pi$ :  $Y | \pi \sim \text{binomial}(2430, \pi)$ . To reflect the Vice-President’s knowledge we decided to place a distribution on  $\pi$  such that  $P(\pi \leq 0.05) = 0.025$  and  $P(\pi \geq 0.15) = 0.025$ . The prior  $\pi \sim \text{beta}(12.05, 116.06)$  satisfies these constraints.

With  $Y = 219$ , computer simulations give the estimated posterior median of  $\hat{\pi} = 0.09$  and a 95% equal-tailed credible interval of (0.08, 0.10). The posterior probability that the proportion of scrap is over 10% is 0.025. In this model it is possible to obtain the posterior distribution directly. Note that, in general, if  $Y | \pi \sim \text{binomial}(n, \pi)$  and  $\pi \sim \text{beta}(a, b)$ , then  $\pi | Y \sim \text{beta}(a + Y, b + n - Y)$ , from Table 1.3. The median of a  $\text{beta}(12.05 + 219, 116.06 + 2430 - 219)$  distribution is 0.0902, the same as the estimate obtained from computer simulations up to two decimal places. The exact credible interval is (0.0795, 0.0998). (0.0795, 0.0998).

### 2.3 Abortion in dairy cattle: Survival data

Bedrick, Christensen, and Johnson (2000) consider data on 45 cows that naturally aborted their fetuses prematurely. It is of interest to dairy managers to determine whether cows infected with *neospora caninum* typically abort later than unaffected cows; 19 of the 45 cows were infected. The times to abortion in the uninfected group are 60, 74, 37, 45, 75, 40, 50, 50, 146, 70, 50, 84, 60, 149, 50, 90, 259, 40, 90, 101, 70, 90, 254, 130, 80, and 40 days. For the infected group the times are 50, 130, 100, 130, 50, 140, 129, 76, 138, 69, 70, 144, 70, 130, 70, 150, 251, 110, and 120 days.

Let  $X_i$  be the time to abortion of the  $i^{\text{th}}$  cow infected with *neospora caninum* and let  $Y_i$  be the time to abortion of the  $i^{\text{th}}$  cow in the uninfected group. We assume that the abortion times in each group are log normal:  $X_1, \dots, X_{19} | \mu_1, \tau_1$  iid log normal( $\mu_1, 1/\tau_1$ ),  $Y_1, \dots, Y_{26} | \mu_2, \tau_2$  iid log normal( $\mu_2, 1/\tau_2$ ). We assume  $\mu_i \sim N(0, 1000)$  and  $\tau_i \sim \Gamma(0.001, 0.001)$  for  $i = 1, 2$ . This is an approximation to Jeffrey's prior on  $\mu_1, \mu_2, \tau_1$ , and  $\tau_2$ , see Sec. 4.6. Given the model parameters, the median times to abortion in the two groups are  $\exp(\mu_1)$  and  $\exp(\mu_2)$  so the difference in medians is  $\Delta = \exp(\mu_2) - \exp(\mu_1)$ . We are interested in the posterior distribution of  $(\Delta | X_1, \dots, X_{19}, Y_1, \dots, Y_{26})$  and specifically whether this difference is reasonably close to zero.

Computer simulations were used to estimate the posterior median of  $\Delta$ , 27.9 days, and the equal-tailed 95% credible interval for  $\Delta$ , (1.6, 54.8) days. With 95% probability, those infected have a median time to abortion between 1.6 and 54.8 days longer than those cows not infected with *neospora caninum*. If we use uniform priors on all parameters, these numbers change only slightly; the posterior median is 28.0 days and the 95% credible interval is 1.8 to 54.9 days.

Often survival data like this is subject to censoring, which is to say that an individual's data might eventually be lost to the study. Censoring is discussed briefly in Section 2.6 and more extensively in Chapter ?.

### 2.4 Armadillo hunting: Poisson data

The Ache tribe of Paraguay are part-time hunter-gatherers and have been in contact with Paraguayan society only since the mid-1970's. McMillan (2001) collected data on many aspects of Ache life including hunting, nutrition status, demographic features, child care, and health. Part of Ache life is spent away from the village on extended forest treks. Each trek lasts from three days to several weeks. Between one and six families participate. While on trek, men typically search for game animals by themselves but come together to capture peccaries, monkeys, and paca in groups. While trekking, the Ache subsist exclusively on foods that they collect on a given day. Meats are agouti paca, peccary, deer, coati, capuchin monkeys, a variety of reptiles, toucans and other birds, and insects. Gathered foods include

palm hearts, palm starch, oranges and other fruit, and honey. However, armadillos comprise the vast majority of food calories consumed by the Ache and it is of interest to quantify the typical number of armadillos killed in a day.

Consider a Poisson model for  $Y_i$ , the number of armadillos killed by 38 Ache men:  $Y_1, \dots, Y_{38} | \lambda \stackrel{iid}{\sim} \text{Poisson}(\lambda)$ . We refer to  $\lambda$  as the *kill rate*. (Something that should probably also be analyzed for Vin Diesel movies.) The broader Ache data include many hunting days for each man. For this analysis one day was randomly selected from each man. Dr. Garnett McMillan, an expert on Ache hunting practices, believes that Ache men typically kill an armadillo every other day and thus provides a “best guess” of  $\lambda$  to be 0.5 armadillos, which we take to be the median of the prior distribution. Dr. McMillan is 95% sure that the mean daily number of kills is no greater than 2 armadillos. A conjugate gamma prior distribution conveniently, but accurately provides a model for prior information on the mean daily number of kills  $\lambda$ . With  $\lambda \sim \Gamma(a, b)$ , we solve the simultaneous equations  $P(\lambda \leq 0.5 | a, b) = 0.50$ ,  $P(\lambda \leq 2 | a, b) = 0.95$  for  $a$  and  $b$  yielding  $a = 1.11$  and  $b = 1.61$ . Thus our model is completed by specifying  $\lambda \sim \Gamma(1.11, 1.61)$ .

In this simple model the posterior for  $\lambda$  may be found analytically. From Table 1.3 we find  $\lambda | Y_1, \dots, Y_n \sim \Gamma(\sum_{i=1}^n Y_i + a, n + b)$ . In the case of the Ache data,  $n = 38$ ,  $\sum_{i=1}^{38} Y_i = 10$ ,  $a = 1.11$ , and  $b = 1.61$ . We obtain  $\lambda | Y_1, \dots, Y_{38} \sim \Gamma(11.11, 39.61)$ . The prior median daily kill rate is 0.497 armadillos per day with a 95% credible interval of (0.024, 2.433). The posterior median daily kill rate is 0.272 armadillos per day with a 95% credible interval of (0.140, 0.468). With probability 0.95, the mean number of armadillos killed per day is between 0.14 and 0.47 of an armadillo, or one armadillo killed per 2 to 7 days. These are estimated from the data and use the expert’s prior.

In this example the posterior focuses tightly on smaller values of  $\lambda$  than the prior (see Figure 2.1). The data indicate that the mean number of kills is less than the expert’s best guess. In fact, the posterior 95% credible interval for  $\lambda$  does not contain 0.5, so we could reasonably reject the value 0.5. Does this suggest that the expert’s opinion is suspect? Not at all! First, the expert’s prior encompasses a wide range of plausible values for  $\lambda$  that are roughly centered at 0.5 and the posterior 95% credible interval falls well within the middle 95% of the prior.

If the posterior credible interval had fallen far outside the prior interval we would have evidence of a discrepancy: the data and the expert would indicate two very different scenarios regarding the daily kill rate of armadillos. This could happen, for example, if the expert’s opinion was based on the historical abundance of armadillos and the population declined shortly before the sample was collected. This divergence suggests a possible rethinking of armadillo abundance or Ache hunting habits by the expert. A

Parameter	Median	95% C.I.
$\beta_0$	-0.7147	(-1.007, -0.433)
$\beta_1$	0.01368	(-0.002525, 0.03085)
$\beta_2$	-0.002683	(-0.004007, -0.001459)
$\sigma$	0.4252	(0.2731, 0.658)

TABLE 2.1. Posterior medians and credible intervals.

strength of the Bayesian paradigm is that it provides a natural forum for the comparison and synthesis of scientific opinion derived from theory and historical information with current data.

## 2.5 Ache Hunting with age trends

We now incorporate random hunter effects into the Ache data analysis along with a tendency for older hunters to be more successful. Data were collected on the daily number of armadillos killed by 38 adult males of an Ache tribe over *several* forest treks; there are 1302 observations  $Y_{ij}$  total where  $i = 1, \dots, 38$  indexes an Ache male and  $j = 1, \dots, N_i$  denotes a trek. A plot of the average number of kills per man by age shows a generally increasing, then decreasing trend; it is of interest to model and quantify how a man's age affects daily kill success. We assume the number of armadillos killed is distributed Poisson and take the log-rate to be a quadratic function of a man's age in years  $a_i$  and a subject-specific random effect  $\delta_i$ . We include a normally distributed random effect for each man to account for the correlation of an individual's daily kills over the several hunting trips in which the data were collected. We also include a simple linear regression for the effect of age on kill rate. Clearly this is an approximation for the data at hand. Hunting effectiveness would eventually begin to decrease when age gets too large. For example, octogenarians would not kill many armadillos but there aren't any octogenarians out trekking.

The sampling model is specified

$$\begin{aligned}
 Y_{ij} | \lambda_i &\stackrel{iid}{\sim} \text{Poisson}(\lambda_i), & i = 1, \dots, 38; j = 1, \dots, N_i, \\
 \log(\lambda_i) &= \beta_0 + \beta_1(a_i - \bar{a}) + \beta_2(a_i - \bar{a})^2 + \delta_i \\
 \delta_i &\stackrel{iid}{\sim} N(0, \tau^{-1}), & i = 1, \dots, 38.
 \end{aligned}$$

Here  $\lambda_i$  is the mean daily kill rate for individual  $i$ . The ages  $a_i$  are fixed, known constants. The parameters  $\beta_0$ ,  $\beta_1$ ,  $\beta_2$ , and  $\tau$  are given the vague prior distributions  $\beta_i \sim N(0, 1000)$   $i = 0, 1, 2$  and  $\tau \sim \Gamma(0.001, 0.001)$ .

We obtain posterior information from simulation; summary information is presented in Table 2.1. The quadratic coefficient  $\beta_2$  is estimated to be



FIGURE 2.1. Prior (dashed) and posterior (solid) distributions for kill rate  $\lambda$ .

-0.0027 and is clearly nonzero. If we exponentiate  $\log(\lambda)$ , we obtain a “plug-in” estimate of the mean daily kill rate as a function of age  $a$ :

$$\left[ \hat{\lambda}(a) = \exp\{-0.7147 + 0.01368(a - \bar{a}) - 0.002683(a - \bar{a})^2\}. \right]$$

We use posterior credible intervals every five years from ages 20 to 70 to obtain Figure 2.2, a plot of the 95% credible intervals across typical ages of Ache hunters. The range in the data is 20 to 66 years. We see that the average kill-rate increases with age up until about 50, perhaps reflecting that hunting experience increases the chance of killing an armadillo, but then declines as the hunter enters his “golden age.” When the model was refit using independent uniform priors on all model parameters the resulting posterior inferences were almost identical to those presented above.

## 2.6 Lung cancer treatment: log-normal regression

Consider data presented in Maksymiuk et al. (1993) on the treatment of limited-stage small lung cancer. The data were analyzed using Bayesian semiparametric models by Walker and Mallick (1999), Kottas and Gelfand (2001), and Hanson and Johnson (2002). In the study, it was of interest to determine which sequencing of the drugs cisplatin and etoposide increased the lifetimes of those with limited-stage small cell lung cancer. Treatment 0 was the administration of cisplatin followed by etoposide, those receiving treatment 1 were given etoposide followed by cisplatin. The 121 patients studied were randomly assigned to the two treatment groups: 62 patients received treatment 0 and 59 patients received treatment 1. The data are the time in days  $T$  that a patient was known to be alive from the start of the treatment regimen, along with explanatory variables treatment ( $TR$ ) and the patient’s age ( $A$ ) at entry into the study. Define  $Y$  to be the time from the start of a treatment regimen until death due to small cell lung cancer. The *survival time*  $Y$  may or may not be the recorded time  $T$ . For some individuals the survival time  $Y$  is *right censored*. That is, the survival time is known only to be greater than the time recorded:  $Y > T$ . This can happen for several reasons. For example if a patient is alive at the end of the study we only know that survival time is longer than the time the individual spent in the study. Each individual has a noncensoring indicator  $d$  that is 0 for a right censored observation and 1 for an uncensored observation.

A portion of the data is reproduced here:



FIGURE 2.2. Mean daily kill by age and 95% CI.

Parameter	Median	95% C.I.
$\beta_0$	7.7	(6.7, 8.7)
$\beta_1$	-0.0184	(-0.0344, -0.0028)
$\beta_2$	-0.4024	(-0.6914, -0.1210)
$\tau$	1.72	(1.25, 2.30)

TABLE 2.2. Posterior medians and credible intervals.

$i$	$A_i$	$T_i$	$d_i$	$TR_i$
1	56	730	1	0
2	70	1980	0	0
3	56	260	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
120	71	251	1	1
121	63	254	1	1

Let  $Y_i$  be the survival time for patient  $i$ . Note that when  $d_i = 1$ ,  $Y_i = T_i$ , but when  $d_i = 0$ ,  $Y_i > T_i$ . Consider the log-normal regression model

$$\begin{aligned} \log Y_i &= \beta_0 + \beta_1 A_i + \beta_2 TR_i + \epsilon_i, \quad i = 1, \dots, 121, \\ \epsilon_i | \tau &\stackrel{iid}{\sim} N(0, \tau^{-1}), \\ \beta_0, \beta_1, \beta_2 &\stackrel{iid}{\sim} N(0, 1000), \\ \tau &\sim \Gamma(0.001, 0.001). \end{aligned}$$

We are placing approximately flat distributions on  $\beta_0$ ,  $\beta_1$ , and  $\beta_2$  by specifying a large variance in the independent normal priors for these parameters. This attempts to reflect an absence of prior information for these parameters. Similarly, we place an approximation to Jeffrey's prior on  $\tau$ .

The posterior estimates, based on simulations, are tabulated in Table 2.2. Group 1 survives  $\exp(-0.4024) = 0.67$  times as long as group 0 on average. The 95% CI for  $\beta_2$  contains only negative values so we are confident that people in treatment group 1 do not survive as long as people in group 2. In either treatment group, adding ten years to one's entry age decreases median survival time by a multiplicative factor of  $\exp(-0.0184 \times 10) = 0.83$ .

## 2.7 Survival with random effects: Ache hunting

We continue our examination of Ache hunting skill by looking at the *time* it takes to find an armadillo. Much current anthropologic theory assigns a great deal of importance to skill in hunting success, which varies from person to person due to experience, age, and cognitive ability. Furthermore, an Ache male's status within the group is highly determined by their skill in hunting.

A search time  $T$  is defined to be the time in minutes from when an Ache hunter starts walking to when an armadillo burrow is found. The search is right censored if it stops before finding an armadillo burrow. This may occur because other animals are encountered, the hunter takes a break, or darkness occurs and the hunter settles in camp for the night. For right censored observations, we only know that the true search time would have been longer than the recorded time had the hunter not been distracted from his task.

Let  $i = 1, \dots, 14$  denotes a specific Ache male, while  $j = 1, \dots, N_i$  indexes a particular attempt at finding an armadillo burrow. We consider a model in which the log of the search time  $T$  is normally distributed.  $\log(T_{ij}) = \mu + \delta_i + \epsilon_{ij}$ , where  $\mu$  is an overall search effect,  $\delta_i$  is a random effect for the skill of the  $i^{\text{th}}$  hunter and  $\epsilon_{ij}$  is an individual random effect for a particular search. The model is further specified  $\epsilon_{ij}|\sigma \stackrel{iid}{\sim} N(0, \sigma^2)$  and  $\delta_i|\sigma_\delta \stackrel{iid}{\sim} N(0, \sigma_\delta^2)$ , with the  $\{\delta_i\}$  independent of the  $\{\epsilon_{ij}\}$ . We assume that the log survival times follow a normal distribution, that is, the survival times follow a log-normal distribution. Furthermore, we have assumed that the *acceleration factors*  $e^{\delta_i}$  follow a log-normal distribution as well. (More on this later.) These assumptions should be carefully checked through a residual analysis, which we illustrate in Section XXX.

We elicited priors for this model from our Ache expert, Dr. McMillan, assuming that model parameters are *a priori* independent. The model with  $\delta_i = 0$  is for those Ache hunters of “average hunting skill.” The expert provided a best guess of 90 minutes for the median hunting time for those hunters of average ability and is 95% sure that this median is under 150 minutes. Noting that for  $\delta_i = 0$ , the median search time is  $e^\mu$ , we use these statements to specify a normal prior for  $\mu$  by solving the simultaneous equations  $P(e^\mu \leq 20) = 0.05$  and  $P(e^\mu \leq 140) = 0.95$ , yielding  $\mu \sim N(4.0, 0.35)$ . A prior for  $\sigma$  is found by considering the random 3<sup>rd</sup> quartile, or 75<sup>th</sup> percentile, of average Ache hunter search times *given a specified median search time*. We denote  $M$  as the median search time for average Ache hunters, so the 3<sup>rd</sup> quartile is given by  $Me^{\sigma z_{0.75}}$  where  $z_{0.75} = 0.6745$  is the 3<sup>rd</sup> quartile of a standard normal. Given that  $M = 90$ , the expert provides a best guess of 180 minutes for 3<sup>rd</sup> quartile and is 95% sure the 3<sup>rd</sup> quartile is below 300 minutes. We translate these statements into  $0.5 = P(Me^{\sigma z_{0.75}} \leq 180|M = 90) = P(\sigma \leq 1.0276)$ , and  $0.95 = P(Me^{\sigma z_{0.75}} \leq 300|M = 90) = P(\sigma \leq 1.785)$ . The gamma prior  $\sigma \sim \Gamma(2.29, 2.92)$  satisfies these conditions.

A randomly selected Ache male has a random *acceleration factor*  $e^\delta$  relative to Ache hunters of average skill (whom have  $\delta = 0$ ) where  $\delta \sim N(0, \sigma_\delta^2)$ . The expert believes that the best Ache hunters are about 1.5 times faster at finding burrows than average Ache hunters and at the most twice as fast. We assume the the best Ache hunters are those in the upper 10% of hunting ability and thus consider the random 90<sup>th</sup> percentile, that is,  $Q$  is defined

Parameter	Prior		Posterior	
	Median	95% C.I.	Median	95% C.I.
$\mu$	4.0	(2.8, 5.2)	3.7	(3.4, 4.0)
$\sigma$	0.67	(0.11, 2.08)	1.24	(1.08, 1.45)
$\sigma_\delta$	0.32	(0.14, 0.59)	0.35	(0.18, 0.51)
$T_p$	54	(5, 575)	42	(3, 562)

TABLE 2.3. Posterior median and 95% credible intervals.

by the relationship  $P(e^\delta \leq Q|\sigma_\delta) = 0.9$ , where  $Q = e^{z_{0.9}\sigma_\delta} = e^{1.282\sigma_\delta}$ . In the expert's opinion, we have  $P(Q \leq 1.5) = 0.5$  and  $P(Q \leq 2) = 0.95$ , so  $P\left(\sigma_\delta \leq \frac{\log(1.5)}{1.282}\right) = 0.5$  and  $P\left(\sigma_\delta \leq \frac{\log(2)}{1.282}\right) = 0.95$ . We fit a gamma prior to  $\sigma_\delta$  by solving the above equations to obtain  $\sigma_\delta \sim \Gamma(8.1, 24.5)$ .

Define  $T_p$  to be the predictive search time for a randomly selected Ache male different from the ones included in the analysis. In Table 2.3 we present posterior results for model parameters and  $T_p$ . The median time to find an armadillo burrow is estimated to be 42 minutes with a plausible range of 3 to 562 minutes.

This simple parametric survival model fits the data quite well. We examine more sophisticated survival models in Chapter ?, including models based on Dirichlet process mixtures, Polya trees, and gamma process.

## 2.8 Rats in Arizona: Multiple discrete time series

Brown et al. (2001) describe an extensive data set that includes the presence of certain rodent species in Portal, Arizona over time. The data were collected as part of an ongoing long-term study begun in 1977 on the growth or decline of species related to environmental factors in the southwestern U.S. desert. The site is comprised of 24 quarter-hectare plots; in 16 of the plots kangaroo rats were removed and fences erected to maintain their absence. The other 8 "control" plots were left as is so kangaroo rats were free to come and go as they please.

To investigate how species interact we considered 51 time periods of 6 months (indexed  $j = 1, 2, \dots, 51$ ) and looked at whether *any* of the species were trapped at each of the 8 control sites (indexed  $i = 1, \dots, 8$ ) in a given time period. The species considered were *Dipodomys merriami* (Merriam's kangaroo rat, identified as  $m_{ij} = 0$  if none were trapped,  $m_{ij} = 1$  if at least one was trapped), *Dipodomys ordii* (Ord's kangaroo rat,  $o_{ij} = 0, 1$ ), *Dipodomys spectabilis* (Banner-tailed kangaroo rat,  $b_{ij} = 0, 1$ ), and *Perognathus baileyi* (Bailey's pocket mouse,  $p_{ij} = 0, 1$ ). For example  $o_{16} = 1$  indicates that at least one Ord's kangaroo rat was trapped at site  $i = 1$  during the  $j = 6^{th}$  6-month time period.

Par.	With site effect	No site effect
$\alpha$	-1.77 (-2.37,-1.21)	-1.77 (-2.29,-1.28)
$\beta_o$	-0.37 (-0.97,0.22)	-0.37 (-0.93,0.20)
$\beta_b$	-1.48 (-3.46,-0.10)	-1.51 (-3.51,-0.12)
$\beta_p$	-0.47 (-2.14,0.90)	-0.60 (-2.25,0.71)
$\beta_m$	2.77 (2.19,3.36)	2.82 (2.26,3.41)
$e^{\beta_o}$	0.69 (0.38,1.25)	0.69 (0.39,1.22)
$e^{\beta_b}$	0.23 (0.03,0.91)	0.22 (0.03,0.89)
$e^{\beta_p}$	0.63 (0.12,2.46)	0.55 (0.10,2.03)
$e^{\beta_m}$	15.9 (8.9,15.9)	16.7 (9.55,30.33)

TABLE 2.4. Posterior results for Portal data.

These four species are prevalent at this location and since they eat similar foods we expect that competition among these four would be strongest relative to other species in area. We are particularly interested in how other species affect the presence or absence of Merriam's kangaroo rat, the most prevalent of the four rodents at Portal, as well as the propensity for Merriam's kangaroo rat to "take hold" once present. To this end we examine a simple lagged logistic regression model in which the probability of seeing Merriam's kangaroo rat,  $\pi_{ij}$ , depends on whether there is currently a captured Ord's kangaroo rat, Banner-tailed kangaroo rat, or Bailey's pocket mouse, and whether there was a Merriam's kangaroo rat trapped in the previous time period.  $m_{ij}|m_{i,j-1} \sim \text{Bernoulli}(\pi_{ij})$ ,  $\text{logit}(\pi_{ij})|m_{i,j-1} = \alpha + \beta_o o_{ij} + \beta_b b_{ij} + \beta_p p_{ij} + \beta_m m_{i,j-1} + \gamma_i$ ,  $\gamma_i \sim N(0, \tau^{-1})$ . Here, a random site effect  $\gamma_i$  was included in the model to account for Merriam's kangaroo rat preferring certain sites. The various species have habitat preferences and the 8 control sites vary somewhat with respect to amount of vegetation, water availability, and other factors. Independent,  $N(0, 1000)$  priors were placed on  $\alpha$ ,  $\beta_o$ ,  $\beta_b$ ,  $\beta_p$ ,  $\beta_m$ , and a  $\Gamma(0.001, 0.001)$  prior placed on  $\tau$ . To examine the strength of a possible site effect, we additionally fit the model without the random effects:  $m_{ij}|m_{i,j-1} \sim \text{Bernoulli}(\pi_{ij})$ ,  $\text{logit}(\pi_{ij})|m_{i,j-1} = \alpha + \beta_o o_{ij} + \beta_b b_{ij} + \beta_p p_{ij} + \beta_m m_{i,j-1}$ . Table 2.4 contains simulated posterior medians and 95% credible intervals for models both with and without the site effect.

The inclusion of site effects change the posterior results very little. The sites with the largest effects (not shown) were the only sites to have Bailey's pocket mouse present. This is the effect,  $\beta_p$ , that changes the most when site effects are left out of the model. We find that the presence of Merriam's kangaroo rat 6 months earlier increases the odds of seeing this same species at the present by about 17 times. The presence of the Banner-tailed kangaroo rat is the only other effect that is statistically significant: the 95% credible interval for  $\beta_b$  does not include zero whereas credible intervals for  $\beta_o$  and  $\beta_p$  do contain zero. If the Banner-tailed kangaroo rat is *not* present, the odds of

seeing Merriam's kangaroo rat is about  $1/0.22 = 4.5$  times as large as when Ord's kangaroo rat is present; a 95% CI is  $(1/0.89, 1/0.03) = (1.1, 33.3)$ , perhaps indicating competition among these two species affects Merriam's kangaroo rat negatively.

## 2.9 Fisheries in Namibia: Nonlinear time series

We re-examine data from Hillborn and Mangel (1997, Chapter 10) on commercial hake fishing in a particular area of the Atlantic off Namibia (on the west coast of Africa). For years this area was fished without restrictions by large trawlers primarily from Spain, South Africa, and the Soviet Union. The catch per unit effort (CPUE) and profits declined dramatically during this time leading to formation of the International Commission for Southeast Atlantic Fisheries (ICSEAF) in the mid 1970's. We examine data encompassing 23 years (1965 - 1988) that includes the formation of the ICSEAF. Of theoretical interest is the "maximum sustainable yield" (MSY), the amount of fish that can be harvested that produces a steady-state in fish population dynamics and therefore maximizes profit without driving hake to extinction.

Let  $C_t$  be the catch in year  $t$  as measured in thousands of tons and let  $I_t$  be tons of fish caught per standardized trawler hour.  $I_t$  is an index of fish abundance. Hillborn and Mangel (1997) suggest modelling the data as follows. The model assumes

$I_t = qB_tV_t$  where  $\log(V_t) \stackrel{iid}{\sim} N(0, \sigma^2)$ .  $B_t$  is a latent (unobserved) variable that is the biomass of hake stock vulnerable to fishing at start of time period  $t$ .  $B_t$  is assumed to relate to  $I_t$  through  $I_t = qB_tV_t$  where  $\log V_t$  are independent  $N(0, \tau^{-1})$ . Future biomass is related to current biomass through  $B_{t+1} = B_t + rB_t(1 - \frac{B_t}{K}) - C_t$ .

The model parameters are  $r$ , the hake growth rate;  $K$ , the equilibrium hake population size in the absence of fishing; and  $q$  a constant of proportionality relating  $I_t$  to  $B_t$ . The  $V_t$ 's are multiplicative log-normal errors. As calculated from the model, the  $MSY$  is given by  $MSY = kr/4$ . Flat priors were placed on all model parameters except for  $\tau$ , which was given the improper prior  $f(\tau) \propto 1/\tau$ . Posterior median and 95% CI's are in Table 2.5.

We estimate the  $MSY$  to be 267,000 tons of hake a year, less than what was actually harvested from about 1968 through 1976, during which time the ICSEAF began regulating hake fishing in the area (Figure 2.3). A visible decline in fish abundance  $I_t$  is noted throughout these years, possibly due to overfishing. From 1977 through 1988, we see the abundance slowly climb towards 200,000 tons per year, perhaps in part due to the fishing regulations. The observed and predicted abundance, along with 95% prediction intervals are found in Figure 2.4; most observed values fall well within the prediction intervals indicating some amount of modelling success.



FIGURE 2.3. Catch and abundance by year.  $C_t$  is dashed and  $300I_t$  is solid.

Par.	Posterior Median	95% CI
$q \times 10000$	4.4	(3.5, 5.4)
$k$	2730	(2375, 3280)
$r$	0.39	(0.30,0.50)
$s$	0.13	(0.10,0.18)
$MSY$	267	(246,287)

TABLE 2.5. Posterior results: hake fishing example.

This model is quite difficult to fit as a frequentist using maximum likelihood methods. Additionally, one cannot easily obtain the spectrum of inferences available from a Bayesian analysis. For example, prediction intervals for catch  $C_t$  at each year  $t$  are difficult to obtain as a frequentist (Figure 2.4). Hillborn and Mangel (1997) used a profile likelihood analysis to fit the model and asymptotic approximations to obtain standard errors. Although we did not fit this model in WinBUGS, a simple “Metropolis within Gibbs” approach (Tierney, 1994) worked well here and was readily coded in *Mathematica*.



FIGURE 2.4. Observed abundance  $I_t$  and 95% prediction intervals.