Introduction to Applied Bayesian Statistics and Estimation for Social Sciences: Answers to Exercises (and error corrections therein)

Scott M. Lynch[1]

July 7, 2009

[1]Scott M. Lynch is Associate Professor, Department of Sociology and Office of Population Research, Princeton University, Princeton NJ, 08544 (email: slynch@princeton.edu).

Answers to Appendix A Exercises, pages 335-336

Calculus Exercises

1. $\frac{1}{\infty} = 0$
2. Technically, as written, there is no limit—the function is discontinuous at 0. The limit, however, can be taken from the right (∞) or left ($-\infty$).
3. $\frac{\infty}{3} = \infty$
4. $-\frac{1}{5}$
5. $e^{-x} = \frac{1}{e^x}$. So, $\lim_{x\to\infty} \frac{1}{e^x} = \frac{1}{\infty} = 0$.
6. $\frac{1}{e^0} = 1$
7. $\frac{x}{2x} = \frac{1}{2}$.
8. $\frac{1}{\infty}e^0 = 0$
9. $\frac{dy}{dx} = 9x^2 + 10x + 2$. Evaluated at 3 is 113.
10. $\int_0^5 (3x^3 + 5x^2 + 2x + 10)dx = (3/4)x^4 + (5/3)x^3 + x^2 + 10x \mid_0^5 = 752.08$

Matrix Algebra Exercises

1. $\begin{bmatrix} 7 & 2 & 2 \\ 11 & 4 & 13 \\ 16 & 12 & 8 \end{bmatrix}$
2. $\begin{bmatrix} 1 & 0 & 4 \\ 3 & -2 & 3 \\ -4 & 4 & 6 \end{bmatrix}$
3. Does not conform for multiplication (number of columns in first matrix does not equal the number of rows in second matrix)
4. $2 + 63 + 12 = 77$
5. $\begin{bmatrix} 2 & 14 & 8 \\ 9 & 63 & 36 \\ 3 & 21 & 12 \end{bmatrix}$
6. Does not conform
7. $\begin{bmatrix} 4 & 7 & 6 \\ 1 & 1 & 8 \\ 3 & 8 & 7 \end{bmatrix}$
8. $(3)(3) - (2)(-1) = 11$

9. $9 + 50 + (-16) - [-30 + 4 + 60] = 9$

Step 1: Replace elements with det(*minor*)	Step 2: Sign appropriately	Step 3: Transpose to get adjoint
$\begin{bmatrix} -57 & 1 & 50 \\ -17 & 10 & 26 \\ 5 & 11 & -3 \end{bmatrix}$	$\begin{bmatrix} -57 & -1 & 50 \\ 17 & 10 & -26 \\ 5 & -11 & -3 \end{bmatrix}$	$\begin{bmatrix} -57 & 17 & 5 \\ -1 & 10 & -11 \\ 50 & -26 & -3 \end{bmatrix}$

10.

Step 4: Get determinant of original matrix	Step 5: Multiply by adjoint to obtain inverse
$28 + 48 + 168 - [18 + 256 + 49] = -79$	$\begin{bmatrix} 57/79 & -17/79 & -5/79 \\ 1/79 & -10/79 & 11/79 \\ -50/79 & 26/79 & 3/79 \end{bmatrix}$

11. $3 + 3 + 1 = 7$

12. The area of a unit square is 1. The volume of a unit cube is 1. The hypervolume of any multidimensional unit hypercube is also 1.

Answers to Chapter 2 Exercises, pages 44-45

Probability Exercises

1. Must find c so that:
$$\frac{1}{c} = \int_r^s (mx + b)dx$$
Evaluate integral:
$$\frac{1}{c} = \frac{mx^2}{2} + bx\Big|_r^s.$$
$$\frac{1}{c} = \frac{ms^2}{2} + bs - \left(\frac{mr^2}{2} + br\right).$$
Simplifying:
$$\frac{2}{c} = ms^2 + 2bs - mr^2 - 2br.$$
Further simplification:
$$\frac{2}{c} = m(s + r)(s - r) + 2b(s - r)$$
Finally:
$$c = \frac{2}{(s - r)[m(s + r) + 2b]}.$$

2. $p(x = 3|p = .5) = \binom{3}{3}(.5)^3(1 - .5)^{3-3} = .125$

3. $p(x = 3|p = .7) = \binom{3}{3}(.7)^3(1 - .7)^{3-3} = .343$

4. probability of three heads is .125; probability of three tails is .125. Probability of one or the other is .125+.125=.25.

5. $p(x = 3|n = 4, p = .5) = \binom{4}{3}(.5)^3(1 - .5)^1 = .25$

6. $\mu = (200)(.5) = 100$; $\sigma^2 = 200(.5)(1 - .5) = 50$
$z = \frac{130-100}{7.07} = 4.24$; $p(z > 4.24) = 0$

7. The following R code will produce the plot:
`plot(seq(0,10,by=.1),dnorm(seq(0,10,by=.1),mean=5,sd=2),type="l")`

8. see above; change parameter values

9. Use the following R code (substitute for k):
`plot(seq(-5,5,by=.1),dt(seq(-5,5,by=.1),df=k),type="l")`

10. The density function for the multivariate normal distribution is:
$$(2\pi)^{-k/2}|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(X - \mu)^T\Sigma^{-1}(X - \mu)\right\}.$$

If the dimensionality is 1, then $|\Sigma|$ is just σ^2, and Σ^{-1} is just $1/\sigma^2$. Thus, the kernel reduces to:

$$\exp\left\{-\frac{1}{2}\frac{(X-\mu)^2}{\sigma^2}\right\},$$

and the normalizing constant becomes:

$$(2\pi)^{-1/2}|\Sigma|^{-1/2} = \frac{1}{\sqrt{2\pi\sigma^2}.}$$

Classical Inference Exercises

1. The MLE is $20/30 = 2/3$.

2. There are multiple ways to answer this question; here's one. Under the assumption that the coin is fair ($p = 1/2$), we can calculate the probability of observing an MLE of $\hat{p} = 2/3$ in a sample of size $n = 30$ by using the z distribution:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{(2/3) - (1/2)}{\sqrt{\frac{(1/2)(1/2)}{30}}} = 1.83.$$

   If $z = 1.83$, then the probability of observing a sample of this size this extreme (two-tail) is: $p(|z| \geq 1.83) = .068$. From a classical perspective, this is not sufficient information to reject the null hypothesis that the coin is fair.

   An alternative approach is to compute the probability of obtaining 20 or more heads in 30 coin flips (and 15 or less) to obtain the probability a random sample from a fair coin would produce such 'extreme' data as 20 heads. This probability is .099.

3. **NOTE:** This exercise is incorrectly worded but can still be answered. First, the question should state that $n = 10$, not $s = 10$; that is, a sample of $n = 10$ students were administered an IQ test resulting in a mean $\bar{x} = 120$. Using that information, we can compute the probability of observing such a sample, under the assumption that college students and the general population have equivalent mean IQs as follows:

$$z = \frac{120 - 100}{\frac{16}{\sqrt{10}}} = 3.95.$$

   So, $p(|z| > 3.95) = .00008$. Given the rarity of obtaining such a mean in a random sample from the general population, we might conclude that college students are more intelligent than average. That is, they do not constitute a sample from the general population, but instead possibly from a population with greater mean IQ.

   Considering the question as it is stated in the text, we could conclude that the sample is rare enough to consider college students more intelligent than average so long as our sample size was larger than 2 people. That is, if we recomputed z from above, we would find that a sample size of 3 gives us $z = 2.17$, which is sufficient to reject the null hypothesis.

4. No. Given $\mu = 100$ and $\sigma = 16$, a person chosen at random from the population has a probability of .211 of having an IQ 1.25 standard deviations from the mean or greater $\left(z = \frac{120-100}{16} = 1.25;\ p(|z| > 1.25) = .211\right)$. This is certainly not a rare event.

5. The conclusions are opposite. It is rarer to obtain a sample mean 20 IQ points from the population mean than it is to find a single individual with an IQ 20 points from the population mean.

6. Start with the first derivative wrt μ on page 42 (after substituting $\tau = \sigma^2$):

$$\frac{\partial LL}{\partial \mu} = \frac{n(\bar{x} - \mu)}{\tau}.$$

Taking the second derivative wrt μ yields:

$$\frac{\partial^2 LL}{\partial \mu^2} = \frac{-n}{\tau},$$

which is the (1,1) element of the Hessian matrix shown on page 43. Taking the second derivative with respect to τ yields:

$$\frac{\partial^2 LL}{\partial \mu \tau} = -n(\bar{x} - \mu)\tau^{-2},$$

which constitutes the diagonal elements of the Hessian matrix. The remaining element is only slightly more difficult to obtain. Start with the log-likelihood in Eq. 2.39, but replace $-n\ln(\sigma)$ with $(-n/2)\ln(\tau)$ and replace $\sum(x_i - \mu)^2$ with v. The log-likelihood is then:

$$LL = \frac{-n\ln(\tau)}{2} - \frac{v\tau^{-1}}{2}.$$

Then,

$$\frac{\partial LL}{\partial \tau} = \frac{-n\tau^{-1}}{2} + \frac{v\tau^{-2}}{2},$$

and

$$\frac{\partial^2 LL}{\partial \tau^2} = \frac{n\tau^{-2}}{2} - v\tau^{-3},$$

which, after simplification and substitution, is the (2,2) element of the Hessian matrix shown on page 43.

7. Poisson likelihood:

$$L(\lambda|x) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x!} = \left(\frac{1}{\prod x_i}\right) e^{-n\lambda}\lambda^{\sum x_i}.$$

Log-likelihood:

$$LL = \ln\left(\frac{1}{\prod x_i}\right) - n\lambda + \sum x_i \ln\lambda.$$

Derivative wrt λ

$$\frac{\partial LL}{\partial \lambda} = -n + \frac{\sum x_i}{\lambda}.$$

Set to 0 and solve:

$$\lambda = \frac{\sum x_i}{n} = \bar{x}.$$

8. Back up to next to last step from previous item and take second derivative:

$$\frac{\partial^2 LL}{\partial \lambda^2} = -\frac{\sum x_i}{\lambda^2}.$$

Expectation of $\sum x_i$ is $n\lambda$, so:

$$I(\lambda) = \frac{n\lambda}{\lambda^2}.$$

Invert, simplify, and square root to obtain the standard error: $\sqrt{\lambda/n}$.

Answers to Chapter 3 Exercises, pages 74-75

1. Consider the Venn diagram in Fig. 2.1. What we would like to know is $p(B|A)$, but suppose what we know is $p(A|B)$, and we know $p(B)$. If we also know $p(A|B^c)$ and we know $p(B^c)$, then we can compute the total probability of A: It is simply the sum of the probability of being in A when B is true (i.e., the portion of A that is within the B circle—$p(A, B)$) and being in A when B is false (i.e., the portion of A that is *not* within the B circle. This is the denominator of Bayes rule; it simply rescales the sample space to the desired region (the totality of A). What we want, then, is the joint probability of being in both A and B as a proportion of the total reduced sample space A. This is the numerator of the formula and follows from the basic conditional probability rule. In other words, Bayes theorem simply rescales the joint probability as a proportion of a region that is known.

2. If false positives and false negatives occur with equal probability, and the accuracy rate defined in this manner is 90%, then $p(+|\text{preg})$ is still .9 but $p(+|\text{not preg}) = .1$ Thus, the only change is in the denominator of the formula. Substituting yields:

$$p(\text{preg}|+) = \frac{(.9)(.15)}{(.9)(.15) + (.1)(.85)} = .614.$$

   This posterior probability is much greater as a result of the much reduced probability of a positive test under the more likely scenario of not being pregnant.

3. $$p(\text{cancer}|+) = \frac{(.9)(.00001)}{(.9)(.00001) + (.9)(.99999)} = .00001.$$

   Interestingly, if the test has a 90% false positive rate and a 90% accuracy rate for positive cases, then the posterior probability is the same as the prior probability—nothing is gained from the test. In other words, a positive test is just as likely to occur whether the individual has cancer as not, and so the test is noninformative. So, under these conditions, no one of any age should be tested, unless, of course, the false positive rate is lower for older males. In reality, the false positive rate of the PSA rate is *not* that high, but, given the low incidence of prostate cancer among young males, the posterior probability is not high for them even with a positive test.

4. If the likelihood is binomial with $x = 5$ successes out of $n = 10$ tries, and the prior is beta with parameters α and β, then the posterior is Beta($\alpha + 5$, $\beta + 5$). The following R code will produce plots for a beta distribution with $\alpha = 5 + a$ and $\beta = 5 + b$:

   ```
   plot(seq(0,1,by=.01),dbeta(seq(0,1,by=.01),5+a,5+b),type="l")
   ```

   Obviously, the larger the beta prior parameters, the larger the effect on the posterior. If α and β are equal in the prior, just as the count of successes and failures are equal in the likelihood, the only effect of the prior is to shrink the posterior's width. On the other hand, if α and β are not set equal, the posterior becomes skewed.

5. Assume the variance is known. What we are interested in is only the posterior distribution for the mean. As shown on page 64, if the variance is known, given the sample values for $\bar{x}$ and σ^2, the posterior distribution for the mean is normal with a mean of $(144M + \tau^2(169)(100))/((169)\tau^2 + 144)$ and a variance of $(144\tau^2)/(169\tau^2+144)$. Thus, we only need to specify values for M and τ^2 in order to be able to plot the posterior distribution.

   The following R code will plot the posterior distribution for normal data:

```
plot(seq(90,110,by=.1),dnorm(seq(90,110,by=.1),
     mean=((144*m+16900*t)/(169*t+144)),
     sd=sqrt((144*t)/(169*t+144))),type="l",ylab="f(x)",xlim=c(95,105))
```

   If you substitute various values for m and t—the prior mean and variance—you will find that the value of the prior variance seems to have the largest effect on the posterior. Our choice for the prior mean is practically irrelevant if the prior variance is large. For example, set $t = 10000$ and try different values for m. Of course, you could also argue that the mean matters considerably if the prior variance is small. Because the posterior is a mixture of the prior and likelihood, having a small prior variance with a prior mean that is extreme causes considerable shrinkage toward the prior.

6. For the sake of space, I do not replicate all 10 tests; doing so is straightforward: Simply substitute the posterior probability from one test into Bayes' formula as the prior for the next test. Regarding the second part of the question, if we consider all tests to be taken simultaneously, then Bayes' formula is:

$$p(\text{preg}|+10 \text{ times}) = \frac{p(+10 \text{ times}|\text{preg})p(\text{preg})}{p(+10 \text{ times}|\text{preg})p(\text{preg}) + p(+10 \text{ times}|\text{not preg})p(\text{not preg})}.$$

   Under the original scenario for accuracy and false positives, the probability of testing positive 10 times in a row if one is pregnant is simply $.9^{10}$. Additionally, the probability of testing positive 10 times in a row if one is not pregnant is simply $.5^{10}$. Substitution yields:

$$p(\text{preg}|+10 \text{ times}) = \frac{(.9^{10})(.15)}{(.9^{10})(.15) + (.5^{10})(.85)} = .984,$$

   which is the same result obtained by sequentially updating the posterior probability one test at a time.

   Note that we can represent the probabilities of testing positive 10 times in a row either directly, as here, or we can represent them via the binomial distribution. Indeed, $.9^{10}$ is what the binomial distribution simplifies to when all trials are successes (i.e., the combinatorial reduces to 1, as does $(1-p)^0$).

   For the third part of the question, we no longer consider the prior to be a point estimate of .15. Instead, we are to construct a beta prior distribution. In order to

pick an appropriate prior distribution, let's assume we want the distribution to be concentrated over .15—the most likely value for the prior pregnancy probability—but we recognize that this value is not exact, especially for the *individual's* pregnancy probability, and so we wish this probability to vary considerably. The Beta(.15,.85) distribution suits this purpose. Most of its mass is centered over .15, but 25% of its mass extends between .2 and 1. If we choose this prior, our posterior, after obtaining 10 successful pregnancy tests is Beta(10.15,.85). The mean of this beta distribution is .92, and the median is .95. A 95% credible interval for these estimates is [.72,1]. Thus, there is a large probability that the woman's pregnancy probability is large.

7. Yes, the posterior probability distribution for a Kerry victory could have been updated at every poll, just as the last exercise showed—the end result is the same whether the posterior is updated sequentially or all at once. The result of sequentially updating the posterior, at least with the CNN polls shown, would have shown the posterior distribution slowly shifting in Kerry's favor. This is much more informative than assuming each poll is 'new' and that its information has no bearing on or relationship with the previous data.

8. (1) A normal distribution prior with variance 0 in a normal distribution problem. In that case, the posterior mean will be the prior mean, and the posterior variance will be 0. More generally, any time a degenerate prior is used—that is, any time the prior places all probability for a parameter on a single point—the posterior distribution will be determined by the prior. (2) A normal distribution prior with infinite mean. In that case, the posterior mean will be be the prior mean also—∞.

9. The multinomial mass function is:

$$pr(x_1 \dots x_k | n, p_1 \dots p_k) = \frac{n!}{x_1! x_2! \dots x_k!} p_1^{x_1} p_2^{x_2} \dots p_1^{x_1} p_2^{x_2} \dots p_k^{x_2}.$$

Given the constraints that $\sum p = 1$ and $\sum x = n$, if we only have two possible outcomes, we have p and $1-p$, and we have x and $n-x$. Thus, the density reduces to:

$$\frac{n!}{x!(n-x)!} p^x (1-p)^{n-x},$$

which is the binomial mass function. For the Dirichlet distribution, with two possible outcomes, replace α_2 with β, and replace x_1 with p and x_2 with $1-p$. Substitution shows that this reduces to the beta distribution.

10. The density for the Wishart is:

$$f(X) \propto |X|^{(v-d-1)/2} \exp\left\{ -\frac{1}{2} \text{tr}(S^{-1}X) \right\}$$

If the dimensionality is 1, the trace of X is just x, $(v-2)/2 = \alpha - 1$, and so $\alpha = v/2$. Within the exponential, s and x are scalars, so the trace function can be eliminated,

and we are left with $-(1/2s) = -\beta$, or $\beta = 1/2s$. This is proportional to a gamma distribution.

11. The inverse chi square distribution is simply a special case of the inverse gamma distribution with parameter $\alpha = v/2$ and $\beta = 1/2$. As we have seen, in a normal distribution, the marginal distribution for a variance is an inverse gamma distribution with parameters $\alpha = (n-1)/2$ and $\beta = (n-1)\text{var}(x)/2$. This looks like an inverse chi square distribution with $v = n-1$, with the β parameter scaled by $(n-1)\text{var}(x)$. In fact, the variance is considered a scaled inverse chi square random variable.

12. Because a uniform density gives equal weight to all values of a random variable over an interval, the uniform density is simply a constant. Because proportionality is simply a matter of multiplication, any constant is proportional to a uniform density over some interval; the constant can be multiplied by a scalar to make the density proper.

Answers to Chapter 4 Exercises, pages 105

1. The question should ask for F^{-1} of $x|y$. $F^{-1}(y|x)$ is derived on page 90. For $x|y$:

$$u = \int_0^z \frac{2x + 3y + 2}{6y + 8} dx,$$

giving us $u(6y + 8) = x^2 + (3y + 2)x \,|_0^z$, and so

$$u(6y + 8) = z^2 + (3y + 2)z.$$

This equation can be solved by completing the square in z

$$u(6y + 8) + \left(\frac{3y + 2}{2}\right)^2 = \left(z + \frac{3y + 2}{2}\right)^2$$

Taking the square root of both sides and solving for z yields:

$$z = \sqrt{u(6y + 8) + \left(\frac{3y + 2}{2}\right)^2} - \frac{3y + 2}{2}.$$

2. This density (as stated previously) is bounded by 0 and 2 in both dimensions, making the maximum value of the unnormalized function 12. A rejection sampler can be constructed by (1) sampling values from a bivariate uniform distribution on the (0,2,0,2) square (function $g(x)$ in step 1 on page 84), (2) computing the ratio R as shown in step 2 on page 84, with $m = 48$, and (3) following the rule in step 3. Here is an R program that will do this:

```
count=0; items=0
x=matrix(NA,1000,2)

while(items<1000)
{
count=count+1
u=as.matrix(runif(2,min=0,max=2))

r=(2*u[1]+3*u[2]+2)/12
if(runif(1,0,1)<r){
items=items+1
x[items,]=t(u)
}
```

3. The question does not place limits on the density, so assume the density is bounded on the interval $[r, s]$. So, the normalizing constant is (see page 15):

$$\frac{2}{(s-r)[5(s+r)+4]}.$$

For inversion, we first draw a $U(0,1)$ random variable u, which represents area under the curve. This implies the equation:

$$u = \int_r^z \frac{2}{(s-r)[5(s+r)+4]}(5x+2)dx,$$

which we wish to solve for z. After moving the constant and taking the integral, we obtain:

$$\frac{u(s-r)[5(s+r)+4]}{2} = (5/2)x^2 + 2x \mid_r^z,$$

which equals:

$$\frac{u(s-r)[5(s+r)+4]}{2} = (5/2)z^2 + 2z - (5/2)r^2 - 2r.$$

Next, we can move all terms not containing z to the left side of the equation and multiply through by $(2/5)$ to obtain:

$$\frac{u(s-r)[5(s+r)+4]}{5} + r^2 + (4/5)r = z^2 + (4/5)z.$$

After completing the square in z, square rooting both sides, and solving for z, we obtain:

$$z = \sqrt{\frac{u(s-r)[5(s+r)+4]}{5} + r^2 + (4/5)r + (4/10)^2} - (4/10).$$

Constructing an R routine to perform inversion sampling from this linear density can be done as follows (say, for 1000 draws):

```
r=?
s=?
u=runif(1000,min=0,max=1)
z= sqrt(u*(s-r)*(5*(s+r)+4)/5 + r^2 +(4/5)*r + (4/10)^2)-(4/10)
```

4. In R, answering this question simply involves simulation of n draws from a $gamma(\alpha = 1498, \beta = 3017)$ distribution as follows:

```
x=rgamma(1000,shape=1498,rate=3017)
```

The only difficulty arises when trying to determine whether β is a rate or scale parameters. Try examining the help documentation for `rgamma` to see why it is a rate parameter.

5. (Note: I use `set.seed(413)` to set the random number generator seed for the purpose of replication.)
   Step 1: Simulating 20 N(0,1) draws. In R, simply type: `x=rnorm(20,0,1)`.

   Step 2: Use the R program on page 103, but now `x` is already defined.

   Step 3: Use the last 1000 draws from the posterior distribution for μ and generate a histogram. We can obtain the height of the t density function using Equation 2.28. In this equation, substitute the mean of μ for μ, the standard deviation of μ for σ, and 19 for v (the degrees of freedom) and then plot. In R:

```
plot(density(mu[1001:2000],n=50),type="h")

w=gamma(10)/(gamma(19/2)*sqrt(19*pi)*sd(mu[1001:2000]))*

(1+(1/19)*((seq(-3,3,by=.1)-mean(mu[1001:2000]))/sd(mu[1001:2000]))^2)^(-10)

lines(seq(-3,3,by=.1),w,lty=2)
```

   The two densities appear to be very similar. The reason is that, although the conditional distribution for $\mu|\sigma^2$ is normal, sampling μ across the distribution for σ^2 ultimately yields the full marginal distribution for μ. That distribution is t.

6. See footnote 1 on page 150.

Answers to Chapter 5 Exercises, pages 129-130

1. As stated, the identity claimed in the question is not true. The question should say: Show that $A^TBA = tr(AA^TB)$.

   Consider the left side of this identity:

$$[a_1 \quad a_2 \quad a_3] \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

   This product will ultimately yield a scalar. After multiplying the first two matrices, we are left with a 1×3 row vector:

$$[(a_1b_{11} + a_2b_{21} + a_3b_{31}) \quad (a_1b_{12} + a_2b_{22} + a_3b_{32}) \quad (a_1b_{13} + a_2b_{23} + a_3b_{33})]$$

   Carrying out the next multiplication yields:

$$[a_1(a_1b_{11} + a_2b_{21} + a_3b_{31}) + a_2(a_1b_{12} + a_2b_{22} + a_3b_{32}) + a_3(a_1b_{13} + a_2b_{23} + a_3b_{33})]$$

   Now consider the right side of the identity:

$$tr\left(\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} [a_1 \quad a_2 \quad a_3] \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix} \right)$$

   The first product yields a 3×3 matrix:

$$\begin{bmatrix} a_1^2 & a_1a_2 & a_1a_3 \\ a_2a_1 & a_2^2 & a_2a_3 \\ a_3a_1 & a_3a_2 & a_3^2 \end{bmatrix}$$

   The second product then yields a 3×3 matrix. However, given that we ultimately take the trace of this matrix, we need only focus on the diagonal elements and sum them:

$$(a_1^2b_{11} + a_1a_2b_{21} + a_1a_3b_{31}) + (a_2a_1b_{12} + a_2^2b_{22} + a_2a_3b_{32}) + (a_3a_1b_{13} + a_3a_2b_{23} + a_3^2b_{33})$$

   This result is identical to the result above.

2. In Chapter 2, we saw that the normalizing constant for the linear density was a function of the domain of x and the parameters b and m. We also saw that the normalizing constant served to rescale the intercept and slope so as to preserve the "relative frequency" of occurrence of values of x while keeping consistent with the rule that the total area under the density function is 1. For the purpose of this problem, we can combine the process of finding the normalizing constant and the process of showing

how m and b are related as follows. First, we know the density with rescaled m and b parameters must integrate to 1 over the $[r, s]$ interval:

$$1 = \int_r^s (mx + b)dx.$$

After integration, we obtain

$$1 = \frac{mx^2}{2} + bx \mid_r^s,$$

And after evaluation and simplification, we find

$$2 = ms^2 - mr^2 + 2b(s - r).$$

On page 115, we saw the solution of this equation in terms of b, but we can just as easily solve it for m:

$$m = \frac{2 - 2b(s - r)}{(s^2 - r^2)}$$

If we wish to have the program simulate b, and then compute m, we only need to change two lines in the program—the first two lines within the `for` loop:

```
b[i]=b[i-1] + rnorm(1,mean=0,sd=.002)
m[i]=(2-10*b[i])/25
```

If you now rerun the MH algorithm for both approaches, you'll find that the results are extremely similar. However, you'll also find that the acceptance rate is higher for the "b first" approach, given the relatively narrow variance of the proposal distribution and the wider posterior distribution for b.

3. If you allow both parameters to be updated, both parameters increase seemingly without bound, regardless of how long you run the MH algorithm. If you add independent beta priors for both parameters, you'll find that you can control how fast the parameters increase, and, ultimately, you can keep it from happening, but only if the priors are strong enough to outweigh the data/likelihood. If you use a bivariate normal prior distribution with a corrrelation of .5, standard deviations of .01 for both parameters, and 0 for the means, you can get the MH algorithm to stabilize, but the parameter values are unacceptable. The slope parameter, m, becomes negative a large, while the intercept parameter, b, remains close to 0. Clearly this produces a density with negative values as x increases. All in all, if the constraint that b is determined by m is ignored, it is difficult or impossible to obtain reasonable results.

4. In the MH algorithm on page 124, only two parameters are updated using Gibbs sampling: the means of x and y. Thus, the only changes that need to be made to make the algorithm entirely MH is before the comment "`#update sums of squares`." The changes we need to make include: (1)the replacement of the simulation step for generating `mx[]` and `my[]`(now we simulate a candidate from a proposal); and (2)

repeated updating of the sums of squares. As the program snippet below shows, I now draw candidates from a uniform distribution of width .02 centered over the previous value for the parameter, and I compute the sums of squares and cross products using the current value of the mean parameters as they are updated.

```
#update sums of squares
sx2=sum((x-mx[i-1])^2); sy2=sum((y-my[i-1])^2);
sxy=sum((x-mx[i-1])*(y-my[i-1]))

mx[i]=mx[i-1]+runif(1,-.01,.01)
#update sums of squares
sx2i=sum((x-mx[i])^2); sy2i=sum((y-my[i-1])^2);
sxyi=sum((x-mx[i])*(y-my[i-1]))

if((lnpost(r[i-1],mx[i],my[i-1],sx[i-1],sy[i-1],sx2i,sy2i,sxyi)
   -lnpost(r[i-1],mx[i-1],my[i-1],sx[i-1],sy[i-1],sx2,sy2,sxy))
   <log(runif(1,min=0,max=1)))
{acc=0; mx[i]=mx[i-1]}
accmx=accmx+acc

#update sums of squares
sx2=sum((x-mx[i])^2); sy2=sum((y-my[i-1])^2);
sxy=sum((x-mx[i])*(y-my[i-1]))

my[i]=my[i-1]+runif(1,-.01,.01)
#update sums of squares
sx2i=sum((x-mx[i])^2); sy2i=sum((y-my[i])^2);
sxyi=sum((x-mx[i])*(y-my[i]))

if((lnpost(r[i-1],mx[i],my[i],sx[i-1],sy[i-1],sx2i,sy2i,sxyi)
   -lnpost(r[i-1],mx[i],my[i-1],sx[i-1],sy[i-1],sx2,sy2,sxy))
   <log(runif(1,min=0,max=1)))
{acc=0; my[i]=my[i-1]}
accmy=accmy+acc
```

5. The first issue to consider for this problem is how the intercept is determined once the values of the slopes have been selected. To find this, we need to solve the following for b:

$$
\begin{aligned}
1 &= \int_0^5 \int_0^5 (m_1 x + m_2 y + b) dy dx. \\
1 &= \int_0^5 (m_1/2)x^2 + m_2 xy + bx \mid_0^5 \\
1 &= \int_0^5 (25/2)m_1 + 5m_2 y + 5b \\
1 &= (25/2)m_1 y + (5/2)m_2 y^2 + 5by \mid_0^5 \\
1 &= (125/2)m_1 + (125/2)m_2 + 25b \\
2 &= 125m_1 + 125m_2 + 50b \\
b &= (2 - 125m_1 - 125m_2)/50
\end{aligned}
$$

Given this result, an MH algorithm can be written as follows:

```
#data already stored, subtract 1 so range is 0-5
x=x-1; y=y-1;

m1=matrix(0,50000); m2=matrix(0,50000); b=matrix(.04,50000)
acctot=0

for(i in 2:50000){

#draw candidate
m1[i]=m1[i-1]+runif(1,min=-.0003,max=.0003)
m2[i]=m2[i-1]+runif(1,min=-.0003,max=.0003)
b[i]=(2-125*m1[i]-125*m2[i])/50

#compute posterior at current and previous values
acc=tot=1
for(j in 1:1377){
tot=tot*((x[j]*m1[i] + y[j]*m2[i] + b[i])/
         (x[j]*m1[i-1] + y[j]*m2[i-1] + b[i-1]))
}

#evaluate for rejection
if(runif(1,min=0,max=1)>tot || b[i]<0){
m1[i]=m1[i-1]; m2[i]=m2[i-1]; b[i]=b[i-1]; acc=0
}
acctot=acctot+acc

if(i%%100==0){print(c(i,m1[i],m2[i],b[i],acctot/i))}
}
```

6. There are minimal changes that need to be made to the algorithm in the text. Below are the changes, including that the scale matrix `sc` is now computed only once using the sample means, and that the simulation of Σ is removed from the loop for simulating

the mean vector (and with df=1376). Given that the `riwish` function only makes one random draw at a time, we must loop it, and so it really need not be extracted from the loop simulating the mean vector. (Note that the `riwish()` function is not part of the base R system—instead, it is part of the MCMC pack, which can be downloaded and installed.)

```
.
.
.
mn=matrix(c(mean(d[,1]),mean(d[,2])),2)

#simulate s
 e[,1]=d[,1]-mn[1]; e[,2]=d[,2]-mn[2]

 sc=t(e)%*%(e)

for(i in 1:10000){s[i,,]=riwish(1376,sc)}

for(i in 1:10000)
{
 #simulate m
 u=rnorm(2,mean=0,sd=1)
 m[i,]=t(mn) + t(u)%*%chol(s[i,,]/1377)
 corr[i]=s[i,1,2]/sqrt(s[i,1,1]*s[i,2,2])

 if(i%%100==0){print(c(i,corr[i]))}
}
```