SDS 383D: Exercise 1

Diego Garcia-Olano

February 5, 2017

Problem 1. Bayesian inference in simple conjugate families

We start with a few of the simplest building blocks for complex multivariate statistical models: the beta/binomial, normal, and inverse- gamma conjugate families.

(A) Suppose that we take independent observations $X_1, \ldots, X_N$ from a Bernoulli sampling model with unknown probability w. That is, the X_i are the results of flipping a coin with unknown bias. Suppose that w is given a Beta(a, b) prior distribution:

$$p(w) = \Gamma(a+b)w^{a-1}(1-w)^{b-1},$$

where $\Gamma(\cdot)$ denotes the Gamma function. Derive the posterior distribution $p(w|x_1, \ldots, x_N)$.

By Bayes rule we know,

$$p(w|y) = \frac{p(w) * p(y|w)}{p(y)}$$

where $p(w) \sim Beta(a, b)$ is the prior probability distribution (belief) of results for coin flips, $p(y|w) \sim Binom(w)$ is the sampling model probability distribution with density function

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

and $p(y)$ is the marginal distribution, where

$$p(y) = \int p(y, w)d\theta = \int p(w)p(y|w)$$

Plugging in the values for the prior, the sampling probability and the marginal we get:

$$p(w|y) = \frac{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} w^{a-1}(1-w)^{b-1}\binom{n}{y}w^y(1-w)^{n-y}}{\int p(w)p(y|w)}$$

$$= \frac{w^{a+y-1}(1-w)^{b+(n-y)-1}}{\int w^{a+y-1}(1-w)^{b+(n-y)-1}}$$

$$= \frac{w^{a+y-1}(1-w)^{b+(n-y)-1}}{Beta(a+y, b+n-y)}$$

$$= \frac{1}{Beta(a+y, b+n-y)} w^{a+y-1}(1-w)^{b+(n-y)-1}$$

This is the $Beta(a+y, b+ny)$ distribution.

Thus $p(w|y) \sim Beta(a+y, b+ny)$.

Thus a prior Bernouli leads to a posterior Beta, ie the **Bernoulli-Beta** model.

...

(B) The probability density function (PDF) of a gamma random variable, $X \sim Ga(a, b)$, is

$$p(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$$

Suppose that $X_1 \sim Ga(a_1, 1)$ and that $X_2 \sim Ga(a_2, 1)$.

Define two new random variables $y1 = x1/(x1 + x2)$ and $y2 = x1 + x2$.

Find the joint density for (y1, y2) using a direct PDF transformation (and its Jacobian). Use this to characterize the marginals p(y1) and p(y2), and propose a method that exploits this result to simulate beta random variables, assuming you have a source of gamma random variables.

The density functions for X_1 and X_2 are $f_{x_1}(x_1) = \frac{x_1^{a_1-1}}{\Gamma(a_1)e^{x_1}}$ and $f_{x_2}(x_2) = \frac{x_2^{a_2-1}}{\Gamma(a_2)e^{x_2}}$.

The joint distribution for x_1,x_2 is then

$$f(x_1, x_2) = f_{x_1}(x_1) f_{x_2}(x_2) = \frac{x_1^{a_1-1} x_2^{a_2-1}}{\Gamma(a_1)\Gamma(a_2)e^{x_1+x_2}}$$

Rewrite $y1 = \frac{x_1}{x_1+x_2}$ so $x_1 = y_1(x_1 + x_2)$ and by substitution $x_1 = y_1 y_2$

similarly $y_2 = x_1 + x_2$ so $x_2 = y_2 - x_1$ and by substitution $x_2 = y_2 - (y_1 y_2)$

now let $v_1(y_1, y_2) = y_1 y_2$

and $v_2(y_1, y_2) = y_2 - y_1 y_2)$

Find Jacobin for system with v_1 and v_2,

$$|J| = \left|\begin{matrix} \frac{\partial v_1}{\partial y_1} & \frac{\partial v_1}{\partial y_2} \\ \frac{\partial v_2}{\partial y_1} & \frac{\partial v_2}{\partial y_2} \end{matrix}\right| = \left|\begin{matrix} y_2 & y_1 \\ -y_2 & 1 - y_1 \end{matrix}\right|$$

$$= (1 - y_1)(y_2) - (-y_2)(y_1)$$

$$= y_2 - y_2 y_1 + y_2 y_1 = y_2$$

Now using the change of variables equation, for the joint density of y_1 and y_2 we have

$$g(y_1, y_2) = |J| f[v_1(y_1, y_2), v_2(y_1, y_2)]$$

$$= y_2(f[y_1 y_2, y_2 - y_1 y_2])$$

$$= y_2 \frac{((y_1 y_2)^{a_1-1}(y_2 - y_1 y_2)^{a_2-1}}{\Gamma(a_1)\Gamma(a_2)e^{y_1 y_2 + y_2 - y_1 y_2}}$$

$$= \frac{y_2 (y_1 y_2)^{a_1-1}(y_2 - y_1 y_2)^{a_2-1}}{\Gamma(a_1)\Gamma(a_2)e^{y_2}}$$

$$= \frac{y_2 (y_1 y_2)^{a_1-1}(y_2 - y_1 y_2)^{a_2-1}}{\Gamma(a_1)\Gamma(a_2)e^{y_2}}$$

$$= \frac{y_1^{a_1-1}(1 - y_1)^{a_2-1} y_2^{a_1+a_2-1}}{\Gamma(a_1)\Gamma(a_2)e^{y_2}}$$

We can group terms and rewrite as:

$$= \frac{1}{\Gamma(a_1)\Gamma(a_2)} y_1^{a_1-1}(1-y_1)^{a_2-1} * y_2^{a_1+a_2-1} e^{y_2}$$

We observe the left hand side looks like the $Beta(a1, a2)$ distribution

and that the right hand side looks like $Gamma(a_1 + a_2, 1)$.

In fact the joint density $f(y_1, y_2) = Beta(a_1, a_2) * Gamma(a_1 + a_2, 1)$

and thus $y_1 \sim Beta(a_1, a_2)$ and y_2 $Gamma(a_1 + a_2, 1)$

Because $y_1 \sim Beta(a_1, a_2)$ and y_1 is linear combination of x_1 and x_2

and $x_1, x_2 \sim Gamma$, we can generate x_1 and x_2 samples to obtain y_1 Beta ones.

...

(C) Suppose that we take independent observations $x_1, \ldots, x_N$ from a normal sampling model with unknown mean θ and known variance σ^2: $x_i \sim N(\theta, \sigma^2)$. Suppose that θ is given a normal prior distribution with mean m and variance v. Derive the posterior distribution $p(\theta|x_1, ..., x_N)$.

Our sampling model $\sim N(\theta, \sigma^2)$ has pdf $p(X|\theta) = \prod_{i=1}^{n} \frac{1}{(2\sigma^2\pi)^{1/2}} e^{-\frac{(x_i-\theta)^2}{2\sigma^2}}$.

prior $\theta \sim N(m, v)$ with pdf $p(\theta|m, v) = \frac{1}{(2v\pi)^{1/2}} e^{-\frac{(\theta-m)^2}{2v}}$

marginal $p(x_1, \ldots, x_n) = \int p(x_1, \ldots, x_n, \theta) = \int p(\theta)p(x_1, \ldots, x_n|\theta)$

thus by bayes rule,

the posterior $p(\theta|x_1, \ldots, x_N) = \frac{p(\theta)p(X|\theta)}{p(\theta)}$

which is proportional to the numerator, so

$p(\theta|x_1, \ldots, x_N) \propto p(\theta)p(X|\theta)$

$$\begin{aligned} p(\theta|x_1, \ldots, x_N) &\propto \frac{1}{(2v\pi)^{1/2}} e^{-\frac{(\theta-m)^2}{2v}} \prod_{i=1}^{n} \frac{1}{(2\sigma^2\pi)^{1/2}} e^{-\frac{(x_i-\theta)^2}{2\sigma^2}} \\ &\propto e^{-\frac{(\theta-m)^2}{2v^2}} e^{\sum_{i=1}^{n} -\frac{(x_i-\theta)^2}{2\sigma^2}} \\ &= e^{\frac{-1}{2v}\theta^2 + m^2 + 2m\theta} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n} x_i^2 + n\theta^2 + 2\theta\sum_{i=1}^{n} x_i} \end{aligned} \tag{1}$$

now we can pull out constants

$$\begin{aligned} p(\theta|x_1, \ldots, x_N) &\propto e^{\frac{-n}{2v}\theta^2 - 2m\theta} e^{\frac{-1}{2\sigma^2}\theta^2 - 2\bar{x}\theta} \\ &= e^{\frac{-n}{2v\sigma^2}\left(\frac{\sigma^2}{n}(\theta^2 - 2m\theta) + v(\theta^2 - 2\bar{x}\theta)\right)} \end{aligned} \tag{2}$$

after grouping, and simplifying we get

$$p(\theta|x_1, \ldots, x_N) \propto e^{\frac{-nv+\sigma^2}{2\sigma^2 v}\left(\theta - \frac{\bar{x}nv + \sigma^2 m}{nv+\sigma^2}\theta\right)^2} \tag{3}$$

and hence our posterior $\sim N\big(\frac{v}{v+\sigma^2/n}\bar{x} + \frac{\sigma^2/n}{v+\sigma^2/n}m, (\frac{n}{\sigma^2} + \frac{1}{v})^{-1}\big)$

Thus for a Normal sampling model with known variance and unknown mean θ that has a normal prior of the form N(m,v) we get a Normal posterior.

*Note: The precision, which is the inverse of variance, is additive in Gaussian models: ie, the posterior precision is the sum of the prior precision (1/v) and the sample data precision (n/σ^2). The mean, moreover, is a weighted average of prior mean m and of sample average θ, whose weights are the precisions related to them.

(D) Suppose that we take independent observations $x_1, \ldots, x_N$ from a normal sampling model with known mean θ but unknown variance σ_2. (This seems even more artificial than the last, but is conceptually important.) To make this easier, we will re-express things in terms of the precision, or inverse variance $\omega = \frac{1}{\sigma_2}$

$$p(x_i|\theta, \omega) = \big(\frac{\omega}{2\pi}\big)^{1/2} e^{-\frac{\omega}{2}(x_i-\theta)^2}$$

Suppose that ω has a gamma prior with parameters a and b, implying that σ^2 has what is called an inverse-gamma prior, written $\sigma^2 \sim IG(a, b)$. Derive the posterior distribution $p(\omega|x_1, ..., x_N)$. Re-express this as a posterior for σ^2, the variance.

given prior Gamma(a,b) has pdf $p(\omega) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$

and our sampling model above, the posterior:

$$\begin{aligned} p(\omega|x_1, ..., x_N) &\propto p(\omega)p(x_i|\theta, \omega) \\ &= \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} \big(\frac{\omega}{2\pi}\big)^{1/2} e^{-\frac{\omega}{2}(x_i-\theta)^2} \\ &\propto \omega^{a-1} e^{-b\omega} \big[\big(\frac{\omega}{2\pi}\big)^{1/2}\big]^N e^{-\omega/2 \sum_{i=1}^N (x_i-\theta)^2} \\ &= \omega^{a+N/2-1} e^{-b(\omega+1/2 \sum_{i=1}^N (x_i-\theta)^2)} \end{aligned} \tag{4}$$

This equation is proportional to the Gamma pdf. From (B) we know Ga(c,d) has pdf

$$p(x) = \frac{d^c}{\Gamma(c)} x^{c-1} e^{-dx}$$

which is equivalent (4) if we let $c = a + N/2$, and $d = b + 1/2 \sum_{i=1}^N (x_i - \theta)^2$

thus $p(\omega|x_1, ..., x_N) \propto Ga(a + N/2, b + 1/2 \sum_{i=1}^N (x_i - \theta)^2)$

and therefore

$$p(\sigma^2|X) \sim IG(a + N/2, b + 1/2 \sum_{i=1}^N (x_i - \theta)^2)$$

This is Normal-inverse gamma model for sampling model $\sim N(\theta, \sigma^2)$ with known mean θ and unknown variance $\sigma^2 \sim IG$.

(E) Suppose that as above, we take independent observations $x_1, \ldots, x_N$ from a normal sampling model with unknown, common mean θ. This time, however, each observation has its own idiosyncratic (but known) variance: $x_i \sim N(\theta, \sigma^2)$. Suppose that θ is given a normal prior distribution with mean m and variance v. Derive the posterior distribution $p(\theta|x_1, \ldots, x_N)$. Express the posterior mean in a form that is clearly interpretable as a weighted average of the observations and the prior mean.

Given prior θ $N(m, v)$ and sampling method $p(X|\theta)$ $N(\theta, \sigma^2)$, our posterior:

$$\begin{aligned} p(\theta|X) &\propto p(X|\theta)p(\theta) \\ &= \prod_{i=1}^{N} \big((\frac{1}{2\pi\sigma_i^2})^{1/2} e^{-1/2\sigma_i^2(x_i-\theta)^2} \big) \frac{1}{(2\pi v)^{1/2}} e^{-1/2v(\theta-m)^2} \\ &\propto e^{\frac{-1}{2}\left[(\sum_{i=1}^{N} \frac{(x_i-\theta)^2}{\sigma_i^2}) + \frac{(\theta-m)^2}{v}\right]} \\ &= e^{\frac{-1}{2}\left[\sum_{i=1}^{N}(\frac{x_i^2}{\sigma_i^2} + \frac{\theta_i^2}{\sigma_i^2} - 2\frac{x_i\theta}{\sigma_i^2}) + \frac{\theta^2+m^2-2m\theta}{v}\right]} \\ &\propto e^{\frac{-1}{2}\left[\theta^2(\frac{1}{v} + \sum_{i=1}^{N}\frac{1}{\sigma_i^2}) - 2\theta(\frac{m}{v} + \sum_{i=1}^{N}\frac{x_i}{\sigma_i^2})\right]} \end{aligned} \tag{5}$$

Therefore, $p(\theta|X) \sim N(\frac{s}{t}, (t)^{-1})$

where $s = \frac{m}{v} + \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2}$ and $t = \frac{1}{v} + \sum_{i=1}^{N} \frac{1}{\sigma_i^2}$

* This is the Normal-Normal model for unknown mean and known idiosyncratic variances.

The precision ($(\frac{1}{v} + \sum_{i=1}^{N} \frac{1}{\sigma_i^2})^2$ is the sum of the prior precision and of each single idiosyncratic precision of the data. The mean is again an average of the prior mean and of the weighted average of the data (weighted by the precisions). This heteroschedastic model is useful when dealing with robust regression.

(F) Suppose that $(x|\sigma^2) \sim N(0, \sigma^2)$, and that $1/\sigma^2$ has a Gamma(a,b) prior, defined as above. Show that the marginal distribution of x is a $Student's\ t$.
This is why the t distribution is often referred to as a *scale mixture of normals*.

Our sampling model $(x|\sigma^2) \sim N(0, \sigma^2)$ with prior $\frac{1}{\sigma^2} \sim Gamma(a, b)$
can be rewritten as $(x|\omega) \sim N(0, \omega)$ with prior $\omega \sim Gamma(a, b)$ given precision $\omega = \frac{1}{\sigma^2}$
thus, the marginal of x is

$$\begin{aligned} p(x) &= \int f(x|\omega) f(\omega) d\omega \\ &= \int (\frac{\omega}{2\pi})^{1/2} e^{-\frac{\omega}{2}x^2} \cdot \frac{b^a}{\Gamma(a)} \omega^{a-1} e^{-b\omega} \\ &= (\frac{1}{2\pi})^{1/2} \frac{b^a}{\Gamma(a)} \int \omega^{(a+1/2)-1} e^{-(b+1/2x^2)} d\omega \end{aligned} \tag{6}$$

The section under the integral is the kernel for a $Beta(a + 1/2, b + 1/2x^2)$ so we can write:

$$\begin{aligned} &= (\frac{1}{2\pi})^{1/2} \frac{b^a}{\Gamma(a)} \frac{\Gamma(a+\frac{1}{2})}{(b+\frac{1}{2}x^2)^{a+\frac{1}{2}}} \\ &= (\frac{1}{2\pi})^{1/2} \frac{b^a}{\Gamma(a)} \Gamma(a+\frac{1}{2})(b+\frac{1}{2}x^2)^{-1(a+\frac{1}{2})} \end{aligned} \tag{7}$$

This is in the form of the Student's t distribution

Problem 2. The multivariate normal distribution

Basics
We all know the univariate normal distribution, whose long history began with de Moivre?s 18th-century work on approximating the (ana- lytically inconvenient) binomial distribution. This led to the probability density function

$$p(x) = \frac{1}{(2\pi v)^{1/2}} exp\{ - \frac{(x-m)^2}{2v} \}$$

for the normal random variable with mean m and variance v, written $x \sim N(m, v)$.
Here is an alternative characterization of the univariate normal distribution in terms of moment-generating functions. a random variable x has a normal distribution if and only if $E\{exp(tx)\} = exp(mt + vt2/2)$ for some real m and positive real v. Remember that $E(\cdot)$ denotes the expected value of its argument under the given probability distribution. We will generalize this definition to the multivariate normal.

(A) First, some simple moment identities. The covariance matrix cov(x) of a vector-valued random variable x is defined as the matrix whose (i, j) entry is the covariance between x_i and x_j. In matrix notation, $cov(x) = E\{(x-\mu)(x-\mu)^T\}$, where μ is the mean vector whose i-th component is $E(x_i)$. Prove the following: (1) $cov(x) = E(xxT) - \mu\mu^T$ and(2) $cov(Ax + b) = Acov(x)A^T$ for matrix A and vector b.

* see images **multivariatenorm-a1.JPG** and **multivariatenorm-a2.JPG**

(B) Consider the random vector $z = (z1, ..., zp)^T$, with each entry having an independent standard normal distribution (that is, mean 0 and variance 1). Derive the probability density function (PDF) and moment-generating function (MGF) of z, expressed in vector notation. We say that z has a standard multivariate normal distribution. * Remember that the MGF of a vector- valued random variable x is the expected value of the quantity $e^{t^T x}$, as a function of the vector argument t.

* see images **multivariatenorm-b.JPG**

(C) A vector-valued random variable x = (x1, . . . , xp) variate normal distribution if and only if every linear combination of its components is univariate normal. That is, for all vectors a not

identically zero, the scalar quantity z = aT x is normally distributed. From this definition, prove that x is multivariate normal, written x ? N(?, S), if and only if its moment-generating function is of the form E(exptT x) = exp(tT ? + tT St/2). Hint: what are the mean, variance, and moment-generating function of z, expressed in terms of moments of x?

* see images **multivariatenorm-c.JPG**

(D) Another basic theorem is that a random vector is multivariate nor- mal if and only if it is an affine transformation of independent univariate normals. You will first prove the ?if? statement. Let z have a standard multivariate normal distribution, and define the random vector x = Lz + ? for some p ? p matrix L of full column rank.6 Prove that x is multivariate normal. In addition, use the mo- ment identities you proved above to compute the expected value and covariance matrix of x.

* see images **multivariatenorm-d.JPG**

(E) Now for the ?only if.? Suppose that x has a multivariate normal distribution. Prove that x can be written as an affine transformation of standard normal random variables. (Note: a good way to prove that something can be done is to do it!) Use this insight to propose an algorithm for simulating multivariate normal random variables with a specified mean and covariance matrix.

todo

(F) Use this last result, together with the PDF of a standard multi- variate normal, to show that the PDF of a multivariate normal x ? N(?, S) takes the form p(x) = C exp Q(x ?)/2 for some constant C and quadratic form Q(x ?).

* see images **multivariatenorm-f.JPG**

(G) Letx ? N(? ,S)andx ? N(? ,S),wherex andx are 11122212 independent of each other. Let y = Ax1 + Bx2 for matrices A, B of full column rank and appropriate dimension. Note that x1 and x2 need not have the same dimension, as long as Ax1 and Bx2 do. Use your previous results to characterize the distribution of y.

* see images **multivariatenorm-g.JPG**

Conditionals and marginals

Suppose that $x \sim N(\mu, \Sigma)$ has a multivariate normal distribution. Let x1 and x2 denote an arbitrary partition of x into two sets of components. Because we can relabel the components of x without changing their distribution, we can safely assume that x1 comprises the first k elements of x, and x2 the last p - k . We will also assume that μ and Σ have been partitioned conformably with x:

$\mu = (\mu_1, \mu_2)^T$ and $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$

Clearly $\Sigma_{21} = \Sigma_{12}^T$, as Σ is a symmetric matrix.

(A) Derive the marginal distribution of x1. (Remember your result about affine transformations.)

* see image **conditionals-marginals-a-and-b-1.JPG**

(B) Let $\Omega = \Sigma^{-1}$ be the inverse covariance matrix, or precision matrix, of x, and partition Ω just as you did Σ:

$$\Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}$$

Using (or deriving!) identities for the inverse of a partitioned matrix, express each block of Ω in terms of blocks of Σ.

* see images **conditionals-marginals-a-and-b-1.JPG** and **conditionals-marginals-b-2.JPG**

(C) Derive the conditional distribution for x_1, given x_2, in terms of the partitioned elements of x, μ, and Σ. There are several keys to inner peace: work with densities on a log scale, ignore constants that don?t affect x_1, and remember the cute trick of completing the square from basic algebra. (see note 8) Explain briefly how one may interpret this conditional distribution as a linear regression on x_2, where the regression matrix can be read off the precision matrix.

* see images **conditionals-marginals-c.JPG**

Problem 3. Multiple regression: three classical principles for inference

(A) Show that all three of these principles lead to the same estimator.

* see images **multipleregression-a.JPG**

(B) Now suppose you trust some observations more than others, and will estimate b by minimizing the weighted sum of squared errors,

$$\hat{B} =$$

where the wi are weights. (Trustworthy observations have large weights.) Derive this estimator, and show that it corresponds to the maximum-likelihood solution under heteroscedastic Gaussian error:

$$\hat{B} =$$

Make sure you explicitly connect the weights wi and the idiosyncratic variances σ_2.

* see images **multipleregression-b.JPG**

Problem 4. Quantifying uncertainty: some basic frequentist ideas

In linear regression

In frequentist inference, inferential uncertainty is usually characterized by the sampling distribution, which expresses how one's estimate is likely to change under repeated sampling. The idea

is simple: unstable estimators should not be trusted, and should therefore come with large error bars. This should be a familiar concept, but in case it is not, consult the tutorial on sampling distributions in this chapter's references.
Suppose, as in the previous section, that we observe data from a linear regression model with Gaussian error:

$$y = XB + \epsilon, \epsilon\ N(0, \sigma^2 I)$$

(A) Derive the sampling distribution of your estimator for b from the previous problem.

* see images **quantifying-uncertainity-a.JPG** and **quantifying-uncertainity-a-inclass.JPG**

(B) This sampling distribution depends on σ^2, yet this is unknown. Suppose that you still wanted to quantify your uncertainty about the individual regression coefficients. Propose a strategy for calculating standard errors for each B_j. Then consult the data set on ozone concentration in Los Angeles, where the goal is to regress daily ozone concentration on a set of other atmospheric variables. This is available from the R package "mlbench," with my R script "ozone.R" giving you a head start on processing things.

Calculate standard errors using your method, and then using the pre-packaged lm function in R. Note: you may have an essentially correct strategy for calculating standard errors that yields some- thing slightly different from the lm function. If so, that?s OK, can you explain the discrepancy?

* see images **quantifying-uncertainity-b-inclass.JPG** and **ozone.R**

Problem 5. Propagating uncertainty

Suppose you have taken data and estimated some parameters $\theta_1, ..., \theta_P$ of a multivariate statistical model, for example, the regression model of the previous problem. Call your estimate $\hat{\theta} = (\hat{\theta_1}, ..., \hat{\theta_P})^T$. Suppose that you also have an estimate of the covariance matrix of the sampling distribution of $\hat{\theta}$:

$$\hat{\Sigma} \approx cov(\hat{\theta}) = E[(\hat{\theta} - \bar{\theta})(\hat{\theta} - \bar{\theta})^T]$$

where the expectation is under the sampling distribution for the data, given the true parameter θ. Here $\bar{\theta}$ denotes the mean of the sampling distribution.

If you want to report uncertainty about the $\hat{\theta}_j$'s you can do so by peeling off the diagonal of the estimated covariance matrix: $\hat{\Sigma}_{jj} = \hat{\sigma}_j^2$ is the square of the ordinary standard error of $\hat{\theta}_j$. But what if you want to report uncertainty about some function involving multiple components of the estimate $\hat{\theta}$?

(A) Start with the trivial case where you want to estimate

$$f(\theta) = \theta_1 + \theta_2$$

Calculate the standard error of $f(\hat{\theta})$, and generalize this to the case where f is the sum of all p components of $\hat{\theta}$.

* see images **propagatinguncertainity-a-b.JPG**

(B) What now if f is a nonlinear function of the $\hat{\theta}_j$'s? Propose an approximation for $var\{f(\hat{\theta})\}$, where f is any sufficiently smooth function. (As above, the variance is under the sampling distribution of the data, given the true parameter.) There are obviously many potential strategies that might work, but here?s one you might find fruitful: try a first-order Taylor approxi- mationof f(q?)aroundtheunknowntruevalueq.Trytoboundthe size of the likely error of the approximation, or at least talk gener- ally about what kinds of assumptions or features of f or p(q? | q) might be relevant. You should also reflect on some of the potential caveats of this approach. *If you?re unfamiliar with the idea of the bootstrap, consult he review paper by Efron and Gong entitled ?A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation.?

* see images **propagatinguncertainity-a-b.JPG**

Problem 6. Bootstrapping

The basic idea is to simulate the process of repeated sampling from the population by re-sampling from your sample (with replacement). The ties and duplicates in your ?bootstrapped samples? will mimic the sampling variability of the true data-generating process.

(A) Let $\hat{\Sigma}$ denote the covariate matrix of the sampling distribution of $\hat{B}$, the least-squares estimator. Write an R function that will estimate $\hat{\Sigma}$ via bootstrapped resampling for a given design matrix X and response vector y. Use it to compute $\hat{\Sigma}$ for the ozone data set, and compare it to the parametric estimate based on normal theory.

From before, we have

$$\hat{B} = (X^T X)^{-1} X^T Y$$

$$\hat{\Sigma} = cov(\hat{B}) = \sigma^2 (X^T X)^{-} 1$$

(B) Now lets look at a few of these ideas. Write R functions that will accomplish the following:

(B)1. For a specified mean vector μ and covariance matrix Σ, simulate multivariate normal random variables.

(B)2. For a given sample $x_1, ..., x_N$ from a multivariate normal distribution, estimate the mean vector and covariance matrix by maximum likelihood. 12 Remember to work on a log scale. R has many possibilities for optimization; the built-in function **optim** usually works for me.

(B)3. Bootstrap a given sample $x_1, ..., x_N$ to estimate the sampling distribution of the MLE. Try out your code in d = 2 dimensions for a few different sample sizes N. See how well you can recover the true covariance matrix from simulated data.

* see **ozone.R** for A and B responses

Appendix A

R code