SDS 383D: Exercises 1

February 3, 2017

Professor Scott

Spencer Woody

Problem 1

Bayesian inference in simple conjugate families

(A) $X_1, \ldots, X_N | w \overset{\text{iid}}{\sim} \text{Bernoulli}(w)$, $w \sim \text{Beta}(a, b)$. Define $Y := \sum_{i=1}^{n} X_i$, so $Y|w \sim \text{Binomial}(N, w)$.

$$p(y|w) = P(Y = y|w) = \binom{N}{y} w^y (1-w)^{N-y} \tag{1}$$

$$p(w) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} w^{a-1}(1-w)^{b-1} \tag{2}$$

By Bayes' Rule,

$$p(w|y) \propto p(w)p(y|w) \tag{3}$$

$$\propto \left(w^{a-1}(1-w)^{b-1}\right)\left(w^y(1-w)^{N-y}\right) \tag{4}$$

$$= w^{a+y-1}(1-w)^{b+N-y-1}, \tag{5}$$

so $w|y \sim \text{Beta}(a + y, b + N - y)$

(B) We have two independently distributed variables, $X_1 \sim \text{Gamma}(a_1, 1)$ and $X_2 \sim \text{Gamma}(a_2, 1)$. The joint distribution of X_1 and X_2 is

$$f_{X_1,X_2}(x_1, x_2) = \frac{b^{a_1+a_2}}{\Gamma(a_1)\Gamma(a_2)} x_1^{a_1-1} x_2^{a_2-1} \exp\left[-(x_1 + x_2)\right] \tag{6}$$

Then we define the transformation of variables $(X_1, X_2) \mapsto (Y_1, Y_2)$ as follows:

$$Y_1 = \frac{X_1}{X_1 + X_2} \tag{7}$$

$$Y_2 = X_1 + X_2. \tag{8}$$

We can find the joint distribution of Y_1 and Y_2 with

$$f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}(g_1(y_1, y_2), g_2(y_1, y_2)) \, |J| \,, \tag{9}$$

where $x_1 = g_1(y_1, y_2) = y_1 y_2$, $x_2 = g_2(y_1, y_2) = y_2(1 - y_1)$, and J is the determinant of the Jacobian matrix,

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} = \begin{vmatrix} y_2 & y_1 \\ -y_2 & 1 - y1 \end{vmatrix} = y_2(1 - y_1) + y_2 y_1 = y_2. \tag{10}$$

Y_2 is the ratio of two nonnegative variables, so $|J| = |y_2| = y_2$. Now we can write (9) as

$$f_{Y_1,Y_2}(y_1, y_2) = \frac{b^{a_1+a_2}}{\Gamma(a_1)\Gamma(a_2)} (y_1 y_2)^{a_1-1} \left[y_2(1-y_1)\right]^{a_2-1} \exp\left[-(y_1 y_2 + y_2(1-y_1))\right] y_2 \tag{11}$$

$$= \frac{b^{a_1+a_2}}{\Gamma(a_1)\Gamma(a_2)} y_1^{a_1-1}(1-y_1)^{a_2-1} y_2^{a_1+a_2-1} \exp(-y_2). \tag{12}$$

Therefore, $Y_1 \sim \text{Beta}(a_1, a_2)$ independent of $Y_2 \sim \text{Gamma}(a_1 + a_2, 1)$.

Problem 1 continued on next page...

(C) $X_i|\theta \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma^2)$, $i = 1, 2, \ldots, N$ where σ^2 is *known* and $\theta \sim \mathcal{N}(m, v)$ is *unknown*. The posterior distribution of θ given $x_1, \ldots, x_N$ is

$$
\begin{aligned}
f(\theta|x_1, \ldots, x_N) &\propto f(x_1, \ldots, x_N|\theta) f(\theta) && (13)\\
&\propto \left(\prod_{i=1}^{N} \exp\left[-\frac{(x_i - \theta)^2}{2\sigma^2} \right] \right) \exp\left[-\frac{(\theta - m)^2}{2v} \right] && (14)\\
&= \exp\left[-\frac{\sum_{i=1}^{N} (x_i - \theta)^2}{2\sigma^2} - \frac{(\theta - m)^2}{2v} \right] && (15)\\
&\propto \exp\left[-\frac{n\theta^2 - 2n\bar{x}\theta}{2\sigma^2} - \frac{\theta^2 - 2m\theta}{2v} \right] && (16)\\
&= \exp\left[-\frac{\theta^2 - 2\bar{x}\theta}{\frac{2\sigma^2}{n}} - \frac{\theta^2 - 2m\theta}{2v} \right] && (17)\\
&= \exp\left[-\frac{1}{2\frac{\sigma^2 v}{n}} \left(v\theta^2 - 2v\bar{x}\theta + \frac{\sigma^2}{n}\theta^2 - 2\frac{\sigma^2}{n} m\theta \right) \right] && (18)\\
&= \exp\left[-\frac{1}{2\frac{\sigma^2 v}{n}} \left(\left[v + \frac{\sigma^2}{n} \right] \theta^2 - 2\left[v\bar{x} + \frac{\sigma^2}{n} m \right] \theta \right) \right] && (19)\\
&= \exp\left[-\frac{1}{2\frac{\sigma^2 v}{n} \left(\frac{1}{v + \frac{\sigma^2}{n}} \right)} \left(\theta^2 - 2\frac{v\bar{x} + \frac{\sigma^2}{n} m}{v + \frac{\sigma^2}{n}} \theta \right) \right] && (20)\\
&\propto \exp\left[-\frac{1}{2\left(\frac{n}{\sigma^2} + \frac{1}{v} \right)^{-1}} \left(\theta - \frac{v\bar{x} + \frac{\sigma^2}{n} m}{v + \frac{\sigma^2}{n}} \right)^2 \right] && (21)\\
&= \exp\left[-\frac{1}{2\left(\frac{n}{\sigma^2} + \frac{1}{v} \right)^{-1}} \left(\theta - \frac{\frac{\sum_{i=1}^{N} x_i}{\sigma^2} + \frac{m}{v}}{\frac{n}{\sigma^2} + \frac{1}{v}} \right)^2 \right] && (22)\\
&= \exp\left[-\frac{1}{2\left(\frac{1}{v} + \frac{n}{\sigma^2} \right)^{-1}} \left(\theta - \frac{\frac{m}{v} + \frac{\sum_{i=1}^{N} x_i}{\sigma^2}}{\frac{1}{v} + \frac{n}{\sigma^2}} \right)^2 \right], && (23)
\end{aligned}
$$

so

$$
\theta|x_1, \ldots, x_2 \sim \mathcal{N}\left(\frac{\frac{m}{v} + \frac{\sum_{i=1}^{N} x_i}{\sigma^2}}{\frac{1}{v} + \frac{n}{\sigma^2}}, \left[\frac{1}{v} + \frac{n}{\sigma^2} \right]^{-1} \right). \tag{24}
$$

(D) $X_i|\sigma^2 \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma^2)$, $i = 1, 2, \ldots, N$ where θ is *known* and $\sigma^2 \sim \text{IG}(a, b)$ is *unknown*. Let $w = \sigma^{-2}$ so

$w \sim \text{Gamma}(a, b)$. The posterior distribution of w given $x_1, \dots, x_N$ is

$$\begin{aligned}
f(w|x_1, \dots, x_2) &\propto f(x_1, \dots, x_2|w) f(w) && (25)\\
&\propto \left(\prod_{i=1}^{N} w^{1/2} \exp\left[-\frac{w}{2}(x_i - \theta)^2 \right] \right) w^{a-1} \exp(-bw) && (26)\\
&= w^{n/2} \exp\left[-\frac{w}{2} \sum_{i=1}^{N} (x_i - \theta^2) \right] w^{a-1} \exp(-bw) && (27)\\
&= w^{a+n/2-1} \exp\left[-\left(b + \frac{\sum_{i=1}^{N}(x_i - \theta^2)}{2} \right) w \right], && (28)
\end{aligned}$$

so

$$\begin{aligned}
w|x_1, \dots, x_2 &\sim \text{Gamma}\left(a + \frac{n}{2}, b + \frac{\sum_{i=1}^{N}(x_i - \theta^2)}{2} \right) && (29)\\
\sigma^2|x_1, \dots, x_2 &\sim \text{IG}\left(a + \frac{n}{2}, b + \frac{\sum_{i=1}^{N}(x_i - \theta^2)}{2} \right) && (30)
\end{aligned}$$

(E) $X_i \sim \mathcal{N}(\theta, \sigma_i^2)$, $i = 1, 2, \dots, n$ where each $X_i \perp\!\!\!\perp X_j, i \neq j$ is observed once and has a *known* unique variance σ_i^2 and $\theta \sim \mathcal{N}(m, v)$ is *unknown*. The posterior distribution of θ is

$$\begin{aligned}
f(\theta|x_1, \dots, x_N) &\propto f(x_1, \dots, x_N|\theta) f(\theta) && (31)\\
&\propto \left(\prod_{i=1}^{N} \exp\left[-\frac{(x_i - \theta)^2}{2\sigma_i^2} \right] \right) \exp\left[-\frac{(\theta - m)^2}{2v} \right] && (32)\\
&= \exp\left[-\frac{1}{2} \left(\sum_{i=1}^{n} \frac{(\theta - x_i)^2}{\sigma_i^2} + \frac{(\theta - m)^2}{v} \right) \right] && (33)\\
&\propto \exp\left[-\frac{1}{2} \left(\sum_{i=1}^{N} \frac{1}{\sigma_i^2} \cdot \theta^2 - 2 \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \cdot \theta + \frac{1}{v}\theta^2 - 2\frac{m}{v}\theta \right) \right] && (34)\\
&= \exp\left[-\frac{1}{2} \left(\left[\frac{1}{v} + \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \right] \theta^2 - 2 \left[\frac{m}{v} + \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \right] \theta \right) \right] && (35)\\
&= \exp\left[-\frac{1}{2\left(\frac{1}{v} + \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \right)^{-1}} \left(\theta^2 - 2 \left[\frac{\frac{m}{v} + \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2}}{\frac{1}{v} + \sum_{i=1}^{N} \frac{1}{\sigma_i^2}} \right] \theta \right) \right] && (36)\\
&\propto \exp\left[-\frac{1}{2\left(\frac{1}{v} + \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \right)^{-1}} \left(\theta - \frac{\frac{m}{v} + \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2}}{\frac{1}{v} + \sum_{i=1}^{N} \frac{1}{\sigma_i^2}} \right)^2 \right], && (37)
\end{aligned}$$

so,

$$\theta|x_1, \dots, x_N \sim \mathcal{N}\left(\frac{\frac{m}{v} + \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2}}{\frac{1}{v} + \sum_{i=1}^{N} \frac{1}{\sigma_i^2}}, \left(\frac{1}{v} + \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \right)^{-1} \right). \qquad (38)$$

(F) $X|\sigma^2 \sim \mathcal{N}(0, \sigma^2)$, $w = \frac{1}{\sigma^2} \sim \text{Gamma}(a, b)$. The marginal distribution of X is

$$
\begin{aligned}
f(x) &= \int_0^\infty f(x, w) dw && (39)\\
&= \int_0^\infty f(x|w) f(w) dw && (40)\\
&\propto \int_0^\infty w^{1/2} \exp\left(-\frac{w}{2}x^2\right) w^{a-1} \exp\left(-bw\right) dw && (41)\\
&= \int_0^\infty w^{a-1/2} \exp\left[-\left(b + \frac{x^2}{2}\right) w\right] dw \quad * \text{kernel of Gamma} \left(a + \frac{1}{2}, b + \frac{x^2}{2}\right) && (42)\\
&= \frac{\Gamma\left(a + \frac{1}{2}\right)}{\left(b + \frac{x^2}{2}\right)^{a+1/2}} && (43)
\end{aligned}
$$

Problem 2

The multivariate normal distribution

Basics

(A) Here we prove two properties of the covariance of a vector of random variables. First, note that $E(Ax + b) = A\mu + b$.

1.

$$
\begin{aligned}
\text{cov}(x) &= E\left((x - \mu)(x - \mu)^T\right) && (44)\\
&= E\left((x - \mu)(x^T - \mu^T)\right) && (45)\\
&= E\left(xx^T - x\mu^T - \mu x^T + \mu\mu^T\right) && (46)\\
&= E(xx^T) - E(x)\mu^T - \mu E(x^T) + \mu\mu^T && (47)\\
&= E(xx^T) - \mu\mu^T - \mu\mu^T + \mu\mu^T && (48)\\
&= E(xx^T) - \mu\mu^T && (49)
\end{aligned}
$$

2.

$$
\begin{aligned}
\text{cov}(Ax + b) &= E\left((Ax + b - (A\mu + b))(Ax + b - (A\mu + b))^T\right) && (50)\\
&= E\left((Ax - A\mu)(Ax - A\mu)^T\right) && (51)\\
&= E\left((Ax - A\mu)\left(x^T A^T - \mu^T A^T\right)\right) && (52)\\
&= E\left(Axx^T A - Ax\mu^T A^T - A\mu x^T A^T + A\mu\mu^T A^T\right) && (53)\\
&= E\left(Axx^T A^T\right) - E\left(Ax\mu^T A^T\right) - E\left(A\mu x^T A^T\right) + \left(A\mu\mu^T A^T\right) && (54)\\
&= AE\left(xx^T\right) A^T - A\mu\mu^T A^T - A\mu\mu^T A^T + A\mu\mu^T A^T && (55)\\
&= AE\left(xx^T\right) A^T - A\mu\mu^T A^T && (56)\\
&= A\left(E\left(xx^T\right) - \mu\mu^T\right) A^T && (57)\\
&= A\text{cov}(x) A^T && (58)
\end{aligned}
$$

(B) Define the vector $z = (z_1, \ldots, z_p)$ where $z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1), i = 1, 2, \ldots, p$. Because each component is iid, the joint probability density function (PDF) for z is

$$f(z) = \prod_{i=1}^{p} (2\pi)^{-1/2} \exp\left(-z_i^2/2\right) \tag{59}$$
$$= (2\pi)^{-p/2} \exp\left(-z^T z/2\right). \tag{60}$$

For each component z_i, the moment generating function (MGF) is

$$M_{z_i}(t_i) = E\left(\exp\left(t_i z_i\right)\right) \tag{61}$$
$$= \int_{-\infty}^{+\infty} \exp\left(t_i z_i\right) \cdot f(z_i) dz_i \tag{62}$$
$$= \int_{-\infty}^{+\infty} \exp\left(t_i z_i\right) \cdot (2\pi)^{-1/2} \exp\left(-z_i^2/2\right) dz_i \tag{63}$$
$$= \int_{-\infty}^{+\infty} (2\pi)^{-1/2} \exp\left(-z_i^2/2 + t_i z_i\right) dz_i \tag{64}$$
$$= \exp\left(t_i^2/2\right) \int_{-\infty}^{+\infty} (2\pi)^{-1/2} \exp\left(-\left[z_i - t_i\right]^2/2\right) dz_i \tag{65}$$
$$= \exp\left(t_i^2/2\right). \tag{66}$$

The MGF for the full vector z is

$$M_z(t) = E\left(\exp\left(t^T z\right)\right) \tag{67}$$
$$= \prod_{i=1}^{p} E\left(\exp\left(t_i z_i\right)\right) \tag{68}$$
$$= \prod_{i=1}^{p} \exp\left(t_i^2/2\right) \tag{69}$$
$$= \exp\left(\sum_{i=1}^{p} t_i^2/2\right) \tag{70}$$
$$= \exp\left(t^T t/2\right) \tag{71}$$

(C) We are trying to show that $x = (x_1, \ldots, x_p)$ is a multivariate normal distribution with mean vector μ and covariance matrix Σ. Let $a^T x = z \sim \mathcal{N}(m, v)$ (z is now a scalar random variable). Let t be a scalar, a is a vector of length p, and $b = ta$ is also a vector of length p. The MGF of z is

$$M_z(t) = E\left(\exp\left(tz\right)\right) \tag{72}$$
$$= E\left(\exp\left(ta^T x\right)\right) \tag{73}$$
$$= E\left(\exp\left(bx\right)\right) \tag{74}$$
$$= M_x(b) = \exp\left(mt + vt^2/2\right), \tag{75}$$

by the MGF definition of the univariate normal distribution. We can solve for m and v in terms of μ and Σ by using the first and second moments of z. The first moment of z is equal to $E(z) = m$, and can also be expressed as

$$E(z) = E\left(a^T x\right) \tag{76}$$
$$= a^T E(x) \tag{77}$$
$$= a^T \mu = m. \tag{78}$$

Note that $m^2 = (a^T\mu)^2 = a^T\mu\mu^T a$. Next, the second moment of z is equal to $E(z^2) = \text{var}(z) + E(z)^2 = v + m^2$, which can also be expressed as

$$\begin{aligned}
E(z^2) &= E\,(z \cdot z) && (79)\\
&= E\left(a^T x x^T a\right) && (80)\\
&= a^T E(xx^T) a && (81)\\
&= a^T(\text{cov}(x) + \mu^T\mu)a && (82)\\
&= a^T(\Sigma + \mu^T\mu)a && (83)\\
&= a^T\Sigma a^+ a^T\mu^T\mu a = v + m^2 = v + a^T\mu\mu^T a && (84)\\
&\Rightarrow v = a^T\Sigma a && (85)
\end{aligned}$$

Now we return to the (75) to write the MGF of x as

$$\begin{aligned}
M_x(b) &= \exp\left(mt + vt^2/2\right) && (86)\\
&= \exp\left(ta^T\mu + t^2 a^T\Sigma a^T/2\right) && (87)\\
&= \exp\left(ta^T\mu + (ta^T)\Sigma(ta)/2\right) && (88)\\
&= \exp\left(b^T\mu + b^T\Sigma b/2\right) && (89)\\
&\text{Q.E.D.} && (90)
\end{aligned}$$

(D) The p-length vector $z \sim \mathcal{N}_p(0, I_p)$ follows the standard multivariate normal distribution. We will full prove that the vector $x = Lz + \mu$, where L is a $p \times p$ matrix of full column rank, is multivariate normal. The MGF of x is,

$$\begin{aligned}
M_x(t) &= E\left(\exp\left(t^T x\right)\right) && (91)\\
&= E\left(\exp\left(t^T(Lz + \mu)\right)\right) && (92)\\
&= E\left(\exp(t^T Lz + t^T\mu)\right) && (93)\\
&= \exp\left(t^T\mu\right) E\left(\exp(t^T Lz)\right) && (94)\\
&= \exp\left(t^T\mu\right) M_z(t^T L) && (95)\\
&= \exp\left(t^T\mu\right) \exp\left[\frac{1}{2}\left(t^T L\right) I_p \left(t^T L\right)^T\right] && (96)\\
&= \exp\left(t^T\mu + t^T LL^T t/2\right). && (97)
\end{aligned}$$

Therefore x follows a multivariate normal distribution with mean vector μ and covariance matrix LL^T, $x \sim \mathcal{N}_p\left(\mu, LL^T\right)$.

(E) By definiton, $x = Lz + \mu$ is an affine transformation of a vector of standard normal random variables, z. To generate random numbers from $x \sim \mathcal{N}_p(\mu, \Sigma)$, first perform the Cholesky decomposition of Σ to obtain a lower triangle matrix L such that $\Sigma = LL^T$, generate p iid scalar normal random numbers to make the z vector, and finally compute $x = Lz + \mu$.

(F) Before we begin, let us first show that

$$\det\left(L^{-1}\right) = [\det(\Sigma)]^{-1/2}.$$

This may be shown by

$$\begin{aligned}
L^{-1}L &= I \\
\det\left(L^{-1}L\right) &= \det(I) \\
\det\left(L^{-1}\right)\det(L) &= 1 \\
\det\left(L^{-1}\right) &= [\det(L)]^{-1}
\end{aligned}$$

and

$$\begin{aligned}
\Sigma &= LL^T \\
\det(\Sigma) &= \det\left(LL^T\right) \\
\det(\Sigma) &= \det(L)\det\left(L^T\right) \\
\det(\Sigma) &= \det(L)^2 \\
[\det(\Sigma)]^{1/2} &= \det(L) \\
[\det(\Sigma)]^{-1/2} &= [\det(L)]^{-1} \\
[\det(\Sigma)]^{-1/2} &= \det\left(L^{-1}\right).
\end{aligned}$$

Now we can derive the PDF of the multivariate normal $x \sim \mathcal{N}(\mu, \Sigma)$. Define the transformation $f : z \mapsto x$, $x = f(z) = Lz + \mu$, and its inverse transformation, $f^{-1} = g : x \mapsto z$, $z = g(x) = L^{-1}(x - \mu)$, where z follows the standard multivariate distribution. The PDF of x is

$$f_x(x) = f_z(g(x)) \cdot |J(y)|, \tag{98}$$

where $J(y)$ is the Jacobian determinant of the transformation g, which in this case is just $\det\left(L^{-1/2}\right)$,

$$\begin{aligned}
f_x(x) &= (2\pi)^{-p/2}\exp\left(-\frac{1}{2}\left[L^{-1}(x-\mu)\right]^T\left[L^{-1}(x-\mu)\right]\right)\left|\det\left(L^{-1}\right)\right| && (99) \\
&= (2\pi)^{-p/2}\exp\left(-\frac{1}{2}(x-\mu)^T\left(L^{-1}\right)^T L^{-1}(x-\mu)\right)[\det(\Sigma)]^{-1/2} && (100) \\
&= (2\pi)^{-p/2}[\det(\Sigma)]^{-1/2}\exp\left(-\frac{1}{2}(x-\mu)^T\left(L^T\right)^{-1}L^{-1}(x-\mu)\right) && (101) \\
&= (2\pi)^{-p/2}[\det(\Sigma)]^{-1/2}\exp\left(-\frac{1}{2}(x-\mu)^T\left(LL^T\right)^{-1}(x-\mu)\right) && (102) \\
&= (2\pi)^{-p/2}[\det(\Sigma)]^{-1/2}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) && (103)
\end{aligned}$$

(G) Let $x_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$ independent of $x_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$, and define $y = Ax_1 + Bx_2$. The MGFs of x_1 and x_2 are, respectively,

$$\begin{aligned}
M_{x_1}(s) &= E\left(\exp\left[s^T x_1\right]\right) = \exp\left(s^T\mu_1 + s^T\Sigma_1 s/2\right) && (104) \\
M_{x_2}(s) &= E\left(\exp\left[s^T x_2\right]\right) = \exp\left(s^T\mu_2 + s^T\Sigma_2 s/2\right). && (105)
\end{aligned}$$

Problem 2 continued on next page. . .

We will characterize y by its MGF,

$$
\begin{aligned}
M_y(t) &= E\left(\exp\left[t^T y\right]\right) && (106)\\
&= E\left(\exp\left[t^T\left(Ax_1 + Bx_2\right)\right]\right) && (107)\\
&= E\left(\exp\left[t^T Ax_1 + t^T Bx_2\right]\right) && (108)\\
&= E\left(\exp\left[t^T Ax_1\right]\exp\left[t^T Bx_2\right]\right) && (109)\\
&= E\left(\exp\left[t^T Ax_1\right]\right) E\left(\exp\left[t^T Bx_2\right]\right) \Leftarrow x_1 \perp\!\!\!\perp x_2 && (110)\\
&= M_{x_1}(A^T t) M_{x_2}(B^T t) && (111)\\
&= \exp\left(t^T A\mu_1 + t^T A\Sigma_1 A^T t/2\right)\exp\left(t^T B\mu_2 + t^T B\Sigma_2 B^T t/2\right) && (112)\\
&= \exp\left(t^T A\mu_1 + t^T A\Sigma_1 A^T t/2 + t^T B\mu_2 + t^T B\Sigma_2 B^T t/2\right) && (113)\\
&= \exp\left(t^T(A\mu_1 + B\mu_2) + t^T(A\Sigma_1 A^T + B\Sigma_2 B^T)t/2\right). && (114)
\end{aligned}
$$

Therefore, $y \sim \mathcal{N}(A\mu_1 + B\mu_2, A\Sigma_1 A^T + B\Sigma_2 B^T)$.

Conditionals and marginals

(A) Let $x \sim \mathcal{N}_p(\mu, \Sigma)$ and x_1 is a vector of the first k elements of x, and x_2 is the remaining elements of x. We can also parition μ and Σ into

$$
\mu = (\mu_1, \mu_2)^T \text{ and } \Sigma = \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix} = \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{12}^T & \Sigma_{22}\end{pmatrix}, \tag{115}
$$

where μ_1 is a vector of the first k elements of μ, μ_2 is the vector of remaining elements, Σ_{11} is a $k \times k$ matrix partition of Σ, Σ_{22} is a $(p-k) \times (p-k)$ matrix partition of Σ, Σ_{12} is a $k \times (p-k)$ matrix partition of Σ, and Σ_{21} is a $(p-k) \times k$. We know that $\Sigma_{21} = \Sigma_{12}^T$ because Σ is symmetric. Define the matrix

$$
M = \begin{pmatrix}\mathcal{I}_k & \mathcal{O}_{k\times(p-k)}\end{pmatrix}, \tag{116}
$$

where $\mathcal{I}_k$ is the $k \times k$ identity matrix, and $\mathcal{O}_{k\times(p-k)}$ is the $k \times (p-k)$ matrix of all zero elements. Then,

$$
x_1 = Mx. \tag{117}
$$

We know from the previous problem that $x_1 \sim \mathcal{N}_k(M\mu, M\Sigma M^T) = \mathcal{N}_k(\mu_1, \Sigma_{11})$. This is the marginal distribution of x_1.

(B) Let $\Omega = \Sigma^{-1}$ be the inverse covariance matrix, or precision matrix, of x, which may be partitioned in the same manner as done to the covariance matrix,

$$
\Omega = \begin{pmatrix}\Omega_{11} & \Omega_{12}\\ \Omega_{12}^T & \Omega_{22}\end{pmatrix}. \tag{118}
$$

Now we will derive each block of Ω is terms of blocks from Σ, starting with the identity

$$
\begin{aligned}
\Omega &= \Sigma^{-1} && (119)\\
\Sigma\Omega &= \mathcal{I}_p && (120)\\
\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{12}^T & \Sigma_{22}\end{pmatrix}\begin{pmatrix}\Omega_{11} & \Omega_{12}\\ \Omega_{12}^T & \Omega_{22}\end{pmatrix} &= \begin{pmatrix}\mathcal{I}_k & \mathcal{O}_{k\times(p-k)}\\ \mathcal{O}_{(p-k)\times k} & \mathcal{I}_{p-k}\end{pmatrix} && (121)\\
\begin{pmatrix}\Sigma_{11}\Omega_{11} + \Sigma_{12}\Omega_{12}^T & \Sigma_{11}\Omega_{12} + \Sigma_{12}\Omega_{22}\\ \Sigma_{12}^T\Omega_{11} + \Sigma_{22}\Omega_{12}^T & \Sigma_{12}^T\Omega_{12} + \Sigma_{22}\Omega_{22}\end{pmatrix} &= \begin{pmatrix}\mathcal{I}_k & \mathcal{O}_{k\times(p-k)}\\ \mathcal{O}_{(p-k)\times k} & \mathcal{I}_{p-k}\end{pmatrix}. && (122)
\end{aligned}
$$

From here, we have a system of equations,

$$\Sigma_{11}\Omega_{11} + \Sigma_{12}\Omega_{12}^T = \mathcal{I}_k \tag{123}$$
$$\Sigma_{11}\Omega_{12} + \Sigma_{12}\Omega_{22} = \mathcal{O}_{k\times(p-k)} \tag{124}$$
$$\Sigma_{12}^T\Omega_{11} + \Sigma_{22}\Omega_{12}^T = \mathcal{O}_{(p-k)\times k} \tag{125}$$
$$\Sigma_{12}^T\Omega_{12} + \Sigma_{22}\Omega_{22} = \mathcal{I}_{p-k}. \tag{126}$$

From (124) and we have,

$$\Sigma_{11}\Omega_{12} + \Sigma_{12}\Omega_{22} = \mathcal{O}_{k\times(p-k)} \tag{127}$$
$$\Omega_{12} = -\Sigma_{11}^{-1}\Sigma_{12}\Omega_{22} \tag{128}$$

and from (125) we have,

$$\Sigma_{12}^T\Omega_{11} + \Sigma_{22}\Omega_{12}^T = \mathcal{O}_{(p-k)\times k} \tag{129}$$
$$\Omega_{12}^T = -\Sigma_{22}^{-1}\Sigma_{12}^T\Omega_{11}. \tag{130}$$

Now, from (123),

$$\Sigma_{11}\Omega_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T\Omega_{11} = \mathcal{I}_k \tag{131}$$
$$\left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T\right)\Omega_{11} = \mathcal{I}_k \tag{132}$$
$$\Omega_{11} = \left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T\right)^{-1}, \tag{133}$$

and from (126),

$$-\Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12}\Omega_{22} + \Sigma_{22}\Omega_{22} = \mathcal{I}_{p-k} \tag{134}$$
$$\left(\Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12}\right)\Omega_{22} = \mathcal{I}_{p-k} \tag{135}$$
$$\Omega_{22} = \left(\Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1}. \tag{136}$$

We now have all the pieces to write the Ω is terms of partitions of Σ,

$$\Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{12}^T & \Omega_{22} \end{pmatrix} = \begin{pmatrix} \left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T\right)^{-1} & -\Sigma_{11}^{-1}\Sigma_{12}\left(\Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1} \\ -\Sigma_{22}^{-1}\Sigma_{12}^T\left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T\right)^{-1} & \left(\Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1} \end{pmatrix}. \tag{137}$$

(C) For convenience, define the vector m as

$$m = x - \mu \tag{138}$$
$$\begin{pmatrix} m_1 \\ m_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \tag{139}$$

Now we will find the conditional distribution of x_1, given x_2, which may be found with

$$f(x_1|x_2) = \frac{f(x_1, x_2)}{f(x_2)} \tag{140}$$
$$\log f(x_1|x_2) = \log f(x_1, x_2) - \log f(x_2). \tag{141}$$

Next, note that the joint PDF of x_1 and x_2 is

$$
\begin{aligned}
f(x_1, x_2) &= f(x) && (142)\\
&\propto \exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right] && (143)\\
&= \exp\left[-\frac{1}{2}(x-\mu)^T\Omega(x-\mu)\right]. && (144)
\end{aligned}
$$

On the log-scale, this becomes

$$
\begin{aligned}
\log f(x_1, x_2) &= -\frac{1}{2}(x-\mu)^T\Omega(x-\mu) && (145)\\
&= -\frac{1}{2}m^T\Omega m && (146)\\
&= -\frac{1}{2}\left(m_1^T m_2^T\right)\begin{pmatrix}\Omega_{11} & \Omega_{12}\\ \Omega_{12}^T & \Omega_{22}\end{pmatrix}\begin{pmatrix}m_1\\ m_2\end{pmatrix} && (147)\\
&= -\frac{1}{2}\left(m_1^T\Omega_{11}m_1 + m_2^T\Omega_{12}^T m_1 + m_1^T\Omega_{12}m_1 + m_2^T\Omega_{22}m_2\right) && (148)\\
&= -\frac{1}{2}\left(m_1^T\Omega_{11}m_1 + 2m_2^T\Omega_{12}^T m_1 + m_2^T\Omega_{22}m_2\right) && (149)\\
&= -\frac{1}{2}\left[(x_1-\mu_1)^T\Omega_{11}(x_1-\mu_1) + 2(x_2-\mu_2)^T\Omega_{12}^T(x_1-\mu_1)\right] + C && (150)\\
&= -\frac{1}{2}\left[x_1^T\Omega_{11}x_1 - 2\mu_1^T\Omega_{11}x_1 + 2(x_2-\mu_2)^T\Omega_{12}^T x_1\right] + C && (151)\\
&= -\frac{1}{2}\left(x_1^T\Omega_{11}x_1 - 2\left[\mu_1^T\Omega_{11} - (x_2-\mu_2)^T\Omega_{12}^T\right]x_1\right) + C && (152)
\end{aligned}
$$

dropping some constants C which do not contain x_1. Let $A = \Omega_{11}$ and $b^T = \mu_1^T\Omega_{11} - (x_2-\mu_2)^T\Omega_{12}^T$, so $b = \Omega_{11}\mu_1 - \Omega_{12}(x_2-\mu_2)$. Then (152) becomes

$$
\begin{aligned}
\log f(x_1, x_2) &= -\frac{1}{2}\left(x_1^T A x_1 - 2b^T x_1\right) + C && (153)\\
&= -\frac{1}{2}\left(x_1^T A x_1 - 2b^T x_1 + b^T A^{-1}b - b^T A^{-1}b\right) + C && (154)\\
&= -\frac{1}{2}\left[(x_1 - A^{-1}b)^T A(x_1 - A^{-1}b) - b^T A^{-1}b\right] + C && (155)\\
&= -\frac{1}{2}(x_1 - A^{-1}b)^T A(x_1 - A^{-1}b) + C && (156)\\
&= -\frac{1}{2}(x_1 - \Omega_{11}^{-1}[\Omega_{11}\mu_1 - \Omega_{12}(x_2-\mu_2)])^T\Omega_{11}(x_1 - \Omega_{11}^{-1}[\Omega_{11}\mu_1 - \Omega_{12}(x_2-\mu_2)]) && (157)\\
&= -\frac{1}{2}(x_1 - [\mu_1 - \Omega_{11}^{-1}\Omega_{12}(x_2-\mu_2)])^T\Omega_{11}(x_1 - [\mu_1 - \Omega_{11}^{-1}\Omega_{12}(x_2-\mu_2)]) && (158)
\end{aligned}
$$

We can see that the conditional distribution of x_1 given x_2 is

$$
x_1|x_2 \sim \mathcal{N}_k\left(\mu_1 - \Omega_{11}^{-1}\Omega_{12}(x_2-\mu_2), \Omega_{11}^{-1}\right), \tag{159}
$$

and we can simplify a bit further using the fact that the inverse of a symmetric matrix is also symmetic,

$$\begin{aligned}
\Omega_{11}^{-1}\Omega_{12} &= ((\Omega_{11}^{-1}\Omega_{12})^T)^T && (160)\\
&= (\Omega_{12}^T\Omega_{11}^{-1})^T && (161)\\
&= \left(-\Sigma_{22}^{-1}\Sigma_{12}^T\left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T\right)^{-1}\left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T\right)\right)^T && (162)\\
&= \left(-\Sigma_{22}^{-1}\Sigma_{12}^T\right)^T && (163)\\
&= -\Sigma_{12}\Sigma_{22}^{-1}, && (164)
\end{aligned}$$

so we can finally write the conditional of x_1 in terms of partitions of μ and Σ as,

$$x_1|x_2 \sim \mathcal{N}_k\left(\mu_1+\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2), \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T\right). \tag{165}$$

R code is shown on the following page.

Problem 3

Multiple regression: three classical principles for inference

(A)

$$\begin{aligned}
y_i &= x_i^T\beta+\epsilon_i, \;\; \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma^2) && (166)\\
y &= X\beta+\epsilon, \;\; \epsilon \sim \mathcal{N}_n(0,\sigma^2 I_n) && (167)
\end{aligned}$$

Least squares

$$\begin{aligned}
\hat{\beta} &= \arg\min_{\beta\in\mathcal{R}^p}\left\{\sum_{i=1}^n (y_i - x_i\beta)\right\} && (168)\\
&= \arg\min_{\beta\in\mathcal{R}^p}\left\{(y-X\beta)^T(y-X\beta)\right\} && (169)\\
&= \arg\min_{\beta\in\mathcal{R}^p}\left\{\frac{1}{2}(X\beta-y)^T(X\beta-y)\right\} && (170)\\
&= \arg\min_{\beta\in\mathcal{R}^p}\left\{\frac{1}{2}\beta^TX^TX\beta-\beta^TX^Ty+\frac{1}{2}y^Ty\right\} && (171)\\
&= \arg\min_{\beta\in\mathcal{R}^p}\left\{\frac{1}{2}\beta^TX^TX\beta-\beta^TX^Ty\right\}. && (172)
\end{aligned}$$

Now we find the gradient with respect to β of the objective and set it to zero

$$\begin{aligned}
\nabla_\beta\left(\frac{1}{2}\beta^TX^TX\beta-\beta^TX^Ty\right) &= X^TX\hat{\beta}-X^Ty = 0 && (173)\\
&\Rightarrow \hat{\beta} = (X^TX)^{-1}X^Ty && (174)
\end{aligned}$$

Maximum likelihood under Gaussianity

$$\hat{\beta} = \arg\max_{\beta \in \mathcal{R}^p} \left\{ \prod_{i=1}^{n} p(y_i|\beta, \sigma^2) \right\} \tag{175}$$

$$= \arg\max_{\beta \in \mathcal{R}^p} \left\{ \prod_{i=1}^{n} \exp\left[-\frac{1}{2\sigma^2}(y_i - x_i^T\beta)^2 \right] \right\} \tag{176}$$

$$= \arg\max_{\beta \in \mathcal{R}^p} \left\{ \sum_{i=1}^{n} -\frac{1}{2}(y_i - x_i^T\beta)^2 \right\} \tag{177}$$

$$= \arg\min_{\beta \in \mathcal{R}^p} \left\{ \sum_{i=1}^{n} (y_i - x_i^T\beta)^2 \right\}, \tag{178}$$

so we have the same solution $\hat{\beta} = (X^TX)^{-1}X^Ty$ as from the previous section.

Method of moments

Assume, without loss of generality, that the sum over all the entries in a feature of X, x_j, is $E(x_j) = 0$. Further, assume that $\bar{\epsilon} = 0$. We choose a $\hat{\beta}$ such that the sample covariance between the errors and each of the p predictors is exactly zero. For one predictor j, the sample covariance is

$$\hat{\text{cov}}(x_j, \epsilon) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(\epsilon_i - \bar{\epsilon}) \tag{179}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} x_{ij}\epsilon_i \tag{180}$$

$$= \frac{1}{n-1}x_j^T\epsilon = 0,\ j \in \{1, 2, \ldots, p\} \tag{181}$$

$$\Rightarrow X^Te = 0 \tag{182}$$

$$\Rightarrow X^T(y - X\beta) = 0 \tag{183}$$

$$\Rightarrow \hat{\beta} = (X^TX)^{-1}X^Ty \tag{184}$$

(B) Define the diagonal matrix $W = \text{diag}(w_1, \ldots, w_n)$, where each w_i is a weight associated with a given observation y_i. Now we look for the solution to the minimum weighted least squares problem,

$$\hat{\beta} = \arg\min_{\beta \in \mathcal{R}^p} \left\{ \sum_{i=1}^{n} w_i(y_i - x_i^T\beta)^2 \right\} \tag{185}$$

$$= \arg\min_{\beta \in \mathcal{R}^p} \left\{ (X\beta - y)^TW(X\beta - y) \right\} \tag{186}$$

$$= \arg\min_{\beta \in \mathcal{R}^p} \left\{ \frac{1}{2}(X\beta - y)^TW(X\beta - y) \right\} \tag{187}$$

$$= \arg\min_{\beta \in \mathcal{R}^p} \left\{ \frac{1}{2}\beta^TWX\beta - \beta^TX^TWy + y^TWy \right\} \tag{188}$$

$$= \arg\min_{\beta \in \mathcal{R}^p} \left\{ \frac{1}{2}\beta^TWX\beta - \beta^TX^TWy \right\} \tag{189}$$

From here we will take the gradient of the objective function,

$$\nabla_\beta \left(\frac{1}{2}\beta^TWX\beta - \beta^TX^TWy \right) = X^TWX\hat{\beta} - X^TWy = 0 \tag{190}$$

$$\Rightarrow \hat{\beta} = (X^TWX)^{-1}X^TWy. \tag{191}$$

Problem 3 continued on next page. . .

We can show that this is the maximum-likelihood solution under heteroscedastic Gaussian error too,

$$\hat{\beta} = \arg\max_{\beta \in \mathcal{R}^p} \left\{ \prod_{i=1}^{n} p(y_i | \beta, \sigma_i^2) \right\} \tag{192}$$

$$= \arg\max_{\beta \in \mathcal{R}^p} \left\{ \prod_{i=1}^{n} \exp\left[-\frac{1}{2\sigma_i^2} (y_i - x_i^T \beta)^2 \right] \right\} \tag{193}$$

$$= \arg\max_{\beta \in \mathcal{R}^p} \left\{ \sum_{i=1}^{n} -\frac{1}{2\sigma_i^2} (y_i - x_i^T \beta)^2 \right\} \tag{194}$$

$$= \arg\min_{\beta \in \mathcal{R}^p} \left\{ \sum_{i=1}^{n} \frac{1}{\sigma_i^2} (y_i - x_i^T \beta)^2 \right\}, \tag{195}$$

with the relation of $w_i = \sigma_i^{-2}$. In other words, each observation is weighted by the precision of its residual.

Problem 4

Quantifying uncertainty: some basic ideas

In linear regression

(A) As before, we assume

$$y = X\beta + \epsilon, \;\; \epsilon \sim \mathcal{N}(0, \sigma^2 I)$$

, so $y \sim \mathcal{N}(X\beta, \sigma^2 I)$. Our estimate $\hat{\beta} = (X^T X)^{-1} X^T y$ is a transformation of a multivariate normally distributed variable, y, so that means that $\hat{\beta}$ is also normally distributed, specifically,

$$\begin{aligned} \hat{\beta} &\sim \mathcal{N}((X^T X)^{-1} X^T X\beta, (X^T X)^{-1} X^T (\sigma^2 I)((X^T X)^{-1} X^T)^T) && (196) \\ &\sim \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1}) && (197) \\ &\sim \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1}) && (198) \end{aligned}$$

(B) We can estimate σ^2 with an average, taking into account the degrees of freedom $n - p$ after estimating p parameters,

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} (y_i - X\hat{\beta})^2. \tag{199}$$

Check the appendix for R code for implementing a linear model for the ozone dataset.

Propogating uncertainty

Now we try to estimate the covariance matrix of the sampling distribution of $\hat{\theta}$:

$$\hat{\Sigma} \approx \text{cov} = E\left\{ (\hat{\theta} - \bar{\theta})(\hat{\theta} - \bar{\theta})^T \right\} \tag{200}$$

(A) Define the function

$$\begin{aligned} f(\theta) &= \theta_1 + \theta_2 && (201) \\ f(\hat{\theta}) &= \hat{\theta}_1 + \hat{\theta}_2. && (202) \end{aligned}$$

We can calculate the standard error of $f(\hat{\theta})$ with

$$\begin{aligned} \left(\text{SE}(f(\hat{\theta}))\right)^2 &= \text{var}\left(f(\hat{\theta})\right) && (203) \\ &= \text{var}\left(\hat{\theta}_1 + \hat{\theta}_2\right) && (204) \\ &= \text{var}(\hat{\theta}_1) + \text{var}(\hat{\theta}_2) + 2\text{cov}(\hat{\theta}_1, \hat{\theta}_2). && (205) \end{aligned}$$

More generally, if we have a function which is a summation of p components of θ,

$$g(\theta) = \sum_{i=1}^{p} \theta_i, \tag{206}$$

then the standard error of $g(\hat{\theta})$ will be

$$\begin{aligned} \left(\text{SE}(g(\hat{\theta}))\right)^2 &= \text{var}\left(g(\hat{\theta})\right) && (207) \\ &= \sum_{i=1}^{p} \text{var}(\hat{\theta}_i) + 2\sum_{i<j} \text{cov}(\hat{\theta}_i, \hat{\theta}_j). && (208) \end{aligned}$$

(B) Now consider some nonlinear function $f(\theta)$. First, write the first-order Taylor approximation,

$$f(\hat{\theta}) = f(\theta) + f'(\theta)(\hat{\theta} - \theta) + \mathcal{O}((\hat{\theta} - \theta)^2) \tag{209}$$
$$\text{var}\left\{f(\hat{\theta}\right\} \approx \text{var}\left\{f(\theta) + f'(\theta)(\hat{\theta} - \theta)\right\} \tag{210}$$
$$= (f'(\theta))^2 \cdot \text{var}(\hat{\theta}) \tag{211}$$

Bootstrapping

(A) Let $\hat{\Sigma}$ denote the covariance matrix of the sampling distribution of $\hat{\beta}$. There are two ways which we may estimate $\hat{\Sigma}$ via the bootstrap. Method 1 samples the residuals after estimating the OLS β with replacement, and Method 2 samples points (x_i, y_i) with replacement.

1. Calculate $\hat{\beta} = \arg\min_\beta RSS = (X^T X)X^T y$, then calculate the residual vector $\hat{\epsilon} = y - X\hat{\beta}$. Sample n times with replacement from the empirical distribution of $\hat{\epsilon}$, each time yielding ϵ_i^* and then calculate

$$y_i^* = x_i^T \hat{\beta} + \epsilon_i^*.$$

Each bootstrap simluation yields $\hat{\beta}^* = \arg\min_\beta (y^* - X\beta)^T (y^* - X\beta)$. Compute B simluations of $\hat{\beta}^*$, and from this we can estimate $\hat{\Sigma}$.

2. For each bootstrap simulation, sample with replacement n pairs of (x_i, y_i) to give X_* and y_*. Then calculate each $\hat{\beta}^* = \arg\min_\beta (y_* - X_*\beta)^T (y_* - X_*\beta)$, compute B simluations of $\hat{\beta}^*$, and from this we can estimate $\hat{\Sigma}$.

Here are the results of these two methods, along with the parametric estimate:

	int	V5	V6	V7	V8	V9	V10	V11	V12	V13
int	1.39e+03	-2.62e-01	-1.88e+00	-1.40e-01	3.75e-01	1.46e+00	1.04e-03	-4.87e-02	4.03e-01	-3.26e-03
V5	-2.62e-01	4.98e-05	3.46e-04	2.16e-05	-5.88e-05	-2.63e-04	-3.91e-07	7.96e-06	-1.26e-04	-5.63e-07
V6	-1.88e+00	3.46e-04	2.84e-02	-6.02e-04	-1.26e-03	-1.66e-03	-5.03e-06	-1.32e-04	-1.90e-04	-1.04e-04
V7	-1.40e-01	2.16e-05	-6.02e-04	5.31e-04	5.12e-05	-1.67e-04	5.94e-07	-1.79e-04	-1.55e-04	3.88e-05
V8	3.75e-01	-5.88e-05	-1.26e-03	5.12e-05	4.70e-03	-3.39e-03	-4.59e-06	-5.04e-04	-1.78e-03	-2.24e-05
V9	1.46e+00	-2.63e-04	-1.66e-03	-1.67e-04	-3.39e-03	1.47e-02	-1.99e-05	8.28e-05	-8.46e-03	8.61e-05
V10	1.04e-03	-3.91e-07	-5.03e-06	5.94e-07	-4.59e-06	-1.99e-05	1.45e-07	1.05e-06	3.68e-05	-1.61e-07
V11	-4.87e-02	7.96e-06	-1.32e-04	-1.79e-04	-5.04e-04	8.28e-05	1.05e-06	2.11e-04	5.70e-04	-1.85e-06
V12	4.03e-01	-1.26e-04	-1.90e-04	-1.55e-04	-1.78e-03	-8.46e-03	3.68e-05	5.70e-04	1.35e-02	-1.74e-05
V13	-3.26e-03	-5.63e-07	-1.04e-04	3.88e-05	-2.24e-05	8.61e-05	-1.61e-07	-1.85e-06	-1.74e-05	2.27e-05

Table 1: Estimated covariance matrix of sampling distribution of $\hat{\beta}$ using Method 1 (sampling residuals)

	int	V5	V6	V7	V8	V9	V10	V11	V12	V13
int	1.27e+03	-2.41e-01	-2.38e+00	-3.17e-02	4.57e-01	1.55e+00	1.10e-03	-7.78e-02	1.61e-01	3.62e-04
V5	-2.41e-01	4.62e-05	4.45e-04	-9.17e-07	-7.81e-05	-2.73e-04	-4.28e-07	1.44e-05	-8.82e-05	-7.27e-07
V6	-2.38e+00	4.45e-04	2.55e-02	-4.63e-04	-1.99e-03	-2.53e-03	-1.75e-06	-1.98e-04	2.86e-04	-1.58e-04
V7	-3.17e-02	-9.17e-07	-4.63e-04	5.92e-04	1.73e-04	-6.30e-04	2.55e-06	-2.54e-04	4.46e-04	1.57e-05
V8	4.57e-01	-7.81e-05	-1.99e-03	1.73e-04	4.36e-03	-2.16e-03	-3.89e-06	-4.99e-04	-2.26e-03	1.05e-05
V9	1.55e+00	-2.73e-04	-2.53e-03	-6.30e-04	-2.16e-03	1.34e-02	-2.11e-05	2.44e-04	-8.45e-03	5.58e-05
V10	1.10e-03	-4.28e-07	-1.75e-06	2.55e-06	-3.89e-06	-2.11e-05	1.41e-07	4.83e-07	3.79e-05	-8.39e-08
V11	-7.78e-02	1.44e-05	-1.98e-04	-2.54e-04	-4.99e-04	2.44e-04	4.83e-07	2.39e-04	3.54e-04	4.36e-06
V12	1.61e-01	-8.82e-05	2.86e-04	4.46e-04	-2.26e-03	-8.45e-03	3.79e-05	3.54e-04	1.39e-02	-3.22e-05
V13	3.62e-04	-7.27e-07	-1.58e-04	1.57e-05	1.05e-05	5.58e-05	-8.39e-08	4.36e-06	-3.22e-05	1.59e-05

Table 2: Estimated covariance matrix of sampling distribution of $\hat{\beta}$ using Method 2 (sampling points)

	int	V5	V6	V7	V8	V9	V10	V11	V12	V13
int	1.47e+03	-2.77e-01	-2.06e+00	-1.53e-01	3.59e-01	1.59e+00	1.11e-03	-4.08e-02	4.19e-01	-3.04e-03
V5	-2.77e-01	5.26e-05	3.78e-04	2.37e-05	-5.51e-05	-2.87e-04	-4.22e-07	6.40e-06	-1.33e-04	-6.77e-07
V6	-2.06e+00	3.78e-04	3.03e-02	-5.69e-04	-1.20e-03	-2.16e-03	-4.70e-06	-2.00e-04	-4.36e-05	-1.16e-04
V7	-1.53e-01	2.37e-05	-5.69e-04	5.65e-04	2.82e-05	-1.53e-04	6.20e-07	-1.85e-04	-1.66e-04	4.25e-05
V8	3.59e-01	-5.51e-05	-1.20e-03	2.82e-05	4.80e-03	-3.52e-03	-4.67e-06	-5.09e-04	-1.81e-03	-2.32e-05
V9	1.59e+00	-2.87e-04	-2.16e-03	-1.53e-04	-3.52e-03	1.56e-02	-2.11e-05	7.94e-05	-8.97e-03	9.48e-05
V10	1.11e-03	-4.22e-07	-4.70e-06	6.20e-07	-4.67e-06	-2.11e-05	1.56e-07	1.14e-06	3.91e-05	-1.71e-07
V11	-4.08e-02	6.40e-06	-2.00e-04	-1.85e-04	-5.09e-04	7.94e-05	1.14e-06	2.18e-04	6.02e-04	-2.81e-06
V12	4.19e-01	-1.33e-04	-4.36e-05	-1.66e-04	-1.81e-03	-8.97e-03	3.91e-05	6.02e-04	1.42e-02	-2.21e-05
V13	-3.04e-03	-6.77e-07	-1.16e-04	4.25e-05	-2.32e-05	9.48e-05	-1.71e-07	-2.81e-06	-2.21e-05	2.40e-05

Table 3: Parametric estimated covariance matrix of sampling distribution of $\hat{\beta}$

Problem 4 continued on next page. . .

(B)

R code for `myfuns.R`

```
###########################################################
######### Created by Spencer Woody on 31 Jan 2017 #########
###########################################################

my.lm <- function(X, y) {
    #  Custom function for linear regression
    #
    #  Note: this function assumes that X already has an intercept term
    # (or doesn't, if we want to force OLS through the origin)
    #
    # INPUTS:
    # X is the design matrix
    # y is the response vector
    #
    # OUTPUTS
    # a list of...
    #  Beta.hat is a vector of estimates of the coefficients
    #  Beta.SE is a vector of the standard errors of the coefficients
    #  Beta.t is a vector of t-scores of the coefficients
    #  Beta.p is the p-value for each coefficient
    #  RSS is the residual sum of squares
    #  Var.hat is the estimated variance of homoscedastic residuals
    #  R.sq is the R-squared value
    #  R.sqadj is the adjusted R-squared value
    #

    N <- nrow(X)
    p <- ncol(X)

    XtX <- crossprod(X)

    # Calculate beta.hat
    beta.hat <- solve(XtX, crossprod(X, y))

    # Calculate predicted values and residuals
    y.hat <- crossprod(t(X), beta.hat)
    res <- y - y.hat

    rss <- sum(res^2)

    # Calculate \hat{sigma^2}
    var.hat <- rss / (N - p)

    # Calculate covariance matrix of beta and SE's of beta
    var.beta <- var.hat * solve(XtX)
    beta.SE <- diag(var.beta) ^ 0.5

    # Calculate t-score of each beta
    beta.t <- beta.hat / beta.SE

    # Calculate p-values for coefficients
    beta.p <- 2 * (1 - pt(abs(beta.t), N - p))
```

```
    # Calculate r-squared and adjusted r-squared
    r.sq <- 1 - rss / sum((y - mean(y))^2)
    r.sqadj <- r.sq - (1 - r.sq) * (p - 1) / (N - p - 2)

    # Create a list of calculated values, return it back
    mylist <- list(Beta.hat = beta.hat, Beta.SE = beta.SE,
                   Beta.t = beta.t, Beta.p = beta.p, RSS = rss, Var.hat = var.hat,
                   R.sq = r.sq, R.sqadj = r.sqadj, Res = res)
    return(mylist)
}

my.boot1 <- function(X, y, B = 10000){
    #  Give bootstrapped estimate of covariance matrix of betas by
    #  SAMLING **RESIDUALS**
    #  Note: this function assumes that X already has an intercept term
    # (or doesn't, if we want to force OLS through the origin)
    #
    # INPUTS:
    # X is the design matrix
    # y is the response vector
    # N is the number of bootstrap simulations
    #
    # OUTPUT:
    # cov.star is the estimated covariance matrix of beta-hat
    #
    #
    #

    N <- nrow(X)
    p <- ncol(X)

    XtX    <- crossprod(X)
    XtXinv <- solve(XtX)

    # Calculate beta.hat
    beta.hat <- solve(XtX, crossprod(X, y))

    # Calculate predicted values and residuals
    y.hat <- crossprod(t(X), beta.hat)
    res <- y - y.hat

    # Run bootstrap
    beta.star <- matrix(nrow = B, ncol = p)

    for(i in 1:B) {
        sample.i <- sample(1:N, N, replace = T)
        res.star <- res[sample.i]

        y.star <- y.hat + res.star

        beta.star[i, ] <- crossprod(XtXinv, crossprod(X, y.star))
    }
```

```r
    cov.star <- cov(beta.star)

    return(cov.star)
}

my.boot2 <- function(X, y, B = 10000){
    #  Give bootstrapped estimate of covariance matrix of betas by
    #  SAMLING **POINTS x & y**
    #  Note: this function assumes that X already has an intercept term
    # (or doesn't, if we want to force OLS through the origin)
    #
    # INPUTS:
    # X is the design matrix
    # y is the response vector
    # N is the number of bootstrap simulations
    #
    # OUTPUT:
    # cov.star is the estimated covariance matrix of beta-hat
    #
    #
    #

    N <- nrow(X)
    p <- ncol(X)

    # Run bootstrap
    beta.star <- matrix(nrow = B, ncol = p)

    for(i in 1:B) {
        sample.i <- sample(1:N, N, replace = T)

        X.star <- X[sample.i, ]
        y.star <- y[sample.i, ]

        XtX.star <- crossprod(X.star)

        beta.star[i, ] <- solve(XtX.star, crossprod(X.star, y.star))
    }

    cov.star <- cov(beta.star)

    return(cov.star)
}


loglik <- function(X = NULL, y = NULL, params = NULL) {
    return(TRUE)
}
```

R code for exercises01.R

```r
############################################################
######### Created by Spencer Woody on 29 Jan 2017 #########
############################################################

library(microbenchmark)
library(ggplot2)
library(mlbench)
library(xtable)

source("myfuns.R")

############################################################
#### Linear regression

# Import the data and remove missing values
ozone = data(Ozone, package='mlbench')
ozone = na.omit(Ozone)[,4:13]

# Create response vector and design matrix (with intercept)
y <- as.matrix(ozone[,1])
X <- as.matrix(ozone[,2:10])

N   <- nrow(X)
int <- rep(1, N)
X   <- cbind(int, X)

microbenchmark(
    model1 <- lm(formula = y ~ X - 1),
    model2 <- my.lm(X, y)
    )
# my code runs about six times as fast :)

summary(model1)

model2$Beta.hat
model2$Beta.SE
model2$Beta.t
model2$Beta.p

############################################################
#### Bootstrapping

# Bootstrap estimate of covariance matrix of sampling distribution
# of betahat, resampling residuals
my.cov1 <- my.boot1(X, y)
xtable(my.cov1, display = rep("e", 11), digits = 2)

# Bootstrap estimate of covariance matrix of sampling distribution
# of betahat, resampling pairs x and y
my.cov2 <- my.boot2(X, y)
xtable(my.cov2, display = rep("e", 11), digits = 2)
```

```r
# Parametric estimate of covariance matrix of sampling distribution of betahat
cov.para <- model2$Var.hat * solve(crossprod(X))
xtable(cov.para, display = rep("e", 11), digits = 2)
```