**Bayesian inference in simple conjugate families**

**A)** Let $\mathbf{x} := \langle x_1, \ldots, x_N \rangle$, and let $k$ denote the number of successes in $\mathbf{x}$. Then

$$\begin{aligned} p(w|\mathbf{x}) &\propto p(\mathbf{x}|w)p(w) \\ &\propto w^k(1-w)^{N-k} w^{a-1}(1-w)^{b-1} \\ &= w^{a+k-1}(1-w)^{b+(n-k)-1}, \end{aligned}$$

which is the kernel of a Beta distribution with parameters $a+k-1$ and $b+(n-k)-1$. Thus, the posterior distribution is $\text{Beta}\big(a+k-1, b+(n-k)-1\big)$.

**B)** Let

$$Y_1 = \frac{X_1}{X_1 + X_2} \qquad \text{and} \qquad Y_2 = X_1 + X_2.$$

Then, solving for $X_1$ and $X_2$ in terms of $Y_1$ and $Y_2$, we get

$$X_1 = Y_1 Y_2 \qquad \text{and} \qquad X_2 = (1 - Y_1)Y_2.$$

Since we want to find the joint density of $Y_1$ and $Y_2$, we start by finding the Jacobian of the transformation:

$$\begin{aligned} J &= \begin{vmatrix} \partial x_1/\partial y_1 & \partial x_1/\partial y_2 \\ \partial x_2/\partial y_1 & \partial x_2/\partial y_2 \end{vmatrix} \\ &= \begin{vmatrix} y_2 & y_1 \\ -y_2 & 1-y_1 \end{vmatrix} \\ &= y_2(1-y_1) + y_1 y_2 \\ &= y_2. \end{aligned}$$

Note that since $X_1$ and $X_2$ are never negative, neither is $Y_2$, and therefore $|J| = J$. We can now express the joint distribution function in terms of $f_{X_1,X_2}$ as follows:

$$\begin{aligned} f_{Y_1,Y_2}(y_1, y_2) &\propto (y_1 y_2)^{a_1-1} \cdot e^{-y_1 y_2} \big((1-y_1)y_2\big)^{a_2-1} \cdot e^{-(1-y_1)y_2} \cdot y_2 \\ &= \Big(y_1^{a_1-1}(1-y_1)^{a_2-1}\Big)\Big(y_2^{a_1+a_2-1} \cdot e^{-y_2}\Big) \end{aligned}$$

where the expression in the first pair of parentheses is the kernel of a Beta distribution and that in the second is the kernel of a Gamma distribution. We therefore see that

$$y_1 \sim \text{Beta}(a_1, a_2) \qquad \text{and} \qquad y_2 \sim \text{Gamma}(a_1 + a_2, 1).$$

This provides us with a means to simulate a Beta random variable with parameters $a_1$ and $a_2$ from Gamma random variables by simulating $X_1 \sim \text{Gamma}(a_1, 1)$ and $X_2 \sim \text{Gamma}(a_2, 1)$, and then computing the ratio $X_1/(X_1 + X_2)$.

**C)** Let $X_i \sim N(\theta, \sigma^2)$ be independent where the variance is known and $\theta \sim N(m, v)$. Let $\bar{x}$ denote the mean of $x_1, \ldots, x_N$. Then

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

This last expression is complicated, but has the form $e^{-A\theta^2+B\theta}$, with which we can work until it is further simplified. Note that we still only care about things proportional to the expression involving $\theta$.

$$
\begin{aligned}
\exp\big(-A\theta^2 + B\theta\big) &= \exp\left(\frac{-\theta^2 + (B/A)\theta}{1/A}\right) \\
&\propto \exp\left(\frac{-\big(\theta - (B/2A)\big)^2}{1/A}\right)
\end{aligned}
$$

This is the kernel of a normal distribution with mean and variance as follows:

$$
\mu_{\text{post}} = \frac{B}{2A} = \frac{\dfrac{n\bar{x}}{\sigma^2} + \dfrac{m}{v}}{\dfrac{n}{\sigma^2} + \dfrac{1}{v}}
$$

$$
\sigma^2_{\text{post}} = \frac{1}{2A} = \frac{1}{\dfrac{n}{\sigma^2} + \dfrac{1}{v}}
$$

**D)** Let $X_i \sim N(\theta, \sigma^2)$ be independent where the mean $\theta$ is known and $w := 1/\sigma^2 \sim \text{Gamma}(a, b)$. Let $\hat{\sigma}^2$ denote the standard deviation of $x_1, \ldots, x_N$. Then

$$\begin{aligned}
p(w|x_1, \ldots, x_N) &\propto p(x_1, \ldots, x_N|w)p(w) \\
&= \left(\prod_{i=1}^{N} p(x_i|w)\right)p(w) \\
&\propto w^{N/2} \exp\left(\frac{-w}{2}\sum_{i=1}^{N}(x_i - \theta)^2\right) w^{a-1}\exp(-bw) \\
&= w^{a+N/2-1}\exp\big(-w \cdot N\hat{\sigma}^2/2\big)\exp(-bw) \\
&= w^{a+N/2-1}\exp\big(-(N\hat{\sigma}^2/2 + b)w\big).
\end{aligned}$$

This is the kernel of a Gamma distribution, namely $\text{Gamma}\big(a + N/2, b + N\hat{\sigma}^2/2\big)$. Thus, the posterior distribution for $\sigma^2$ is $\text{IG}\big(a + N/2, b + N\hat{\sigma}^2/2\big)$.

**E)** Let $X_i \sim N(\theta, \sigma_i^2)$ be independent where each $\sigma_i^2$ is known and $\theta \sim N(m, v)$. Then

$$\begin{aligned}
p(\theta|x_1, \ldots, x_N) &\propto p(x_1, \ldots, x_N|\theta)p(\theta) \\
&= \left(\prod_{i=1}^{N} p(x_i|\theta)\right)p(\theta) \\
&\propto \exp\left(-\sum_{i=1}^{N}\frac{(x_i - \theta)^2}{2\sigma_i^2}\right)\exp\left(\frac{-(\theta - m)^2}{2v}\right) \\
&\propto \exp\left(-\sum_{i=1}^{N}\frac{\theta^2 - 2x_i\theta}{\sigma_i^2}\right)\exp\left(\frac{-(\theta^2 - 2m\theta)}{2v}\right) \\
&= \exp\left(-\frac{\theta^2 - 2m\theta}{2v} - \sum_{i=1}^{N}\frac{\theta^2 - 2x_i\theta}{\sigma_i^2}\right) \\
&= \exp\left(-\left(\frac{1}{2v} + \sum_{i=1}^{N}\frac{1}{2\sigma_i^2}\right)\theta^2 + \left(\frac{m}{v} + \sum_{i=1}^{N}\frac{x_i}{\sigma_i^2}\right)\theta\right)
\end{aligned}$$

As in (C) above, this has the form $e^{-A\theta^2 + B\theta}$, and we solve this in the same way. Thereby we find that $\theta|x_1, \ldots, x_N$ is normally distributed with mean and variance given by

$$\mu_{\text{post}} = \frac{\dfrac{m}{v} + \displaystyle\sum_{i=1}^{N}\frac{x_i}{\sigma_i^2}}{\dfrac{1}{v} + \displaystyle\sum_{i=1}^{N}\frac{1}{\sigma_i^2}}$$

$$\sigma^2_{\text{post}} = \frac{1}{\dfrac{1}{v} + \displaystyle\sum_{i=1}^{N}\frac{1}{\sigma_i^2}}$$

**F)** Suppose that $X \sim N(0, \sigma^2)$, and that $1/\sigma^2 \sim \text{Gamma}(a, b)$. Let us find $f(x)$, the marginal distribution of $x$. As above, let $\omega = 1/\sigma^2$. Then:

$$
\begin{aligned}
f(x) &= \int_0^\infty p(x|\omega)\, p(\omega)\, d\omega \\
&\propto \int_0^\infty \omega^{1/2} e^{-\omega x^2/2} \omega^{a-1} e^{-bw}\, d\omega \\
&= \int_0^\infty \omega^{a-1/2} \exp\bigl(-(b + x^2/2)\omega\bigr)\, d\omega \qquad \text{(the integral of the kernel of a Gamma)} \\
&\propto \frac{\Gamma(a + 1/2)}{\bigl(b + x^2/2\bigr)^{a+1/2}} \\
&\propto \frac{1}{\left(1 + \dfrac{x^2}{2b}\right)^{a+1/2}} \\
&\propto \frac{1}{\left(1 + \dfrac{x^2}{(b/a)2a}\right)^{(2a+1)/2}}
\end{aligned}
$$

This is a student-$t$ distribution with $\nu = 2a$ and $\sigma^2 = b/a$.

With the alternative parameterization $x|\omega \sim N(\mu, \omega^{-1})$ and $\omega \sim \text{IG}(d/2, d\,\tau^2/2)$, we get a center $\mu$, scale $\tau$, and degrees of freedom $d$.

**The multivariate normal distribution**

*Basics*

**A)**

$$\begin{aligned}
\operatorname{cov}(x) &:= E\big((x-\mu)(x-\mu)^T\big) \\
&= E\big((x-\mu)(x^T-\mu^T)\big) \\
&= E\big(xx^T - \mu x^T - x\mu^T + \mu\mu^T\big) \\
&= E\big(xx^T\big) - E\big(\mu x^T\big) - E\big(x\mu^T\big) + E\big(\mu\mu^T\big) \\
&= E\big(xx^T\big) - \mu E\big(x^T\big) - E\big(x\big)\mu^T + \mu\mu^T E(I_N) \\
&= E\big(xx^T\big) - \mu\mu^T - \mu\mu^T + \mu\mu^T \\
&= E\big(xx^T\big) - \mu\mu^T
\end{aligned}$$

$$\begin{aligned}
\operatorname{cov}(Ax+b) &:= E\Big(\big(Ax+b-E(Ax+b)\big)\big(Ax+b-E(Ax+b)\big)^T\Big) \\
&= E\Big(\big(Ax-E(Ax)\big)\big(Ax-E(Ax)\big)^T\Big) \\
&= E\Big(A(x-\mu)\big(A(x-\mu)\big)^T\Big) \\
&= E\Big(A(x-\mu)(x-\mu)^T A^T\Big) \\
&= AE\Big((x-\mu)(x-\mu)^T\Big)A^T \\
&= A\operatorname{cov}(x)A^T
\end{aligned}$$

**B)** We want first to find the joint density function for $z$; since its components are independent, the joint density is the product of the individual densities:

$$\begin{aligned}
f(z_1,\ldots,z_p) &= \prod_{i=1}^{p} \frac{e^{-z_i^2/2}}{\sqrt{2\pi}} \\
&= \frac{\exp\Big(\sum_{i=1}^{p} -z_i^2/2\Big)}{(2\pi)^{p/2}} \\
&= \frac{\exp(-z^T z/2)}{(2\pi)^{p/2}}
\end{aligned}$$

Now let us find the moment generating function for $z$. We use the independence of the components of $z$, as well as the fact that the moment generating function of a (univariate) standard normal random variable is $\exp(t^2/2)$:

$$\begin{aligned}
M_z(t) &:= E\big(e^{t^T z}\big) \\
&= E\left(\exp\Big(\sum_{i=1}^{p} t_i z_i\Big)\right) \\
&= E\left(\prod_{i=1}^{p} e^{t_i z_i}\right) \\
&= \prod_{i=1}^{p} E\big(e^{t_i z_i}\big) \\
&= \prod_{i=1}^{p} e^{t^2/2}.
\end{aligned}$$

**C)** Suppose first that the moment generating function of $x$ has the form $\exp\left(t^T\mu + t^T\Sigma t/2\right)$. Let $a$ be a non-zero vector. Then, from Casella-Berger, for example,

$$M_{a^T x}(t) = \exp\Big((a^T t)^T\mu + (a^T t)^T\Sigma(a^T t)/2\Big)$$

which, with

$$\tilde{\mu} := (a^T t)^T\mu, \text{ and}$$

$$\tilde{\sigma}^2 := (a^T t)^T\Sigma(a^T t),$$

we recognize as a normally distributed random variable. This implies that $x$ is multivariate normal.

Now suppose that $x \sim N(\mu, \Sigma)$. We must show that its moment generating function has the desired form. Letting $a$ be any non-zero vector, we have

$$M_{a^T x}(t) = \exp\Big((a^T t)^T\mu + (a^T t)^T\Sigma(a^T t)/2\Big)$$
$$M_{a^T x}(1) = \exp\Big((a^T)^T\mu + (a^T)^T\Sigma(a^T)/2\Big).$$

We can consider this a function of $a^T$, which, after substituting the letter $t$ for $a^T$, gives us

$$M_x(t) = \exp\Big(t^T\mu + t^T\Sigma t/2\Big).$$

Thus, the moment generating function has the desired form.

**D)** Suppose that $z$ has a standard multivariate normal distribution, let $L$ be a $p \times p$ matrix of full rank, and $\mu$ a vector. Let $x := Lz + \mu$. Then for any vector $a$,

$$a^T x = a^T Lz + a^T\mu = (L^T a)^T z + a^T\mu.$$

By assumption, $(L^T a)^T z$ has a normal distribution, and therefore so does $(L^T a)^T z + a^T$. Thus, $x$ has a multivariate normal distribution. By part (A), we know that $\text{cov}(x) = L\text{cov}(z)L^T$.

**E)** Suppose that $x$ is multivariate normal. We want to show that for some $L$ and $\mu$, $x = Lz + \mu$, where $z$ is standard normal. If we can invert $L$, this is equivalent to saying that we can transform $x$ to a standard normal: $z = L^{-1}(x-\mu)$. To that end, let $\mu$ denote the mean of $x$, and let $LL^T$ be the Cholesky decomposition of $\text{cov}(x)$, noting that $L$ is invertible. Then the mean of $L^{-1}(x-\mu)$ is the zero vector, and

$$\begin{aligned}
\text{cov}\big(L^{-1}(x-\mu)\big) &= \text{cov}\big(L^{-1}x - L^{-1}\mu\big) \\
&= \text{cov}\big(L^{-1}x\big) \\
&= L^{-1}\text{cov}(x)(L^{-1})^T \\
&= L^{-1}LL^T(L^{-1})^T \\
&= L^{-1}LL^T(L^T)^{-1} \\
&= I,
\end{aligned}$$

which suffices to show that $L^{-1}(x-\mu)$ is standard normal, completing the proof.

We can therefore simulate random draws from $x \sim N(\mu, \Sigma)$ by simulating random draws from $z$ and then applying the transformation $Lz + \mu$, where $\Sigma = LL^T$.

**F)** The PDF of the standard multivariate normal distribution is

$$f_Z(z) = (2\pi)^{-p/2} \exp\bigl(-z^T z/2\bigr).$$

Suppose that $x$ is multivariate normal. Then $x = Lz + \mu$ for some invertible matrix $L$, and so $z = L^{-1}(x-\mu)$. By the change of variables formula,

$$\begin{aligned} f_X(x) &= f_Z\bigl(L^{-1}(x-\mu)\bigr) \det(J) \\ &= (2\pi)^{-p/2} \exp\Bigl(-(L^{-1}(x-\mu))^T L^{-1}(x-\mu)/2\Bigr) \det(L^{-1}) \\ &= (2\pi)^{-p/2} \det(L^{-1}) \exp\Bigl(-(x-\mu)^T \bigl((L^{-1})^T L^{-1}\bigr)(x-\mu)/2\Bigr) \end{aligned}$$

which has the required form when

$$C = (2\pi)^{-p/2} \det(L^{-1}) \qquad \text{and}$$

$$Q(x-\mu) = -(x-\mu)^T \bigl((L^{-1})^T L^{-1}\bigr)(x-\mu).$$

**G)** Let $x_1 \sim N(\mu_1, \Sigma_1)$ and $x_2 \sim N(\mu_2, \Sigma_2)$ be independent. Let $A$ and $B$ be matrices with full column rank such that $y = Ax_1 + Bx_2$. Then the MGF of $y$ is the product of two known MGFs:

$$\begin{aligned} M_y(t) &:= E\bigl(e^{t^T y}\bigr) \\ &= E\bigl(e^{t^T(Ax_1+Bx_2)}\bigr) \\ &= E\bigl(e^{t^T Ax_1} e^{t^T Bx_2}\bigr) \\ &= E\bigl(e^{t^T Ax_1}\bigr) E\bigl(e^{t^T Bx_2}\bigr) \\ &= E\bigl(e^{(A^T t)^T x_1}\bigr) E\bigl(e^{(B^T t)^T x_2}\bigr) \\ &= \exp\Bigl((A^T t)^T \mu_1 + (A^T t)^T \Sigma_1 (A^T t)/2\Bigr) \exp\Bigl((B^T t)^T \mu_2 + (B^T t)^T \Sigma_2 (B^T t)/2\Bigr) \\ &= \exp\Bigl((A^T t)^T \mu_1 + (B^T t)^T \mu_2 + (A^T t)^T \Sigma_1 (A^T t)/2 + (B^T t)^T \Sigma_2 (B^T t)/2\Bigr) \\ &= \exp\Bigl(t^T A\mu_1 + t^T B\mu_2 + t^T A\Sigma_1 A^T t/2 + t^T B\Sigma_2 B^T t/2\Bigr) \\ &= \exp\Bigl(t^T(A\mu_1 + B\mu_2) + t^T(A\Sigma_1 A^T + B\Sigma_2 B^T)t/2\Bigr) \end{aligned}$$

This is the MGF of a multivariate normal distribution; that is, $y \sim N(A\mu_1 + B\mu_2, A\Sigma_1 A^T + B\Sigma_2 B^T)$.

*Conditionals and marginals*

**A)** Let $x$, $x_1$, $\mu$, and $\Sigma$ be as given. Let $A$ denote the $k \times p$ matrix given by the partition $\big(I_{k\times k}\ 0_{k\times(p-k)}\big)$. Then $x_1 = Ax$, and we can apply the results in (G) above to get

$$\begin{aligned} x_1 &\sim N(A\mu, A\Sigma A^T) \\ &\sim N(\mu_1, \Sigma_{11}) \end{aligned}$$

**B)** We want

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}$$

for the appropriately sized identity and zero matrices. As a system of linear equations, we have

$$\begin{aligned} \Sigma_{11}\Omega_{11} + \Sigma_{12}\Omega_{21} &= I \quad \text{(i)} \\ \Sigma_{11}\Omega_{12} + \Sigma_{12}\Omega_{22} &= 0 \quad \text{(ii)} \\ \Sigma_{21}\Omega_{11} + \Sigma_{22}\Omega_{21} &= 0 \quad \text{(iii)} \\ \Sigma_{21}\Omega_{12} + \Sigma_{22}\Omega_{22} &= I \quad \text{(iv)} \end{aligned}$$

From (iii), we get

$$\Omega_{21} = -\Sigma_{22}^{-1}\Sigma_{21}\Omega_{11} \quad \text{(v)}$$

which we can combine with (i) to get

$$\begin{aligned} I &= \Sigma_{11}\Omega_{11} + \Sigma_{12}\big(-\Sigma_{22}^{-1}\Sigma_{21}\Omega_{11}\big) \\ &= \Big(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Big)\Omega_{11} \\ \Omega_{11} &= \big(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\big)^{-1} \end{aligned}$$

By symmetry, we also therefore have

$$\Omega_{22} = \big(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\big)^{-1}$$

Using this and (v), we get

$$\Omega_{21} = -\Sigma_{22}^{-1}\Sigma_{21}\big(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\big)^{-1}$$

and, again by symmetry,

$$\Omega_{12} = -\Sigma_{11}^{-1}\Sigma_{12}\big(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\big)^{-1}.$$

**C)** By definition, $f(x_1|x_2) = f(x_1, x_2)/f(x_2) = f(x)/f(x_2) \propto f(x)$, so it suffices to express $f(x)$ in terms of $x_1$ and $x_2$ and drop all terms not involving $x_1$. We do this on a log scale, as recommended, and we use the decomposition of $\Omega$ found above. Our goal is to obtain a quadratic expression in $x_1$. We use the notation $=_1$ to indicate equality up to a constant not depending on $x_1$.

$$
\begin{aligned}
2\ln\big(f(x)\big) &= (x-\mu)^T\Omega(x-\mu) \\
&= (x_1-\mu_1)^T\Omega_{11}(x_1-\mu_1) + (x_2-\mu_2)^T\Omega_{21}(x_1-\mu_1) \\
&\quad + (x_1-\mu_1)^T\Omega_{12}(x_2-\mu_2) + (x_2-\mu_2)^T\Omega_{22}(x_2-\mu_2) \\
&=_1 (x_1-\mu_1)^T\Omega_{11}(x_1-\mu_1) + 2(x_2-\mu_2)^T\Omega_{21}(x_1-\mu_1) \\
&=_1 x_1^T\Omega_{11}x_1 - x_1^T\Omega_{11}\mu_1 - \mu^T\Omega_{11}x_1 + 2(x_2-\mu_2)^T\Omega_{21}x_1 \\
&= \ x_1^T\Omega_{11}x_1 - 2\mu^T\Omega_{11}x_1 + 2(x_2-\mu_2)^T\Omega_{21}x_1 \\
&=_1 x_1^T\Omega_{11}x_1 - 2\Big(\mu^T\Omega_{11} + (x_2-\mu_2)^T\Omega_{21}\Big)x_1, \quad \text{and we complete the square to get} \\
&=_1 \Big(x_1 - \Sigma_{11}\big(\mu^T\Omega_{11} + (x_2-\mu_2)^T\Omega_{21}\big)\Big)^T\Omega_{11}\Big(x_1 - \Sigma_{11}\big(\mu^T\Omega_{11} + (x_2-\mu_2)^T\Omega_{21}\big)\Big)
\end{aligned}
$$

Thus,

$$x_1|x_2 \sim N\Big(\Sigma_{11}\big(\mu^T\Omega_{11} + (x_2-\mu_2)^T\Omega_{21}\big),\ \Omega_{11}\Big).$$

**Multiple regression: three classical principles for inference**

**A)** Let us first find the least squares estimate. We can write the quantity that we wish to minimize as

$$f(\beta) = (y - X\beta)^T (y - X\beta),$$

and we can differentiate to get

$$\frac{\partial f}{\partial \beta} = -2X^T (y - X\beta).$$

Assuming that $X^T X$ is invertible, we can ignore the $-2$, set this equal to zero, and solve for $\beta$:

$$\begin{aligned} 0 &= X^T (y - X\beta) \\ X^T y &= X^T X\beta \\ \beta &= (X^T X)^{-1} X^T y \end{aligned}$$

Thus, the least squares estimate for $\beta$ is $\hat{\beta} := (X^T X)^{-1} X^T y$.

Let us now show that the maximum likelihood estimate is the same. Given our assumption about $\varepsilon_i$, note that $p(y_i|\beta, \sigma^2)$ is normal. In matrix notation (and unnormalized), we want to maximize

$$\exp\Big((y - X\beta)^T \Sigma^{-1} (y - X\beta)/2\Big)$$

where $\Sigma$ is the covariance matrix—in this case, the diagonal matrix whose entries are $\sigma^2$. We can maximize this by maximizing the logarithm of it instead, which gives us a function of $\beta$:

$$f(\beta) = (y - X\beta)^T \Sigma^{-1} (y - X\beta)/2$$

which we differentiate and set equal to zero:

$$\begin{aligned} 0 &= X^T \big(\Sigma^{-1} + (\Sigma^{-1})^T\big)(y - X\beta) \\ \beta &= (X^T X)^{-1} X^T y \end{aligned}$$

Note that we can disregard the $\Sigma^{-1} + (\Sigma^{-1})^T$ factor, since is it corresponds to multiplication by a constant. We therefore obtain the same estimate as with least squares.

Finally, let us find the estimate using the method of moments. Here it will be convenient to assume that $\bar{x} := E(x_i) = 0$ for all $i$. We want to have $\operatorname{cov}(\varepsilon, x_j) = 0$ for all $j$, which means

$$\begin{aligned} 0 &= \frac{1}{n-1} \sum_{i=1}^{p} (\varepsilon_i - \bar{\varepsilon})(x_{ij} - \bar{x}_j) \\ &= \sum_{i=1}^{p} \varepsilon_i x_{ij} - \bar{\varepsilon} \sum_{i=1}^{p} x_{ij} - \bar{x} \sum_{i=1}^{p} \varepsilon_i + \bar{x}\bar{\varepsilon} \\ &= \sum_{i=1}^{p} \varepsilon_i x_{ij}. \end{aligned}$$

We can express this equation in matrix form as

$$0 = X^T \varepsilon = X^T (y - X\beta),$$

which yields the same solution as above.

**B)** Suppose now that we want to find the weighted least squares estimate:

$$\hat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} w_i (y_i - x_i^T \beta)^2.$$

As above, we can write this in matrix form:

$$\begin{aligned} f(\beta) &= (y - X\beta)^T W (y - X\beta) \\ &= y^T W y - y^T W X \beta - \beta^T X^T W y + \beta^T X^T W X \beta, \end{aligned}$$

where $W$ is the diagonal matrix of weights. Setting the derivative equal to zero, we get

$$\begin{aligned} 0 &= -y^T W X - X^T W y + 2X^T W X \beta \\ &= -2X^T W y + 2X^T W X \beta \\ X^T W y &= X^T W X \beta \\ \beta &= (X^T W X)^{-1} X^T W y, \end{aligned}$$

assuming that $X^T W X$ is invertible.

Now we must show that this is the same estimate as obtained using maximum likelihood under heteroscedastic Gaussian error. As in part (A), we want to maximize

$$f(\beta) = (y - X\beta)^T \Sigma^{-1} (y - X\beta)/2$$

where $\Sigma$ is the covariance matrix, which is now the diagonal matrix such that $(\Sigma)_{ii} = \sigma_i^2$. We differentiate this and set it equal to zero and solve for $\beta$ to get

$$\begin{aligned} 0 &= X^T \left(\Sigma^{-1} + (\Sigma^{-1})^T\right)(y - X\beta) \\ &= X^T \left(2\Sigma^{-1}\right)(y - X\beta) \\ &= X^T \Sigma^{-1} (y - X\beta) \\ &= X^T \Sigma^{-1} (y - X\beta) \\ X^T \Sigma^{-1} X \beta &= X^T \Sigma^{-1} y \\ \beta &= \left(X^T \Sigma^{-1} X\right)^{-1} X^T \Sigma^{-1} y. \end{aligned}$$

Thus, we get the same solution as with the weighted least squares where the weights are given by the inverse covariance matrix; that is, the weights are the precisions: $w_i = 1/\sigma_i^2$.

**Quantifying uncertainty: some basic frequentist ideas**

*In linear regression*

**A)** Given $y = X\beta + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2 I)$, we want to find the sampling distribution of $\hat{\beta}$. Since $y$ is normally distributed, $\hat{\beta} = (X^T X)^{-1} X^T y$ is a linear combination of multivariate normals, so its sampling distribution is also multivariate normal. Therefore, we need only find the mean and the covariance matrix. Note that $y \sim N(X\beta, \sigma^2 I)$, and thus, since $\hat{\beta} = (X^T X)^{-1} X^T y$, the mean of $\hat{\beta}$ is

$$(X^T X)^{-1} X^T X\beta = \beta.$$

Now we find the covariance matrix:

$$\begin{aligned}
\operatorname{cov}(\hat{\beta}) &= \operatorname{cov}\big((X^T X)^{-1} X^T y\big) \\
&= (X^T X)^{-1} X^T \operatorname{cov}(y) \big((X^T X)^{-1} X^T\big)^T \\
&= (X^T X)^{-1} X^T \sigma^2 I \big((X^T X)^{-1} X^T\big)^T \\
&= \sigma^2 (X^T X)^{-1} X^T X \big((X^T X)^{-1}\big)^T \\
&= \sigma^2 \big((X^T X)^{-1}\big)^T \\
&= \sigma^2 (X^T X)^{-1}
\end{aligned}$$

Note that in the second line, we have used the fact proven above that $\operatorname{cov}(Ax) = A\operatorname{cov}(x)A^T$. Therefore,

$$\hat{\beta} \sim N\big(\beta, \sigma^2 (X^T X)^{-1}\big).$$

**B)** We can use the sample variance

$$\hat{\sigma}^2 := \frac{1}{n - p - 1} \sum_{i=1}^{p} (y_i - \hat{y}_i)^2$$

to approximate $\sigma^2$, whence we get an estimated covariance matrix:

$$\widehat{\operatorname{cov}}(\hat{\beta}) = \hat{\sigma}^2 \big(\operatorname{cov}(\hat{\beta})\big).$$

The standard errors for each $\hat{\beta}_i$ are the diagonals. To find these errors, I added the following lines to the ozone.R script:

```
yhat <- x %*% betahat
sigmahat <- (1/(nrow(x)-ncol(x)-1))*sum((y-yhat)^2)
betacov = sigmahat*solve(t(x) %*% x)
```

The tables below compare the resulting estimated covariance matrix with the estimated matrix obtained by using `lm`.

Results from code shown above

| | V1 | V5 | V6 | V7 | V8 | V9 | V10 | V11 | V12 | V13 |
|---|---|---|---|---|---|---|---|---|---|---|
| V1 | 1476.74028716 | −0.27803845 | −2.07274385 | −0.15401978 | 0.36040930 | 1.59950229 | 0.00112043 | −0.04101012 | 0.42108269 | −0.00305420 |
| V5 | −0.27803845 | 0.00005285 | 0.00038040 | 0.00002384 | −0.00005535 | −0.00028848 | −0.00000042 | 0.00000643 | −0.00013341 | −0.00000068 |
| V6 | −2.07274385 | 0.00038040 | 0.03048385 | −0.00057225 | −0.00120683 | −0.00216768 | −0.00000472 | −0.00020066 | −0.00004378 | −0.00011709 |
| V7 | −0.15401978 | 0.00002384 | −0.00057225 | 0.00056793 | 0.00002838 | −0.00015370 | 0.00000062 | −0.00018582 | −0.00016668 | 0.00004272 |
| V8 | 0.36040930 | −0.00005535 | −0.00120683 | 0.00002838 | 0.00482742 | −0.00354290 | −0.00000469 | −0.00051214 | −0.00181817 | −0.00002328 |
| V9 | 1.59950229 | −0.00028848 | −0.00216768 | −0.00015370 | −0.00354290 | 0.01563449 | −0.00002122 | 0.00007980 | −0.00902163 | 0.00009527 |
| V10 | 0.00112043 | −0.00000042 | −0.00000472 | 0.00000062 | −0.00000469 | −0.00002122 | 0.00000016 | 0.00000115 | 0.00003933 | −0.00000017 |
| V11 | −0.04101012 | 0.00000643 | −0.00020066 | −0.00018582 | −0.00051214 | 0.00007980 | 0.00000115 | 0.00021950 | 0.00060475 | −0.00000282 |
| V12 | 0.42108269 | −0.00013341 | −0.00004378 | −0.00016668 | −0.00181817 | −0.00902163 | 0.00003933 | 0.00060475 | 0.01430462 | −0.00002221 |
| V13 | −0.00305420 | −0.00000068 | −0.00011709 | 0.00004272 | −0.00002328 | 0.00009527 | −0.00000017 | −0.00000282 | −0.00002221 | 0.00002410 |

Results from `lm`

| | V1 | V5 | V6 | V7 | V8 | V9 | V10 | V11 | V12 | V13 |
|---|---|---|---|---|---|---|---|---|---|---|
| V1 | 1469.08878309 | −0.27659784 | −2.06200424 | −0.15322175 | 0.35854189 | 1.59121471 | 0.00111462 | −0.04079763 | 0.41890091 | −0.00303838 |
| V5 | −0.27659784 | 0.00005258 | 0.00037842 | 0.00002372 | −0.00005507 | −0.00028698 | −0.00000042 | 0.00000640 | −0.00013272 | −0.00000068 |
| V6 | −2.06200424 | 0.00037842 | 0.03032590 | −0.00056928 | −0.00120058 | −0.00215645 | −0.00000470 | −0.00019962 | −0.00004355 | −0.00011648 |
| V7 | −0.15322175 | 0.00002372 | −0.00056928 | 0.00056498 | 0.00002824 | −0.00015291 | 0.00000062 | −0.00018485 | −0.00016582 | 0.00004249 |
| V8 | 0.35854189 | −0.00005507 | −0.00120058 | 0.00002824 | 0.00480241 | −0.00352454 | −0.00000467 | −0.00050949 | −0.00180875 | −0.00002316 |
| V9 | 1.59121471 | −0.00028698 | −0.00215645 | −0.00015291 | −0.00352454 | 0.01555349 | −0.00002111 | 0.00007939 | −0.00897489 | 0.00009478 |
| V10 | 0.00111462 | −0.00000042 | −0.00000470 | 0.00000062 | −0.00000467 | −0.00002111 | 0.00000016 | 0.00000114 | 0.00003913 | −0.00000017 |
| V11 | −0.04079763 | 0.00000640 | −0.00019962 | −0.00018485 | −0.00050949 | 0.00007939 | 0.00000114 | 0.00021837 | 0.00060161 | −0.00000281 |
| V12 | 0.41890091 | −0.00013272 | −0.00004355 | −0.00016582 | −0.00180875 | −0.00897489 | 0.00003913 | 0.00060161 | 0.01423050 | −0.00002209 |
| V13 | −0.00303838 | −0.00000068 | −0.00011648 | 0.00004249 | −0.00002316 | 0.00009478 | −0.00000017 | −0.00000281 | −0.00002209 | 0.00002397 |

The magnitudes of the differences are shown below.

Absolute differences

| | V1 | V5 | V6 | V7 | V8 | V9 | V10 | V11 | V12 | V13 |
|---|---|---|---|---|---|---|---|---|---|---|
| V1 | 7.65150408 | 0.00144061 | 0.01073961 | 0.00079803 | 0.00186741 | 0.00828758 | 0.00000581 | 0.00021249 | 0.00218178 | 0.00001582 |
| V5 | 0.00144061 | 0.00000027 | 0.00000197 | 0.00000012 | 0.00000029 | 0.00000149 | 0.00000000 | 0.00000003 | 0.00000069 | 0.00000000 |
| V6 | 0.01073961 | 0.00000197 | 0.00015795 | 0.00000297 | 0.00000625 | 0.00001123 | 0.00000002 | 0.00000104 | 0.00000023 | 0.00000061 |
| V7 | 0.00079803 | 0.00000012 | 0.00000297 | 0.00000294 | 0.00000015 | 0.00000080 | 0.00000000 | 0.00000096 | 0.00000086 | 0.00000022 |
| V8 | 0.00186741 | 0.00000029 | 0.00000625 | 0.00000015 | 0.00002501 | 0.00001836 | 0.00000002 | 0.00000265 | 0.00000942 | 0.00000012 |
| V9 | 0.00828758 | 0.00000149 | 0.00001123 | 0.00000080 | 0.00001836 | 0.00008101 | 0.00000011 | 0.00000041 | 0.00004674 | 0.00000049 |
| V10 | 0.00000581 | 0.00000000 | 0.00000002 | 0.00000000 | 0.00000002 | 0.00000011 | 0.00000000 | 0.00000001 | 0.00000020 | 0.00000000 |
| V11 | 0.00021249 | 0.00000003 | 0.00000104 | 0.00000096 | 0.00000265 | 0.00000041 | 0.00000001 | 0.00000114 | 0.00000313 | 0.00000001 |
| V12 | 0.00218178 | 0.00000069 | 0.00000023 | 0.00000086 | 0.00000942 | 0.00004674 | 0.00000020 | 0.00000313 | 0.00007412 | 0.00000012 |
| V13 | 0.00001582 | 0.00000000 | 0.00000061 | 0.00000022 | 0.00000012 | 0.00000049 | 0.00000000 | 0.00000001 | 0.00000012 | 0.00000012 |

These are all relative differences of 0.52%. This discrepancy was eliminated by removing the $-1$ term from the formula for $\hat{\sigma}^2$.

*Propagating uncertainty*

**A)** If we want to find the variance of the sum of two random variables, we can use the formula

$$\mathrm{Var}(\theta_1 + \theta_2) = \mathrm{Var}(\theta_1) + \mathrm{Var}(\theta_2) + 2\mathrm{cov}(\theta_1, \theta_2).$$

To extend this to the sum of all components of $\theta$, we get

$$\mathrm{Var}\left(\sum_{i=1}^{p} \theta_i\right) = \sum_{i=1}^{p} \mathrm{Var}(\theta_i) + \sum_{i \neq j} \mathrm{cov}(\theta_i, \theta_j).$$

This is the same as the sum of all entries of the covariance matrix $\Sigma$.

**B)** To approximate the variance of a non-linear function of the components $\theta_i$, we can use the delta method, which relies on the function being well-approximated by its first-order Taylor polynomial. In greater generality, we could use a higher-order Taylor polynomial for a better approximation.

*Bootstrapping*

**A)** The following code finds a bootstrap estimation of the covariance, and the results are shown in the table below.

```
library(mlbench)
ozone = data(Ozone, package='mlbench')
ozone = na.omit(Ozone)[,4:13]
y = ozone[,1]
x = as.matrix(ozone[,2:10])
x = cbind(1,x)
betahat = solve(t(x)

#X: design matrix
#y:  observations
#B: how many bootstrap iterations to use
compute_bootstrap_cov <- function(x,y,B) {
    n <- nrow(x)
    p <- ncol(x)
    betahat_boot <- matrix(nrow=B,ncol=p)
    betahat = solve(t(x)
    yhat <- x
    residuals <- y-yhat
    #create bootstrap samples for beta estimation
    for(i in 1:B) {
        boot_sample_indexes <- sample(1:n,n,replace=TRUE)
        xtemp <- x[boot_sample_indexes,]
        ytemp <- y[boot_sample_indexes]
        betahat_boot[i,] <- solve(t(xtemp)
    }
    #find covariance matrix of bootstrap samples
    return(cov(betahat_boot))
}

cov_boot <- compute_bootstrap_cov(x,y,10000)
```

Bootstrap variance approximation

| | V1 | V5 | V6 | V7 | V8 | V9 | V10 | V11 | V12 | V13 |
|---|---|---|---|---|---|---|---|---|---|---|
| V1 | 1251.64846649 | −0.23664214 | −2.44215928 | −0.02053864 | 0.43455310 | 1.54668643 | 0.00100496 | −0.08001760 | 0.11730545 | −0.00214235 |
| V5 | −0.23664214 | 0.00004525 | 0.00045440 | −0.00000302 | −0.00007361 | −0.00027219 | −0.00000041 | 0.00001465 | −0.00008045 | −0.00000027 |
| V6 | −2.44215928 | 0.00045440 | 0.02630838 | −0.00044876 | −0.00177515 | −0.00290728 | −0.00000182 | −0.00025026 | 0.00047022 | −0.00015049 |
| V7 | −0.02053864 | −0.00000302 | −0.00044876 | 0.00058862 | 0.00017420 | −0.00061511 | 0.00000258 | −0.00025084 | 0.00044579 | 0.00001632 |
| V8 | 0.43455310 | −0.00007361 | −0.00177515 | 0.00017420 | 0.00430994 | −0.00212684 | −0.00000416 | −0.00050789 | −0.00230557 | 0.00001002 |
| V9 | 1.54668643 | −0.00027219 | −0.00290728 | −0.00061511 | −0.00212684 | 0.01351445 | −0.00002100 | 0.00025828 | −0.00862845 | 0.00005335 |
| V10 | 0.00100496 | −0.00000041 | −0.00000182 | 0.00000258 | −0.00000416 | −0.00002100 | 0.00000014 | 0.00000056 | 0.00003819 | −0.00000009 |
| V11 | −0.08001760 | 0.00001465 | −0.00025026 | −0.00025084 | −0.00050789 | 0.00025828 | 0.00000056 | 0.00024282 | 0.00036042 | 0.00000529 |
| V12 | 0.11730545 | −0.00008045 | 0.00047022 | 0.00044579 | −0.00230557 | −0.00862845 | 0.00003819 | 0.00036042 | 0.01404178 | −0.00003350 |
| V13 | −0.00214235 | −0.00000027 | −0.00015049 | 0.00001632 | 0.00001002 | 0.00005335 | −0.00000009 | 0.00000529 | −0.00003350 | 0.00001650 |

**B)** The following code simulates drawing samples from a multivariate normal distribution with a given mean and covariance.

```
simulate_mvn <- function(m,v,n) {
    d <- nrow(v)
    cd <- t(chol(v))
    mvn_samp <- matrix(nrow=d,ncol=n)#store samples as rows
    for(i in 1:n){
        std_norm_samp <- rnorm(d)
        mvn_samp[,i] <- cd
    }
    return(mvn_samp)
}
```

To test this, let us take 10,000 samples from the bivariate normal distribution with mean and covariance given as follows:

$$\mu = \langle 4, 5 \rangle \qquad \Sigma = \begin{pmatrix} 5 & 8 \\ 8 & 13 \end{pmatrix}$$

Letting `mvn_samp` denote this sample, we estimate the mean and variance using MLE with the following code

```
mu_hat <- colSums(mvn_samp)/N
sigma_hat <- t(mvn_samp - mu_hat) %*% (mvn_samp - mu_hat)/N
```

This results in the following approximations:

$$\hat{\mu} = \langle 4.003, 5.012 \rangle \qquad \hat{\Sigma} = \begin{pmatrix} 5.465 & 7.912 \\ 7.912 & 13.332 \end{pmatrix}$$

The following code does a bootstrap approximation of the covariance.

```
B <- 10000
n <- nrow(mvn_samp)
p <- ncol(mvn_samp)
mvn_samp_boot <- matrix(nrow=B,ncol=p)
sigma_hat_boot <- matrix(0,nrow=p,ncol=p)
for(i in 1:B) {
    boot_sample_indexes <- sample(1:n,n,replace=TRUE)
    mvn_samp_temp <- mvn_samp[boot_sample_indexes,]
    mu_hat_temp <- colSums(mvn_samp_temp)/N
    sigma_hat_boot <- sigma_hat_boot + (t(mvn_samp_temp - mu_hat_temp)
}
sigma_hat_boot <- sigma_hat_boot/B
```

This results in the following approximation:

$$\hat{\Sigma}_{\text{boot}} = \begin{pmatrix} 5.456 & 7.898 \\ 7.898 & 13.313 \end{pmatrix}$$