A. Suppose that we take independent observations $x_1, ..., x_N$ 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. Derive the posterior distribution $p(w|x_1, ..., x_N)$.

We would like to obtain the posterior distribution for w:

$$\begin{aligned} p(w|x_1, ..., x_N) &\propto \left(\prod_{i=1}^{N} p(x_i|w) \right) p(w) \\ &\propto \left(\prod_{i=1}^{N} w^{x_i}(1-w)^{1-x_i} \right) w^{a-1}(1-w)^{b-1} \\ &\propto \underbrace{w^{\sum_{i=1}^{N} x_i + a - 1}(1-w)^{N - \sum_{i=1}^{N} x_i + b - 1}}_{\text{kernel of Beta}\left(\sum_{i=1}^{N} x_i + a, N - \sum_{i=1}^{N} x_i + b\right)} \end{aligned}$$

Therefore, the posterior distribution for w is

$$p(x_1, ..., x_N) = \text{Beta}\left(\sum_{i=1}^{N} x_i + a, N - \sum_{i=1}^{N} x_i + b \right)$$

B. Suppose that $x_1 \sim Ga(a_1, 1)$ and $x_2 \sim Ga(a_2, 1))$. Define two new random variables $y_1 = \frac{x_1}{x_1+x_2}$ and $y_2 = x_1 + x_2$. Find the joint distribution for (y_1, y_2) using a direct PDF transformation. Use this to characterize the marginals $p(y_1)$ and $p(y_2)$, and propose a method that exploits this result to simulate beta random variables, assuming you have a source of gamma random variables.

We know that

$$\begin{aligned} y_1 &= \frac{x_1}{x_1 + x_2} \\ y_2 &= x_1 + x_2 \end{aligned}$$

implies that

$$\begin{aligned} x_1 &= y_1 y_2 \\ x_2 &= y_2(1 - y_1) \\ J &= \left| \begin{bmatrix} y_2 & y_1 \\ -y_2 & 1 - y_1 \end{bmatrix} \right| \\ &= y_2 - y_1 y_2 + y_1 y_2 = y_2 \end{aligned}$$

Now, we can easily obtain the joint distribution of (y_1, y_2), using the independence of x_1 and x_2:

$$\begin{aligned} f_{y_1,y_2}(y_1, y_2) &= f_{x_1,x_2}(y_1 y_2, y_2(1 - y_1))|J| \\ &= f_{x_1}(y_1 y_2) f_{x_2}(y_2(1 - y_1)) y_2 \\ &\propto e^{-y_1 y_2}(y_1 y_2)^{a_1 - 1} e^{-y_2 + y_1 y_2}(y_2(1 - y_1))^{a_2 - 1} y_2 \\ &\propto \underbrace{y_1^{a_1 - 1}(1 - y_1)^{a_2 - 1}}_{\text{kernel of Beta}(a_1, a_2)} \quad \underbrace{y_2^{a_1 + a_2 - 1} e^{-y_2}}_{\text{kernel of Ga}(a_1 + a_2, 1)} \end{aligned}$$

Therefore, the joint distribution is

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

By definition, $y_1 \perp y_2$ (c.f. Casella & Berger; definition 4.2.5). Thus, we can easily obtain the following marginals:

$$\begin{aligned} p(y_1) &= \text{Beta}(a_1, a_2) \\ p(y_2) &= \text{Gamma}(a_1 + a_2, 1) \end{aligned}$$

Proposed Method

Assume we have gamma random variables, and we want to simulate beta random variables. So long as the rate parameter for the gamma random variables

is 1, we can take a value less than the simulated gamma random variables, divide it by the simulated gamma variables, and then we would obtain a beta random variable with shape parameters summing to the gamma's shape parameter.

C. Suppose that we take independent observations $x_1, ..., x_N$ 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 $p(\theta|x_1, ..., x_N)$.

We want to obtain the posterior distribution for θ:

$$\begin{aligned}
p(\theta|x_1, ..., x_N) &\propto \left(\prod_{i=1}^{N} p(x_i|\theta)\right) p(\theta) \\
&\propto \left(\prod_{i=1}^{N} \exp\left(-\frac{1}{2\sigma^2}(x_i-\theta)^2\right)\right) \cdot \exp\left(-\frac{1}{2v}(\theta-m)^2\right) \\
&\propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\theta)^2 - \frac{1}{2v}(\theta-m)^2\right] \\
&\propto \exp\left[-\frac{1}{2}\left(\frac{\sum_{i=1}^{N} x_i^2 - 2\theta\sum_{i=1}^{N} x_i + N\theta^2}{\sigma^2} + \frac{\theta^2 - 2m\theta}{v}\right)\right] \\
&\propto \exp\left[-\frac{1}{2}\left(-2\left(\underbrace{\frac{\sum_{i=1}^{N} x_i}{\sigma^2} + \frac{m}{v}}_{b^T}\right)\theta + \left(\underbrace{\frac{N}{\sigma^2} + \frac{1}{v}}_{A}\right)\theta^2\right)\right]
\end{aligned}$$

Therefore, we see that the posterior for θ is

$$p(\theta|x_1, ..., x_N) = \mathrm{N}\left(\left(\frac{N}{\sigma^2} + \frac{1}{v}\right)^{-1}\left(\frac{\sum_{i=1}^{N} x_i}{\sigma^2} + \frac{m}{v}\right), \left(\frac{N}{\sigma^2} + \frac{1}{v}\right)^{-1}\right)$$

D. Suppose that we take independent observations x_i from a normal sampling model with known mean θ but unknown variance σ^2. Suppose that precision parameter ω has a gamma prior with parameters a and b, implying that σ^2 has what is called an inverse-gamma prior. Derive the posterior distribution for ω. Re-express this as a posterior for σ^2, the variance.

First, we derive the posterior distribution for ω:

$$\begin{aligned} p(\omega|x_1,...,x_N) &\propto \left(\prod_{i=1}^{N} p(x_i|\omega)\right) p(\omega) \\ &\propto \prod_{i=1}^{N} \left(\frac{\omega}{2\pi}\right)^{1/2} \exp\left[-\frac{\omega}{2}(x_i-\theta)^2\right] \cdot e^{-b\omega}\omega^{a-1} \\ &\propto \underbrace{\omega^{N/2+a-1}\exp\left[-\omega\left(\frac{\sum_{i=1}^{N}(x_i-\theta)^2}{2}+b\right)\right]}_{\text{kernel of Ga}\left(\frac{N}{2}+a,\frac{\sum_{i=1}^{N}(x_i-\theta)^2}{2}+b\right)} \end{aligned}$$

Therefore, the posterior distributions for ω and σ^2 are

$$p(\omega|x_1,...,x_N) = \text{Ga}\left(\frac{N}{2}+a, \frac{\sum_{i=1}^{N}(x_i-\theta)^2}{2}+b\right)$$

$$p(\sigma^2|x_1,...,x_N) = \text{Inv-Ga}\left(\frac{N}{2}+a, \frac{\sum_{i=1}^{N}(x_i-\theta)^2}{2}+b\right)$$

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

We want to find the posterior distribution for θ:

$$\begin{aligned}
p(\theta|x_1,...,x_N) &\propto \left(\prod_{i=1}^{N} p(x_i|\theta)\right) p(\theta) \\
&\propto \left(\prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma_i^2}(x_i-\theta)^2\right]\right) \cdot \exp\left[-\frac{1}{2v}(\theta-m)^2\right] \\
&\propto \exp\left[-\frac{1}{2}\left(\sum_{i=1}^{N}\frac{(x_i-\theta)^2}{\sigma_i^2} + \frac{(\theta-m)^2}{v}\right)\right] \\
&\propto \exp\left[-\frac{1}{2}\left(\sum_{i=1}^{N}\frac{x_i^2}{\sigma_i^2} - 2\theta\sum_{i=1}^{N}\frac{x_i}{\sigma_i^2} + \theta^2\sum_{i=1}^{N}\frac{1}{\sigma_i^2} + \frac{\theta^2-2\theta m}{v}\right)\right] \\
&\propto \exp\left[-\frac{1}{2}\left(-2\left(\underbrace{\sum_{i=1}^{N}\frac{x_i}{\sigma_i^2}+\frac{m}{v}}_{b^T}\right)\theta + \theta^2\left(\underbrace{\sum_{i=1}^{N}\frac{1}{\sigma_i^2}+\frac{1}{v}}_{A}\right)\right)\right]
\end{aligned}$$

Therefore, the posterior distribution for θ is

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

Note that we can express the posterior mean as

$$\left(\sum_{i=1}^{N}\frac{1}{\sigma_i^2}+\frac{1}{v}\right)^{-1}\left(\sum_{i=1}^{N}\frac{x_i}{\sigma_i^2}+\frac{m}{v}\right) = \frac{\sum_{i=1}^{N}\left(\frac{1}{\sigma_i^2}\right)x_i + \frac{1}{v}m}{\sum_{i=1}^{N}\frac{1}{\sigma_i^2}+\frac{1}{v}},$$

which is a weighted average of the observations x_i and the prior mean m with respective weights $\frac{1}{\sigma_i^2}$ and $\frac{1}{v}$.

F. Suppose that $(x|\omega) \sim \mathbf{N}(m, \omega^{-1})$, and that $\omega \sim \mathbf{Ga}(a/2, b/2)$ prior. Show that the marginal distribution of x is Student's t with a degrees of freedom, center m, and scale parameter $(b/a)^{1/2}$. This is why the t distribution is often referred to as a scale mixture of normals.

We want to obtain the marginal data distribution:

$$\begin{aligned}
p(x) &= \int_\Omega p(x, \omega) d\omega \\
&= \int_\Omega p(x|\omega) p(\omega) d\omega \\
&= \frac{1}{\sqrt{2\pi}} \frac{(b/2)^{a/2}}{\Gamma(a/2)} \int_\Omega \underbrace{\omega^{1/2+a/2-1} \exp\left[-\omega\left(\frac{1}{2}(x-m)^2 + \frac{b}{2}\right)\right]}_{\text{kernel of Ga}\left(\frac{a+1}{2}, \frac{1}{2}(x-m)^2 + \frac{b}{2}\right)} d\omega \\
&= \frac{1}{\sqrt{2\pi}} \frac{(b/2)^{a/2}}{\Gamma(a/2)} \cdot \frac{\Gamma((a+1)/2)}{(\frac{1}{2}(x-m)^2 + b/2)^{\frac{a+1}{2}}} \\
&= \frac{\Gamma((a+1)/2)}{\Gamma(a/2)\sqrt{2\pi}} \left(\frac{b}{2}\right)^{a/2-(a+1)/2} \left(\left[\frac{1}{2}(x-m)^2 + \frac{b}{2}\right]\frac{2}{b}\right)^{-\frac{a+1}{2}} \\
&= \frac{\Gamma((a+1)/2)}{\Gamma(a/2)\sqrt{b\pi}} \left(\frac{1}{b}(x-m)^2 + 1\right)^{-\frac{a+1}{2}},
\end{aligned}$$

which uniquely characterizes this distribution as Student's t distribution with a degrees of freedom, center m, and scale parameter $(b/a)^{1/2}$.

A. The covariance matrix $\mathbf{cov}(x)$ 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, $\mathbf{cov}(x) = E[(x-\mu)(x-\mu)^T]$, where μ is the mean vector whose ith component is $E(x_i)$. Prove the following: (1) $\mathbf{cov}(x) = E(xx^T) - \mu\mu^T$; and (2) $\mathbf{cov}(Ax+b) = A\mathbf{cov}(x)A^T$ for matrix A and vector b.

First, let's obtain $\text{cov}(x)$:

$$\begin{aligned}
\text{cov}(x) &= E[(x-\mu)(x-\mu)^T] \\
&= E[(x-\mu)(x^T-\mu^T)] \\
&= E[xx^T - x\mu^T - \mu x^T + \mu\mu^T] \\
&= E[xx^T] - E[x\mu^T] - E[\mu x^T] + E[\mu\mu^T] \\
&= E[xx^T] - E[x]\mu^T - \mu E[x^T] + \mu\mu^T \\
&= E[xx^T] - \mu\mu^T - \mu\mu^T + \mu\mu^T \\
&= E[xx^T] - \mu\mu^T
\end{aligned}$$

Now, let's find the covariance of $Ax+b$:

$$\begin{aligned}
\text{cov}(Ax+b) &= E[(Ax+b-E[Ax+b])(Ax+b-E[Ax+b])^T] \\
&= E[(Ax+b-A\mu-b)(Ax+b-A\mu-b)^T] \\
&= E[(Ax-A\mu)(Ax-A\mu)^T] \\
&= E[A(x-\mu)(x^TA^T-\mu^TA^T)] \\
&= AE[(x-\mu)(x^T-\mu^T)A^T] \\
&= AE[(x-\mu)(x-\mu)^T]A^T \\
&= A\text{cov}(x)A^T
\end{aligned}$$

B. Consider the random vector $z = (z_1, ..., z_p)^T$, with each entry having an independent standard normal distribution. Derive the PDF and MGF of z, expressed in vector notation. We say that z has a standard multivariate normal distribution.

We know that

$$p(z_i) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z_i^2}{2}\right], \ \forall i \in \{1, ..., p\}.$$

Since z_i are independent, we can get the PDF of z in the following way:

$$\begin{aligned} p(z_1, ..., z_p) &= \prod_{i=1}^{p} p(z_i) \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^p \exp\left[-\frac{z_1^2}{2} - \frac{z_2^2}{2} - ... - \frac{z_p^2}{2}\right] \\ &= (2\pi)^{-p/2} \exp\left[-\frac{1}{2}(z_1^2 + z_2^2 + ... + z_p^2)\right] \\ &= (2\pi)^{-p/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{p} z_i^2\right] \end{aligned}$$

In vector notation, we can state the pdf of z as

$$p(z) = (2\pi)^{-p/2} \exp\left[-\frac{z^T z}{2}\right], \ z_i \in \mathbb{R} \ \forall i = 1, ..., p$$

We can obtain z's MGF in the following way:

$$\begin{aligned} M_z(t) &= E(e^{t^T z}) \\ &= E\left[\exp\left(t_1 z_1 + t_2 z_2 + ... + t_p z_p\right)\right] \\ &= E\left[e^{t_1 z_1} \cdot ... \cdot e^{t_p z_p}\right] \end{aligned}$$

By the independence of z_i, we then get

$$\begin{aligned} M_z(t) &= E[e^{t_1 z_1}] \cdot ... \cdot E[e^{t_p z_p}] \\ &= \prod_{i=1}^{p} E(e^{t_i z_i}) = \prod_{i=1}^{p} M_{z_i}(t) \\ &= \prod_{i=1}^{p} e^{t_i^2/2} \\ &= \exp\left[\sum_{i=1}^{p} \frac{t_i^2}{2}\right] \\ &= \exp\left[\frac{t^T t}{2}\right] \end{aligned}$$

C. A vector-valued random variable $x = (x_1, ..., x_p)^T$ has a multivariate normal distribution iff every linear combination of its components is univariate normal. That is, for all vectors a not identically zero, the scalar quantity $z = a^T x$ is normally distributed. From this definition, prove that x is multivariate normal, written $x \sim \mathbf{N}(\mu, \Sigma)$, iff its MGF is of the form $E[e^{t^T x}] = \mathbf{exp}\left[t^T\mu + t^T\Sigma t/2\right]$.

($\Rightarrow$) Let $x \sim \mathrm{N}(\mu, \Sigma)$. By linear combination of normals, $E(z) = a^T\mu$ and $\mathrm{var}(z) = a^T\Sigma a$. We know that $z = a^T x$ has the following mgf

$$\begin{aligned} M_z(\tau) &= E(e^{\tau z}) \\ &= \exp\left[E(z)\tau + \frac{1}{2}\mathrm{var}(z)\tau^2\right] \\ &= \exp\left[a^T\mu\tau + \frac{1}{2}a^T\Sigma a\tau^2\right] \\ &= \exp\left[\tau a^T\mu + \frac{1}{2}(a\tau^T)^T\Sigma a\tau^T\right], \end{aligned}$$

Now, denote $t = a\tau^T$. Then,

$$\begin{aligned} M_x(t) &= E\left[e^{t^T x}\right] \\ &= E\left[e^{\tau a^T x}\right] \\ &= E\left[e^{\tau z}\right] \\ &= M_z(\tau) \\ &= \exp\left[\tau a^T\mu + \frac{1}{2}(a\tau^T)^T\Sigma a\tau^T\right] \\ &= \exp\left[t^T\mu + \frac{1}{2}t^T\Sigma t\right] \end{aligned}$$

($\Leftarrow$) Let $M_x(t) = \exp\left[t^T\mu + \frac{t^T\Sigma t}{2}\right]$. Then,

$$\begin{aligned} M_z(t) &= E[e^{t^T z}] \\ &= E(e^{t^T a^T x}) \\ &= M_x(at) \\ &= \exp\left[(at)^T\mu + \frac{1}{2}(at)^T\Sigma(at)\right] \\ &= \exp\left[t^T a^T\mu + \frac{1}{2}t^T a^T\Sigma at\right] \end{aligned}$$

By uniqueness of MGFs,

$$z \sim \mathrm{N}(a^T\mu, a^T\Sigma a).$$

Therefore, any linear combination of x's components is univariate normal. By the given definition, x is multivariate normal.

D. Another basic theorem is that a random vector is multivariate normal iff 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 $x = Lz + \mu$ for some $p \times p$ matrix L of full-column rank. Prove that x is multivariate normal. In addition, use the moment identities you proved above to compute the expected value and covariance matrix of x.

Let $z \sim \mathrm{N}(0, I_p)$, where $x = Lz + \mu$. We know that $M_z(t) = \exp\left[\frac{t^T t}{2}\right]$. Let's use MGFs to characterize x:

$$\begin{aligned} M_x(t) &= E[e^{t^T x}] \\ &= E[e^{t^T Lz + t^T \mu}] \\ &= E[e^{t^T Lz} e^{t^T \mu}] \\ &= E[e^{t^T Lz}] \cdot E[e^{t^T \mu}] \\ &= e^{t^T \mu} M_z((t^T L)^T) \\ &= \exp\left[t^T \mu + \frac{1}{2}(t^T L)(t^T L)^T\right] \\ &= \exp\left[t^T \mu + \frac{1}{2}(t^T L L^T t)\right], \end{aligned}$$

where the penultimate equality follows from the MGF for z. Note that this MGF uniquely characterizes x, so

$$x \sim \mathrm{N}_p(\mu, LL^T),$$

with expected value μ and covariance matrix LL^T.

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. Use this insight to propose an algorithm for simulating multivariate normal random variables with a specified mean and covariance matrix.

Suppose $x \sim N_p(\mu, \Sigma)$. Since Σ is positive semi-definite, we can use spectral decomposition to obtain

$$\Sigma = P\Lambda P^T,$$

where $\Lambda = \text{diag}(\lambda_1, ..., \lambda_p)$ and P is an orthogonal matrix. Consider the affine transformation $x^* = Lz + \mu$, where z is a vector of standard normal random variables, and $L = P\Lambda^{1/2}$. Note that $r(L) = r(P\Lambda^{1/2}) = p$. From the previous problem, we know that

$$\begin{aligned} x^* &\sim \text{N}_p\left(\mu, P\Lambda^{1/2}(P\Lambda^{1/2})^T\right) \\ &\sim \text{N}_p(\mu, P\Lambda P^T) \\ &\sim \text{N}_p(\mu, \Sigma) \end{aligned}$$

Since mean and variance characterize normal distributions, $x^* = x$. Thus, we showed that we can construct a multivariate normal distribution as an affine transformation of standard normal random variables.

F. Use the previous result and the PDF of a standard multivariate normal to show that the PDF of a multivariate normal $x \sim \mathbf{N}(\mu, \Sigma)$ takes the form $p(x) = C\mathbf{exp}[-Q(x-\mu)/2]$ for some constant C and quadratic form $Q(x-\mu)$.

From the previous problem, we know how to transform a multivariate normal distribution to standard normal variables:

$$x = Lz + \mu \Leftrightarrow z = L^{-1}(x-\mu),$$

where $L = P\Lambda^{1/2}$ and $L^{-1} = (P\Lambda^{1/2})^{-1} = \Lambda^{-1/2}P^T$. Then, $z = \Lambda^{-1/2}P^T(x-\mu)$. We can obtain the PDF for x using our handy-dandy transformation method:

$$\begin{aligned}
f_x(x) &= f_z(\Lambda^{-1/2}P^T(x-\mu))|J| \\
&= \left|\Lambda^{-1/2}P^T\right|(2\pi)^{-p/2}\text{exp}\left[-\frac{1}{2}(x-\mu)^T P\Lambda^{-1/2}\Lambda^{-1/2}P^T(x-\mu)\right] \\
&= \left|\Lambda^{-1/2}\right| \cdot \left|P^T\right|(2\pi)^{-p/2}\text{exp}\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right] \\
&= |\Lambda|^{-1/2}(2\pi)^{-p/2}\text{exp}\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right] \\
&= |\Sigma|^{-1/2}(2\pi)^{-p/2}\text{exp}\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right],
\end{aligned}$$

since $|\Sigma| = |P\Lambda P^T| = |P||\Lambda||P| = |\Lambda|$. Clearly, the above is the desired PDF with

$$\begin{aligned}
C &= (2\pi)^{-p/2}|\Sigma|^{-1/2} \\
Q(x-\mu) &= (x-\mu)^T\Sigma^{-1}(x-\mu)
\end{aligned}$$

G. Let $x_1 \sim \mathbf{N}(\mu_1, \Sigma_1)$ and $x_2 \sim \mathbf{N}(\mu_2, \Sigma_2)$, where x_1 and x_2 are independent of each other. Let $y = Ax_1 + Bx_2$ for matrices A, B of full column rank and appropriate dimension. Note that x_1 and x_2 need not have the same dimension, as long as Ax_1 and Bx_2 do. Use your previous results to characterize the distribution of y.

Since there's a good chance that y will be a multivariate normal distribution, we can attempt to characterize the distribution of y with MGFs. Since $\forall a \in \mathbb{R}^p / \{0\} \implies a^T Ax_1$ and $a^T Bx_2$ are univariate normal random variables, then both Ax_1 and Bx_2 are multivariate normal. That is,

$$M_{Ax_1}(t) = \exp\left[t^T A\mu_1 + \frac{1}{2}t^T A\Sigma_1 A^T t\right]$$
$$M_{Bx_2}(t) = \exp\left[t^T B\mu_2 + \frac{1}{2}t^T B\Sigma_2 B^T t\right]$$

By independence,

$$\begin{aligned} M_{Ax_1+Bx_2}(t) &= M_{Ax_1}(t) \cdot M_{Bx_2}(t) \\ &= \exp\left[t^T(A\mu_1 + B\mu_2) + \frac{1}{2}t^T(A\Sigma_1 A^T + B\Sigma_2 B^T)t\right] \end{aligned}$$

By uniqueness of MGFs, we see that

$$y = Ax_1 + Bx_2 \sim \mathrm{N}\left(A\mu_1 + B\mu_2, A\Sigma_1 A^T + B\Sigma_2 B^T\right)$$

A. Derive the marginal distribution of x_1.

Recall our affine transformation

$$x = Lz + \mu.$$

We can write this as

$$\begin{aligned}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} &= \begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \\ &= \begin{bmatrix} L_{11}z_1 + L_{12}z_2 + \mu_1 \\ L_{21}z_1 + L_{22}z_2 + \mu_2, \end{bmatrix}\end{aligned}$$

where $L = \begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix}$ and z_1, z_2 are independent standard multivariate normal random variables. So, $x_1 = L_{11}z_1 + L_{12}z_2 + \mu_1$. We can obtain its marginal distribution with MGFS, since we have an inkling that it will be a known – most likely Normal– distribution:

$$\begin{aligned} M_{x_1}(t) &= E[e^{t^T x_1}] \\ &= E\left[\exp\left(t^T(L_{11}z_1 + L_{12}z_2 + \mu_1)\right)\right] \\ &= E\left[\exp\left(t^T L_{11}z_1\right)\exp\left(t^T L_{12}z_2\right)\exp\left(t^T\mu_1\right)\right] \\ &= E\left[\exp\left(t^T L_{11}z_1\right)\right] \cdot E\left[\exp\left(t^T L_{12}z_2\right)\right] e^{t^T\mu_1} && \text{(by ind.)} \\ &= M_{z_1}(t^T L_{11}) M_{z_2}(t^T L_{12}) e^{t^T\mu_1} && \text{(def. of MGF)} \\ &= \exp\left[\frac{1}{2}t^T L_{11}L_{11}^T t\right]\exp\left[\frac{1}{2}t^T L_{12}L_{12}^T t\right] e^{t^T\mu_1} && \text{(MGF of standard MVN)} \\ &= \exp\left[t^T\mu_1 + \frac{1}{2}t^T\left(L_{11}L_{11}^T + L_{12}L_{12}^T\right)t\right], \end{aligned}$$

which uniquely characterizes x_1 as a multivariate normal distribution with mean μ_1 and variance $L_{11}L_{11}^T + L_{12}L_{12}^T$. Note that

$$\begin{aligned}\Sigma &= LL^T \\ &= \begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix}\begin{bmatrix} L_{11}^T & L_{21}^T \\ L_{12}^T & L_{22}^T \end{bmatrix} \\ &= \begin{bmatrix} L_{11}L_{11}^T + L_{12}L_{12}^T & L_{11}L_{21}^T + L_{12}L_{22}^T \\ L_{21}L_{11}^T + L_{22}L_{12}^T & L_{21}L_{21}^T + L_{22}L_{22}^T \end{bmatrix} \\ &= \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\end{aligned}$$

Therefore, we concisely write the marginal distribution of x_1 as

$$x_1 \sim \mathrm{N}(\mu_1, \Sigma_{11})$$

B. Let $\Omega = \Sigma^{-1}$ be the inverse covariance matrix, or precision matrix, of x. Using identities for the inverse of a partitioned matrix, express each block of Ω in terms of blocks of Σ.

We know that

$$\begin{aligned}
\Sigma &= \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \\
&= \begin{bmatrix} L_{11}L_{11}^T + L_{12}L_{12}^T & L_{11}L_{21}^T + L_{12}L_{22}^T \\ L_{21}L_{11}^T + L_{22}L_{12}^T & L_{21}L_{21}^T + L_{22}L_{22}^T \end{bmatrix}, \\
\Omega &= \Sigma^{-1} \\
&= \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1}.
\end{aligned}$$

Since Σ is symmetric and positive semi-definite by virtue of it being a covariance matrix, the blocks of Σ are invertible, as well as their respective Schur complements. Therefore, we can use one of the well-established identities to obtain the inverse of a block covariance matrix:

$$\begin{aligned}
\Omega &= \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} \\
&= \begin{bmatrix} \left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)^{-1} & -\Sigma_{11}^{-1}\Sigma_{12}\left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1} \\ -\left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{21}\right)^{-1}\Sigma_{21}\Sigma_{11}^{-1} & \left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1} \end{bmatrix}
\end{aligned}$$

C. Derive the conditional distribution for x_1 given x_2 in terms of the partitioned elements of $x, \mu,$ 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. 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.

First, we "find" the joint distribution of $x = (x_1, x_2)^T$, which was already given. Then, we can ignore all factors that don't involve x_1:

$$\begin{aligned} p(x_1|x_2) &\propto p(x_1, x_2) \\ &\propto \exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right] \end{aligned}$$

Note that we can rewrite

$$\begin{aligned} &(x-\mu)^T\Sigma^{-1}(x-\mu) \\ &= \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}^T \begin{bmatrix} \left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)^{-1} & -\Sigma_{11}^{-1}\Sigma_{12}\left(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1} \\ -\left(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{21}\right)^{-1}\Sigma_{21}\Sigma_{11}^{-1} & \left(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1} \end{bmatrix} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix} \\ &= (x_1-\mu_1)^T\Omega_{11}(x_1-\mu_1) - (x_2-\mu_2)^T\Omega_{21}(x_1-\mu_1) - (x_1-\mu_1)^T\Omega_{12}(x_2-\mu_2) + (x_2-\mu_2)^T\Omega_{22}(x_2-\mu_2) \end{aligned}$$

Now, we only consider terms containing x_1. Thus,

$$\begin{aligned} p(x_1|x_2) &\propto \exp\left[-\frac{1}{2}\left((x_1-\mu_1)^T\Omega_{11}(x_1-\mu_1) - (x_2-\mu_2)^T\Omega_{21}(x_1-\mu_1) - (x_1-\mu_1)^T\Omega_{12}(x_2-\mu_2)\right)\right] \\ &\propto \exp\left[-\frac{1}{2}\left(x_1^T\Omega_{11}x_1 - 2\mu_1^T\Omega_{11}x_1 - x_2^T\Omega_{21}x_1 + \mu_2^T\Omega_{21}x_1 - x_1^T\Omega_{12}x_2 + x_1^T\Omega_{12}\mu_2\right)\right] \\ &\propto \exp\left[-\frac{1}{2}\left(x_1^T\Omega_{11}x_1 - 2\mu_1^T\Omega_{11}x_1 + 2\mu_2^T\Omega_{21}x_1 - 2x_2^T\Omega_{21}x_1\right)\right] \\ &\propto \exp\left[-\frac{1}{2}\left(-2\underbrace{\left(\mu_1^T\Omega_{11} + x_2^T\Omega_{21} - \mu_2^T\Omega_{21}\right)}_{b^T}x_1 + x_1^T\underbrace{\Omega_{11}}_{A}x_1\right)\right] \end{aligned}$$

Therefore, the conditional distribution of x_1 is

$$\begin{aligned} p(x_1|x_2) &= \mathrm{N}(A^{-1}b, A^{-1}), \\ A^{-1} &= \Omega_{11}^{-1} \\ &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}, \\ b &= \Omega_{11}^T\mu_1 + \Omega_{21}^Tx_2 - \Omega_{21}^T\mu_2, \\ A^{-1}b &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2) \end{aligned}$$

Interpretation:

This conditional distribution can be interpreted as a regression model on x_2 because we can write

$$x_1 = \underbrace{\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2}_{\text{intercept}} + \underbrace{\Sigma_{12}\Sigma_{22}^{-1}}_{\text{slope}} \cdot \underbrace{x_2}_{\text{regressor}} + \epsilon,$$

where $\epsilon \sim \mathrm{N}(0, \Omega_{11})$ is the stochastic error based on the precision matrix.

A. Show that MLE, LSE, and MoM all lead to the same estimator under SLR. What is the variance of this estimator under the assumption that each ϵ_i is independent and identically distribution with variance σ^2?

First, we consider the least squares estimator (LSE) for β:

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

which we can derive with respect to β in matrix notation as follows:

$$\begin{aligned} \frac{\partial}{\partial \beta} \left[(Y - X\beta)^T (Y - X\beta) \right] &= \frac{\partial}{\partial \beta} \left[Y^T Y - Y^T X\beta - \beta^T X^T Y + \beta^T X^T X \beta \right] \\ &= -2X^T Y + 2X^T X\beta \stackrel{set}{=} 0 \\ \hat{\beta}_{LSE} &= (X^T X)^{-1} X^T Y \end{aligned}$$

Now, we obtain the maximum likelihood estimate (MLE) by – of course – maximizing the likelihood!

$$\begin{aligned} \hat{\beta}_{MLE} &= \arg\max_{\beta \in \mathbb{R}^p} \left\{ \prod_{i=1}^{n} p(y_i | \beta, \sigma^2) \right\} \\ &= \arg\max_{\beta \in \mathbb{R}^p} \left\{ \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \right] \right\} \\ &= \arg\min_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \right\} \quad (= (1)) \\ &= \hat{\beta}_{LSE} \end{aligned}$$

Last but not least: the method of moments (MoM) estimator. Under MoM, we want to choose the β that satisfies $\text{cov}(\epsilon, x_j) = 0$.

$$\begin{aligned} 0 &= \sum_{i=1}^{n} (e_i - \bar{e})(x_{ij} - \bar{x}_j) \\ &= \sum_{i=1}^{n} \left(e_i x_{ij} - e_i \bar{x}_j - \bar{e} x_{ij} + \bar{e} \bar{x}_j \right) \\ &= \sum_{i=1}^{n} e_i x_{ij} - \bar{x}_j \sum_{i=1}^{n} e_i - \bar{e} \sum_{i=1}^{n} x_{ij} + \bar{e} \bar{x}_j \end{aligned}$$

Suppose we center our data such that $\bar{x}_j = 0$. We also know that $\bar{e} = 0$. There-

fore, the above simplifies to

$$\begin{aligned} 0 &= \sum_{i=1}^{n} e_i x_{ij} \\ &= e^T Xbe \\ &= (Y - X\beta)^T X \\ &= Y^T X - \beta^T X^T X \\ &= X^T Y - X^T X\beta \\ \hat{\beta}_{MoM} &= (X^T X)^{-1} X^T Y, \end{aligned}$$

which is identical to the LSE and MLE estimate. Now, we want to find the variance of these estimates:

$$\begin{aligned} \text{var}(\hat{\beta}) &= \text{var}\left[(X^T X)^{-1} X^T Y\right] \\ &= (X^T X)^{-1} X^T \text{var}(Y) \left((X^T X)^{-1} X^T\right)^T \\ &= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} \end{aligned}$$

As a further reference, one can read section 2.6 in *Plane Answers to Complex Questions* by Ronald Christensen.

B. As mentioned above, the estimator in the previous part corresponds to the assumption that $y \sim \mathbf{N}(X\beta, \sigma^2 I)$. What happens if we instead postulate that $y \sim \mathbf{N}(X\beta, \Sigma)$, where Σ is an arbitrary known covariance matrix, not necessarily proportional to the identity? What is the MLE for β now, and what is the variance of this estimator?

First, we obtain the form of the likelihood and the log-likelihood:

$$f(y|\beta, \sigma^2) = \prod_{i=1}^{n} |2\pi\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(y_i - x_i^T\beta)^T\Sigma^{-1}(y_i - x_i^T\beta)\right]$$

$$\log f(y|\beta, \sigma^2) = \sum_{i=1}^{n} \log\left(|2\pi\Sigma|^{-1/2}\right) - \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^T\Sigma^{-1}(y_i - x_i^T\beta)$$

With the log-likelihood in hand, we can easily obtain the MLE for β by minimizing the second term, since the first is constant with respect to β:

$$\begin{aligned}
\frac{\partial}{\partial\beta}\left[\sum_{i=1}^{n}(y_i - x_i^T\beta)^T\Sigma^{-1}(y_i - x_i^T\beta)\right] &= -2\sum_{i=1}^{n} y_i^T\Sigma^{-1}x_i + 2\sum_{i=1}^{n}\beta^T x_i^T\Sigma^{-1}x_i \overset{set}{=} 0 \\
Y^T\Sigma^{-1}X &= \beta^T X^T\Sigma^{-1}X \\
X^T\Sigma^{-1}Y &= X^T\Sigma^{-1}X\beta \\
\hat{\beta} &= (X^T\Sigma^{-1}X)^- X^T\Sigma^{-1}Y,
\end{aligned}$$

where we use a generalized inverse because a standard inverse is not guaranteed. Now, we find the variance of $\hat{\beta}$:

$$\begin{aligned}
\text{var}\left[(X^T\Sigma^{-1}X)^- X^T\Sigma^{-1}Y\right] &= (X^T\Sigma^{-1}X)^- X^T\Sigma^{-1}\text{var}(Y)\left((X^T\Sigma^{-1}X)^- X^T\Sigma^{-1}\right)^T \\
&= (X^T\Sigma^{-1}X)^- X^T\Sigma^{-1}\Sigma\left((X^T\Sigma^{-1}X)^- X^T\Sigma^{-1}\right)^T \\
&= (X^T\Sigma^{-1}X)^- X^T\Sigma^{-1}X(X^T\Sigma^{-1}X)^- \\
&= (X^T\Sigma^{-1}X)^-,
\end{aligned}$$

where we used a Penrose property to obtain the last equality.

C. Show that in the special case where Σ is a diagonal matrix, that the MLE is the familiar *weighted least squares* estimator. That is, show that $\hat{\beta}$ is the solution to the following linear system of P equations in P unknowns:

$$(X^T W X)\hat{\beta} = X^T W y,$$

where W is a diagonal matrix of weights that you should relate to the σ_i^2's.

First, we start with the given equation:

$$X^T W Y = X^T W X \hat{\beta}$$

Now, consider the form of $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_n^2)$. Recall that the inverse of a diagonal matrix is simply a diagonal matrix with its diagonal entries as reciprocals of the original matrix. That is,

$$\Sigma^{-1} = \text{diag}\left(\frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2}, ..., \frac{1}{\sigma_n^2}\right).$$

By setting $W = \Sigma^{-1}$, the weights in W are simply the reciprocal of the variances in Σ, and we obtain the following:

$$X^T \Sigma^{-1} Y = X^T \Sigma^{-1} X \hat{\beta} \tag{2}$$

Finally, we see that equation (2) resembles our work in the previous problem. That is, the weighted LSE is simply the MLE.

A. Let's continue with the weighted least-squares estimator you just characterized, i.e. the solution to the linear system

$$(X^T W X)\hat{\beta} = X^T W y.$$

One way to calculate $\hat{\beta}$ is to: (1) recognize that, trivially, the solution to the above linear system must satisfy $\hat{\beta} = (X^T W X)^{-1} X^T W y$; and (2) to calculate this directly, i.e. by inverting $X^T W X$. Let's call this the "inversion method" for calculating the WLS solution. Numerically speaking, is the inversion method the fastest and most stable way to actually solve the above linear system? Do some independent sleuthing on this question. Summarize what you find, and provide pseudo-code for at least one alternate method based on matrix factorizations – call it "your method" for short.

Is inversion quicker than factorization?

For this problem, we consider matrix factorization using LU decomposition, inspired by Gunderson's blog post. Solving a linear system directly for $\hat{\beta}$ using factorization is roughly $\frac{2}{3}n^3 + 2n^2$ flops, whereas inverting a matrix, then multiplying the matrix out will require at least $\frac{14}{3}n^3$ flops. Therefore, inversion is much slower than directly solving the system of linear equations; in fact, the inversion method could be up to 7 times slower than using factorization to solve for $\hat{\beta}$.

Is inversion more stable than factorization?

Again, we have a resounding "no!" It may be the case that $X^T W X$ is an ill-conditioned matrix, which would result in sharp decays in singular values that lead to many instabilities. With factorization, we can work with ill-conditioned matrices with more stability, as inverting an ill-conditioned matrix comes with some instabilities. Additionally, $X^T W X$ may be singular, in which case we cannot directly take the inverse of $X^T W X$. Rather, we would have to use a generalized inverse, which are not unique and could lead to different inference.

Matrix Factorization Algorithm

Algorithm 1 "My method"

1: **GOAL:** Solve for $\hat{\beta}$ in $X^T W X \hat{\beta} = X^T W y$
2: Factor $X^T W X$ as LU using LU decomposition, where L and U are lower- and upper-triangular matrices, respectively.
3: Solve for z in $Lz = X^T W y$ using forward substitution.
4: Solve for $\hat{\beta}$ in $U\hat{\beta} = z$ using backward substitution.

B. Code up functions that implement the inversion method and "your method." Simulate some silly data from the linear model for a range of values N and P. Benchmark the performance of the inversion solver and your solver across a range of scenarios.

Given various values for N and P, I initialized the following variables:

- $W = I_n$
- $X_{ii} \sim \mathrm{N}(0, 1)$
- $y \sim \mathrm{N}(0.3 * X[, 1] + 0.5 * X[, 2], 1)$

The mean execution times for both of the methods are reported in the table below under various values for N and P.

N	P	Inversion Time (ms)	Factorization Time (ms)
10	2	0.5518008	0.4652355
100	50	165.27649	11.32176
500	100	1974.4714	112.4707
800	200	31664.8722	800.1262

As can be seen in the above table, the factorization method is more computationally efficient under various scenarios. Therefore, we have experimental evidence that the LU decomposition method is quicker than the inversion method. My personal computer was not able to implement a combination of N and P greater that 800 and 200, respectively; my R session would kill itself. [1] However, even with these various scenarios, we can see that as the parameter space and dimension of data increases, the discrepancy between mean execution time of the two methods increases.

[1] Don't worry- it has since been resurrected.

A. Starting from the "standard" form of each PDF/PMF, show that the following distributions are in an exponential family, and find the corresponding $b, c, \theta,$ and $a(\phi)$.

(i) $Y \sim \mathbf{N}(\mu, \sigma^2)$ **for known** σ^2

Let's begin by writing the PDF for the normal distribution:

$$\begin{aligned} f(y|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\} \\ &= \exp\left\{-\frac{1}{2\sigma^2}(y^2 - 2y\mu + \mu^2)\right\} \cdot \exp\left\{\log\left((2\pi\sigma^2)^{-1/2}\right)\right\} \\ &= \exp\left\{\frac{y^2}{-2\sigma^2} + y\frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\} \\ &= \exp\left\{\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} + \left(-\frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right)\right\} \end{aligned}$$

By letting $\theta = \mu, a(\phi) = \sigma^2, b(\theta) = \frac{1}{2}\mu^2$, and $c(y|\phi) = -\frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$, we can write the normal distribution with fixed variance in the form of an exponential family:

$$f(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y|\phi)\right\}$$

(ii) $Y = Z/N$**, where** $Z \sim \mathbf{Binom}(N, P)$ **for known** N

We can easily obtain the PMF of Y through a transformation of random variables:

$$\begin{aligned} P\left(Y = \frac{Z}{N}\right) &= P\left(Z = NY\right) \cdot \left|\frac{1}{N}\right| \\ &= \binom{N}{NY} P^{NY}(1-P)^{N-NY} \cdot \frac{1}{N} \\ &= \exp\left\{\log\left[\binom{N}{NY} P^{NY}(1-P)^{N-NY} \cdot \frac{1}{N}\right]\right\} \\ &= \exp\left\{\log\left[\binom{N}{NY}\right] + NY\log(P) + N\log(1-P) - NY\log(1-P) - \log(N)\right\} \\ &= \exp\left\{Y\left[N\log\left(\frac{P}{1-P}\right)\right] - N\log\left(\frac{1}{1-P}\right) + \log\left[\binom{N}{NY}\right] - \log(N)\right\} \end{aligned}$$

Let $\theta = N\log\left(\frac{P}{1-P}\right), b(\theta) = N\log\left(\frac{1}{1-P}\right), a(\phi) = 1,$ and $c(y|\phi) = \log\left[\binom{N}{NY}\right] - \log(N)$ to get the form of an exponential family.

(iii) $Y \sim \textbf{Pois}(\lambda)$
You know the drill!

$$\begin{aligned} f(y|\lambda) &= \frac{e^{-\lambda}\lambda^{y}}{y!} \\ &= \exp\left\{\log\left[\frac{e^{-\lambda}\lambda^{y}}{y!}\right]\right\} \\ &= \exp\left\{-\lambda + y\log(\lambda) - \log(y!)\right\} \\ &= \exp\left\{y\log(\lambda) - \lambda + (-\log(y!))\right\} \end{aligned}$$

The above is in the desired exponential family form since we can let $\theta = \log(\lambda), a(\phi) = 1, b(\theta) = \lambda$, and $c(y|\phi) = -\log(y!)$. Therefore, we have shown that we can write the PDF of Y in the desired exponential family form.

B. We want to characterize the mean and variance of a distribution in the exponential family. To do this, we'll take an unfamiliar route, involving a preliminary lemma. Define the score $s(\theta)$ as the gradient of the log-likelihood with respect to θ:

$$s(\theta) = \frac{\partial}{\partial \theta} \log L(\theta).$$

While we think of the score as a function of θ, clearly the score also depends on the data. So a natural question is: what can we say about the distribution of the score over different random realizations of the data under the true data-generating process, i.e., at the true θ? It turns out we can say the following, sometimes referred to as the score equations:

$$E\left[s(\theta)\right] = 0$$
$$\mathcal{I}(\theta) \equiv \mathbf{var}(s(\theta)) = -E\left[H(\theta)\right],$$

where the mean and variance are taken under the true θ. Prove these score equations.

First, we prove $E\left[s(\theta)\right] = 0$. Note that while the score is a function of θ, it's also dependent on the data y. Therefore, we can take the expected value of the score over the sample space $\mathcal{Y}$. Let's write out the form of $E\left[s(\theta)\right]$:

$$\begin{aligned} E\left[s(\theta)\right] &= \int_{\mathcal{Y}} s(\theta) f(y|\theta) dy \\ &= \int_{\mathcal{Y}} \frac{\partial}{\partial \theta} \log L(\theta) \cdot f(y|\theta) dy \\ &= \int_{\mathcal{Y}} \frac{\frac{\partial}{\partial \theta} L(\theta)}{L(\theta)} \cdot f(y|\theta) dy \end{aligned}$$

Now, we use the statistical trick that we can rewrite the likelihood function as a PDF, since we integrate over $\mathcal{Y}$ with PDF $f(y|\theta)$:

$$\begin{aligned} E\left[s(\theta)\right] &= \int_{\mathcal{Y}} \frac{\frac{\partial}{\partial \theta} f(y|\theta)}{f(y|\theta)} \cdot f(y|\theta) dy \\ &= \int_{\mathcal{Y}} \frac{\partial}{\partial \theta} f(y|\theta) dy \\ &= \frac{\partial}{\partial \theta} \int_{\mathcal{Y}} f(y|\theta) dy \\ &= \frac{\partial}{\partial \theta}(1) = 0, \end{aligned}$$

where we assume that any necessary technical conditions are met to be able to switch the order of integration and differentiation.

Now, we prove that $\text{var}(s(\theta)) = -E\left[H(\theta)\right]$. Using the provided hint, suppose we differentiate the first equation with respect to θ^T:

$$\begin{aligned}
\frac{\partial}{\partial \theta^T} E(s(\theta)) &= \frac{\partial}{\partial \theta^T} \int_{\mathcal{Y}} \frac{\partial}{\partial \theta} \log L(\theta) f(y|\theta) dy \\
&= \int_{\mathcal{Y}} \frac{\partial}{\partial \theta^T} \left[\frac{\partial}{\partial \theta} \log L(\theta) f(y|\theta) \right] dy \\
&= \int_{\mathcal{Y}} \frac{\partial}{\partial \theta} \log L(\theta) \cdot \frac{\partial}{\partial \theta^T} f(y|\theta) + f(y|\theta) \frac{\partial^2}{\partial \theta^T \theta} \log L(\theta) dy \\
&= \int_{\mathcal{Y}} \frac{\partial}{\partial \theta} \log L(\theta) \cdot \frac{\partial}{\partial \theta^T} L(\theta) dy + \int_{\mathcal{Y}} f(y|\theta) \frac{\partial^2}{\partial \theta^T \theta} \log L(\theta) dy \\
&= \int_{\mathcal{Y}} \frac{\partial}{\partial \theta} \log L(\theta) \cdot \frac{\partial}{\partial \theta^T} \log L(\theta) \cdot f(y|\theta) dy + E\left[\frac{\partial^2}{\partial \theta^T \theta} \log L(\theta) \right] \\
&= E\left[\frac{\partial}{\partial \theta} \log L(\theta) \cdot \frac{\partial}{\partial \theta^T} \log L(\theta) \right] + E\left[H(\theta)\right] \\
&= E\left[s(\theta) s(\theta)^T \right] + E\left[H(\theta)\right] \\
&\overset{set}{=} \frac{\partial}{\partial \theta^T}(0) = 0
\end{aligned}$$

Before we get to the big reveal, let's acknowledge the nice property that we used to obtain the fifth equality:

$$\begin{aligned}
\frac{\partial}{\partial \theta} \log L(\theta) &= \frac{\frac{\partial}{\partial \theta} L(\theta)}{f(y|\theta)} \\
\implies \frac{\partial}{\partial \theta} L(\theta) &= \frac{\partial}{\partial \theta} \log L(\theta) \cdot f(y|\theta)
\end{aligned}$$

Now, we see from the above that

$$\begin{aligned}
\text{var}\left[s(\theta)\right] &= E[s(\theta) s(\theta)^T] - (E[s(\theta)])^2 \\
&= E[s(\theta) s(\theta)^T] \\
&= -E[H(\theta)]
\end{aligned}$$

C. Use the score equations you just proved to show that, if $Y \sim f(y|\theta, \phi)$ is an exponential family, then

$$E(Y) = b'(\theta)$$
$$\mathbf{var}(Y) = a(\phi)b''(\theta)$$

Thus, the variance of Y is a product of two terms: $b''(\theta)$ depends only on the canonical parameter θ, and hence on the mean, since we showed that $E(Y) = b'(\theta)$; $a(\phi)$ is independent of θ. Note that the most common form of a is $a(\phi) = \phi/w$ where ϕ is called a dispersion parameter and w is a known prior weight that can vary from one observation to another.

Recall that $E[s(\theta)] = E\left[\frac{\partial}{\partial\theta}\log L(\theta)\right] = 0$. Assume without loss of generality that there are n observations. For exponential families, we know that the log-likelihood is

$$\begin{aligned}
\log L(\theta) &= \log\left[\prod_{i=1}^{n}\exp\left\{\frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i|\phi)\right\}\right] \\
&= \sum_{i=1}^{n}\left[\frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i|\phi)\right] \\
&= \frac{\theta}{a(\phi)}\sum_{i=1}^{n} y_i - \frac{nb(\theta)}{a(\phi)} + \sum_{i=1}^{n} c(y_i|\phi)
\end{aligned}$$

By taking the expectation of the gradient of the log-likelihood with respect to θ, we obtain the following:

$$\begin{aligned}
E[s(\theta)] &= \int_{\mathcal{Y}} \frac{\partial}{\partial\theta}\left[\frac{\theta}{a(\phi)}\sum_{i=1}^{n} y_i - \frac{nb(\theta)}{a(\phi)} + \sum_{i=1}^{n} c(y_i|\phi)\right] f(y|\theta)dy \\
&= \int_{\mathcal{Y}} \left[\frac{1}{a(\phi)}\sum_{i=1}^{n} y_i - \frac{nb'(\theta)}{a(\phi)}\right] f(y|\theta)dy \\
&= E\left[\frac{\sum_{i=1}^{n} y_i}{a(\phi)} - \frac{nb'(\theta)}{a(\phi)}\right] \\
&= \frac{1}{a(\phi)}\sum_{i=1}^{n} E(Y) - \frac{nb'(\theta)}{a(\phi)} \\
&\overset{set}{=} 0
\end{aligned}$$

By manipulating the above equations, we get

$$E(Y) = b'(\theta)$$

Now, let's obtain the variance of Y:

$$\begin{aligned}
\text{var}(s(\theta)) &= \text{var}\left[\frac{1}{a(\phi)}\sum_{i=1}^{n} y_i - \frac{nb'(\theta)}{a(\phi)}\right] \\
&= \frac{1}{a(\phi)^2}\sum_{i=1}^{n}\text{var}(Y) \\
&\overset{set}{=} -E\left[H(\theta)\right]
\end{aligned}$$

Note that

$$\begin{aligned}
-E[H(\theta)] &= -E\left[\frac{\partial}{\partial\theta^T}\left(\frac{1}{a(\phi)}\sum_{i=1}^{n} y_i - \frac{nb'(\theta)}{a(\phi)}\right)\right] \\
&= -\int_{\mathcal{Y}}\frac{\partial}{\partial\theta^T}\left(\frac{1}{a(\phi)}\sum_{i=1}^{n} y_i - \frac{nb'(\theta)}{a(\phi)}\right) f(y|\theta)dy \\
&= \int_{\mathcal{Y}}\frac{nb''(\theta)}{a(\phi)} f(y|\theta)dy \\
&= E\left[\frac{nb''(\theta)}{a(\phi)}\right] \\
&= \frac{nb''(\theta)}{a(\phi)}
\end{aligned}$$

By combining the two above derivations, we see that

$$\begin{aligned}
\frac{1}{a(\phi)^2}\sum_{i=1}^{n}\text{var}(Y) &= \frac{nb''(\theta)}{a(\phi)} \\
\implies \text{var}(Y) &= a(\phi)b''(\theta)
\end{aligned}$$

D. To convince yourself that your result in (C) is correct, use these results to compute the mean and variance of the $\mathbf{N}(\mu, \sigma^2)$ distribution.

Recall from (a) that $\theta = \mu, a(\phi) = \sigma^2$, and $b(\theta) = \frac{1}{2}\mu^2$. While (a) *did* assume that σ^2 was known, we see that the result still holds! From (c), we found that

$$\begin{aligned} E(Y) &= b'(\theta) \\ &= \frac{\partial}{\partial \mu}\left(\frac{1}{2}\mu^2\right) \\ &= \mu, \\ \text{var}(Y) &= a(\phi)b''(\theta) \\ &= \sigma^2 \frac{\partial^2}{\partial \mu^2}\left(\frac{1}{2}\mu^2\right) \\ &= \sigma^2 \frac{\partial}{\partial \mu}(\mu) \\ &= \sigma^2, \end{aligned}$$

which are certainly the mean and variance of a $\text{N}(\mu, \sigma^2)$ distribution.

A. Deduce from your results above that, in a GLM,

$$\theta_i = (b')^{-1}\left(g^{-1}(x_i^T\beta)\right),$$
$$\mathbf{var}(Y_i) = \frac{\phi}{w_i}V(\mu_i)$$

for some function V that you should specify in terms of the building blocks of the exponential family model. V is often referred to as the variance function, since it explicitly relates the mean and the variance in a GLM.

Let's start with proving the first equation. Recall that $E(Y_i) = b'(\theta_i)$ and, by definition of GLM, $E(Y_i) = \mu_i$. Additionally, by definition, $g(\mu_i) = x_i^T\beta$. Therefore, we can simply equate these equations in the following way:

$$\begin{aligned} E(Y_i) = b'(\theta_i) &= \mu_i \\ \mu_i &= g^{-1}(x_i^T\beta), \\ \implies b'(\theta_i) &= g^{-1}(x_i^T\beta), \\ \implies \theta_i &= (b')^{-1}\left(g^{-1}(x_i^T\beta)\right), \end{aligned}$$

which is the first equation.

Now, we want to prove the second equation, containing the variance of Y_i. Recall that $\text{var}(Y_i) = a(\phi)b''(\theta)$. In the formulation of the GLM, we see that $a(\phi) = \frac{\phi}{w_i}$. Therefore,

$$\begin{aligned} \text{var}(Y_i) &= \frac{\phi}{w_i}b''(\theta_i) \\ &= \frac{\phi}{w_i}b''\left((b')^{-1}\left(g^{-1}(x_i^T\beta)\right)\right) \\ &= \frac{\phi}{w_i}b''\left((b')^{-1}(\mu_i)\right) \end{aligned}$$

By letting $V(\mu_i) = b''\left((b')^{-1}(\mu_i)\right)$, we get the second equation. Notice how we wrote $V(\mu_i)$ as a function of μ_i using functions of the exponential family (i.e., function $b(\cdot)$).

B. Take two special cases. (1) Suppose that Y is a Poisson GLM, i.e., that the stochastic component of the model is a Poisson distribution. Show that $V(\mu) = \mu$. (2) Suppose that $Y = Z/N$ is a Binomial GLM, i.e., that the stochastic component of the model is a Binomial distribution $Z \sim$ Binom(N, P) and that Y is the fraction of yes outcomes. Show that $V(\mu) = \mu(1-\mu)$.

First special case:

We can always write the stochastic component of the model in the following form:

$$f(y_i|\theta_i, \phi_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} + c(y_i|\phi/w_i)\right\},$$

which resembles an exponential family where $a(\phi) = \frac{\phi}{w_i}$. Note that we can write the Poisson PMF as

$$\begin{aligned} f(y_i|\lambda_i) &= \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \\ &= \exp\left\{\log\left[\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}\right]\right\} \\ &= \exp\left\{-\lambda_i + y_i\log(\lambda_i) - \log(y_i!)\right\}, \end{aligned}$$

which is an exponential family where $\theta_i = \log(\lambda_i)$, $b(\theta_i) = \lambda_i = e^{\theta_i}$, $a(\phi) = \frac{\phi}{w_i} = 1$, and $c(y_i|\phi/w_i) = -\log(y_i!)$. Recall that, for an exponential family,

$$\begin{aligned} E(Y_i) &= b'(\theta_i), \\ \text{var}(Y_i) &= a(\phi)b''(\theta_i), \end{aligned}$$

so we easily obtain

$$\begin{aligned} E(Y_i) &= b'(\theta_i) \\ &= e^{\theta_i} \\ &= \lambda_i, \\ \text{var}(Y_i) &= a(\phi)b''(\theta) \\ &= \frac{\partial}{\partial\theta_i}(e^{\theta_i}) \\ &= \lambda_i, \end{aligned}$$

By definition, $\mu_i = E(Y_i)$, so $\mu_i = \lambda_i$. Now, we set $V(\mu_i) = b''(\theta)$ to satisfy $\text{var}(Y_i) = \frac{\phi}{w_i}V(\mu_i)$. Therefore, $V(\mu_i) = \mu_i$.

Second special case:

If the stochastic component is $Z \sim \text{Binom}(N, P)$, then we can write

$$f(z_i|\theta_i, \phi_i) = \exp\left\{\frac{z_i\theta_i - b(\theta_i)}{\phi/w_i} + c(z_i|\phi/w_i)\right\}$$

Note that since Z follows a binomial distribution, we can write its PMF as

$$\begin{aligned} f(z_i|\theta_i, \phi_i) &= \binom{N}{z_i} P^{z_i}(1-P)^{N-z_i} \\ &= \exp\left\{\log\left[\binom{N}{z_i} P^{z_i}(1-P)^{N-z_i}\right]\right\} \\ &= \exp\left\{z_i \log(P) - z_i \log(1-P) + N\log(1-P) + \log\left[\binom{N}{z_i}\right]\right\} \end{aligned}$$

Note that we can only write this PMF in the form of an exponential family for the GLM when N is known. So, let's assume N is known to show the PMF of z_i can be written in the desired exponential family form:

$$\begin{aligned} f(z_i|\theta_i, \phi_i) &= \exp\left\{z_i \log(P) - z_i \log(1-P) + N\log(1-P) + \log\left[\binom{N}{z_i}\right]\right\} \\ &= \exp\left\{z_i \log\left(\frac{P}{1-P}\right) - N\log\left(\frac{1}{1-P}\right) + \log\left[\binom{N}{z_i}\right]\right\}, \end{aligned}$$

where $a(\phi) = \frac{\phi}{w_i} = 1$, $\theta_i = \log\left(\frac{P}{1-P}\right)$, $b(\theta_i) = N\log\left(\frac{1}{1-P}\right)$, and $c(z_i|\phi/w_i) = \log\left[\binom{N}{z_i}\right]$. Once again, we find the form of $V(\mu_i)$ by setting it equal to $b''(\theta_i)$. Note that

$$\begin{aligned} \theta_i &= \log\left(\frac{P}{1-P}\right) \\ \implies e^{\theta_i} &= \frac{P}{1-P} \\ \implies e^{\theta_i} &= P + Pe^{\theta_i} \\ \implies P(1+e^{\theta_i}) &= e^{\theta_i} \\ \implies P &= \frac{e^{\theta_i}}{1+e^{\theta_i}} \end{aligned}$$

Knowing P, we can obtain the proper form for $b(\theta_i)$:

$$b(\theta_i) = N\log\left(1+e^{\theta_i}\right)$$

Now, we can obtain $b''(\theta_i)$:

$$\begin{aligned} b'(\theta_i) &= \frac{N}{1+e^{\theta_i}} \cdot e^{\theta_i} \\ b''(\theta_i) &= \frac{N}{1+e^{\theta_i}} \cdot e^{\theta_i} - \left(e^{\theta_i}\right)^2 \frac{N}{(1+e^{\theta_i})^2} = \frac{Ne^{\theta_i}}{(1+e^{\theta_i})^2} \end{aligned}$$

We can find μ_i with

$$\begin{aligned} E(Z_i) &= b'(\theta_i) \\ &= \frac{N}{1+\frac{P}{1-P}} \cdot \frac{P}{1-P} \\ &= (1-P)N \cdot \frac{P}{1-P} \\ &= NP \\ &\overset{set}{=} \mu, \end{aligned}$$

which we set to μ by definition of the GLM. Now, we can find $b''(\theta_i)$ in terms of $\mu = NP$:

$$\begin{aligned} b''(\theta_i) &= NP - \left(\frac{P}{1-P}\right)^2 (1-P)^2 N \\ &= NP - NP^2 \\ &= NP(1-P) \end{aligned}$$

Now, notice that all of the above work was in terms of $z_i \sim \text{Binom}(N, P)$; however, this question asks us to show that $V(\mu) = b''(\theta_i) = \mu(1-\mu)$ for $Y = \frac{Z}{N}$. We can simply divide $b''(\theta_i)$ by N to obtain the corresponding $b''(\theta_i)$ for Y. Therefore, $V(\mu) = P(1-P) = \mu(1-\mu)$.

C. To specify a GLM, we must choose the link function $g(\mu_i)$. Recall that g links the predictors with the mean of the response: $g(\mu_i) = x_i^T\beta$. Since you've shown that

$$\theta_i = (b')^{-1}\left\{g^{-1}(x_i^T\beta)\right\},$$

a "simple" choice of link function is one where $g^{-1} = b'$. This is known as the canonical link, in which case the canonical parameter simplifies to $\theta_i = x_i^T\beta$. So under the canonical link $g(\mu) = (b')^{-1}(\mu)$, we have the model

$$f(y_i|\beta,\phi) = \exp\left\{\frac{y_i x_i^T\beta - b(x_i^T\beta)}{\phi/w_i} + c(y_i|\phi/w_i)\right\}.$$

Now return to the two special cases from the previous problem and find the canonical link $g(\mu)$.

First special case:

Recall that $b(\theta_i) = e^{\theta_i}$. We want to find $g(\mu)$ that satisfies $g(\mu) = (b')^{-1}(\mu)$. Note that $b'(\theta_i) = e^{\theta_i}$, so

$$\begin{aligned} & g(\mu) = (b')^{-1}(\mu) \\ \implies & b'(g(\mu)) = \mu \\ \implies & e^{g(\mu)} = \mu \\ \implies & g(\mu) = \log(\mu) \end{aligned}$$

Therefore, the canonical link is the log link, i.e., $g(\mu) = \log(\mu)$.

Second special case:

In this case, we found that

$$\begin{aligned} b(\theta_i) &= \log\left(1 + e^{\theta_i}\right), \\ b'(\theta_i) &= \frac{1}{1+e^{\theta_i}} \cdot e^{\theta_i}. \end{aligned}$$

We want to find $g(\mu)$ that satisfies $g(\mu) = (b')^{-1}(\mu)$:

$$\begin{aligned} & b'(g(\mu)) = \mu \\ \implies & \frac{e^{g(\mu)}}{1+e^{g(\mu)}} = \mu \\ \implies & e^{g(\mu)} = \mu + \mu e^{g(\mu)} \\ \implies & e^{g(\mu)} = \frac{\mu}{1-\mu} \end{aligned}$$

Therefore, our canonical link is $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$.

A. Using the chain rule

$$\frac{\partial}{\partial \beta} = \frac{\partial}{\partial \theta} \cdot \frac{\partial \theta}{\partial \mu} \cdot \frac{\partial \mu}{\partial \beta},$$

show that

$$s(\beta, \phi) \equiv \nabla_\beta \log L(\beta, \phi) = \sum_{i=1}^{n} \frac{w_i(Y_i - \mu_i)x_i}{\phi V(\mu_i) g'(\mu_i)},$$

where x_i is the vector of predictors for case i.

First, let's solve for $\frac{\partial \log(L)}{\partial \theta_i}$, $\frac{\partial \theta_i}{\partial \mu_i}$, and $\frac{\partial \mu_i}{\partial \beta}$:

$$\begin{aligned}
\frac{\partial \log(L)}{\partial \theta_i} &= \frac{\partial}{\partial \theta_i}\left[\frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} + c(y_i|\phi/w_i)\right] \\
&= \frac{y_i - b'(\theta_i)}{\phi/w_i}, \\
\frac{\partial \theta_i}{\partial \mu_i} &= \frac{\partial}{\partial \mu_i}\left[(b')^{-1}(\mu_i)\right] \\
&= \left((b')^{-1}\right)'(\mu_i) \\
&= \frac{1}{b''\left\{(b')^{-1}(\mu_i)\right\}}, \\
\frac{\partial \mu_i}{\partial \beta} &= \frac{\partial}{\partial \beta}\left[g^{-1}(x_i^T\beta)\right] \\
&= (g^{-1})'(x_i^T\beta)x_i^T \\
&= \frac{x_i^T}{g'\left\{g^{-1}(x_i^T\beta)\right\}},
\end{aligned}$$

where we use the fact that $(g^{-1})'(x) = \frac{1}{g'\{g^{-1}(x)\}}$. Now, we multiply these three terms together to obtain $s_i(\beta, \phi) = \frac{\partial l}{\partial \beta}$:

$$\begin{aligned}
s_i(\beta, \phi) &= \left(\frac{y_i - b'(\theta_i)}{\phi/w_i}\right) \cdot \frac{1}{b''\left\{(b')^{-1}(\mu_i)\right\}} \cdot \frac{x_i^T}{g'\left\{g^{-1}(x_i^T\beta)\right\}} \\
&= \left(\frac{y_i - \mu_i}{\phi/w_i}\right) \cdot \frac{1}{b''\{\theta_i\}} \cdot \frac{x_i^T}{g'\{\mu_i\}} \\
&= \frac{w_i(y_i - \mu_i)x_i}{\phi V(\mu_i) g'(\mu_i)}
\end{aligned}$$

Therefore, we have shown that

$$s(\beta, \phi) = \sum_{i=1}^{n} \frac{w_i(y_i - \mu_i)x_i}{\phi V(\mu_i) g'(\mu_i)}$$

B. Show that under the canonical link, $g'(\mu) = 1/V(\mu)$, the score function simplifies to

$$s(\beta, \phi) = \sum_{i=1}^{n} \frac{w_i(Y_i - \mu_i)x_i}{\phi}.$$

Recall from (A) of this section that we were able to write the score as

$$s(\beta, \phi) = \sum_{i=1}^{n} \frac{w_i(y_i - \mu_i)x_i}{\phi V(\mu_i)g'(\mu_i)}$$

By substituting $g'(\mu_i)$ with $\frac{1}{V(\mu_i)}$, we easily obtain the desired result:

$$\begin{aligned} s(\beta, \phi) &= \sum_{i=1}^{n} \frac{w_i(y_i - \mu_i)x_i}{\phi V(\mu_i)/V(\mu_i)} \\ &= \sum_{i=1}^{n} \frac{w_i(y_i - \mu_i)x_i}{\phi} \end{aligned}$$

I assume that this question was actually asking us to show $g'(\mu) = 1/V(\mu)$:

$$\begin{aligned} V(\mu_i) &= b''(\theta_i) \\ &= b''((b')^{-1}(\mu_i)) \\ &= (g^{-1})'(g(\mu_i)) \\ &= \frac{1}{g'(g^{-1}(g(\mu_i)))} \\ &= \frac{1}{g'(\mu_i)} \\ \implies g'(\mu) &= 1/V(\mu) \end{aligned}$$

C. Let's take the specific case of a GLM for a binomial outcome, where $Y_i \sim$ Binom(N_i, μ_i) for known sample size N_i, where $Y_i = Z_i/N_i$ is the observed success fraction, and where μ_i is related to the predictors $x_i \in \mathcal{R}^p$ via the canonical logistic link. This is called the logistic regression model. Write your own function that will fit a logistic regression model by gradient descent. Try to maintain some level of generality to your code, i.e., so that it could also work with different GLMs. Use the "wdbc.csv" dataset from the course website, and use the first 10 features for X.

In a gradient descent algorithm, we can iteratively learn about our parameter of interest (β) by moving along the gradient of the log-likelihood. Suppose we denote the step-size of our gradient descent algorithm as γ. Then, we can update β with the following:

$$\begin{aligned}\beta^{(m+1)} &= \beta^{(m)} + \gamma \nabla_\beta \log L(\beta) \\ &= \beta^{(m)} + \gamma s(\beta, \phi),\end{aligned}$$

where we learned how to efficiently calculate $s(\beta, \phi)$ in (A). In this problem, we use a familiar binomial GLM, where $\phi/w_i = 1$ and the link function is the canonical logistic link. Therefore, our score simplifies to

$$\begin{aligned}s(\beta, \phi) &= \sum_{i=1}^{n} (Y_i - \mu_i) x_i \\ &= \sum_{i=1}^{n} (Y_i - g^{-1}(x_i^T \beta)) x_i\end{aligned}$$

The problem statement specifies that we use the canonical logistic link function, i.e. $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$. Therefore, the inverse of our link function is

$$g^{-1}(x_i^T \beta) = \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}},$$

which gives us

$$s(\beta, \phi) = \sum_{i=1}^{n} \left[Y_i - \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} \right] x_i$$

Now, we can program a function in *R* that obtains the score.

To implement gradient descent, we also need a proper way to obtain γ. While we can specify some fixed value for γ, it may be best to iteratively obtain γ using a line search. Therefore, we can obtain the optimal γ by minimizing the

Algorithm 1 Gradient Descent Algorithm for Logistic Regression Model

1: Read in and scale data X, y
2: Initialize $\log L(\beta) = -100000$, $\beta = \text{rep}(0.1, \text{ncol}(X))$ and conv=FALSE
3: Set tolerance at $1e-5$
4: **while** !conv **do**
5: Set $g = 0$
6: **for** i in 1:nrow(X) **do**
7: Add $g = g + \left[Y_i - \frac{e^{x_i^T \beta^{(i-1)}}}{1+e^{x_i^T \beta^{(i-1)}}} \right] x_i$
8: **end for**
9: Obtain γ by minimizing the negative log-likelihood at $\beta^{(i-1)}$
10: Calculate $\beta^{(i)} = \beta^{(i-1)} + \gamma g$
11: Compute and store log-likelihood at $X\beta^{(i)}$
12: **if** $|\text{log-likelihood}^{(i)} - \text{log-likelihood}^{(i-1)}| <$ tolerance **then**
13: Set conv=TRUE
14: **end if**
15: **end while**

negative log-likelihood given the current value of β, which can be done with the *optim()* function. The steps for this algorithm are reported in Algorithm 1.

After implementing the above algorithm, we obtain a vector of log-likelihood values; Figure 1 is a trace plot of these log-likelihood values. The red line in the plot is the log-likelihood reported by the *glm()* function in *R*. Using the *glm()* function with a binomial family as a means of comparison, we see that the above gradient descent algorithm works very well.

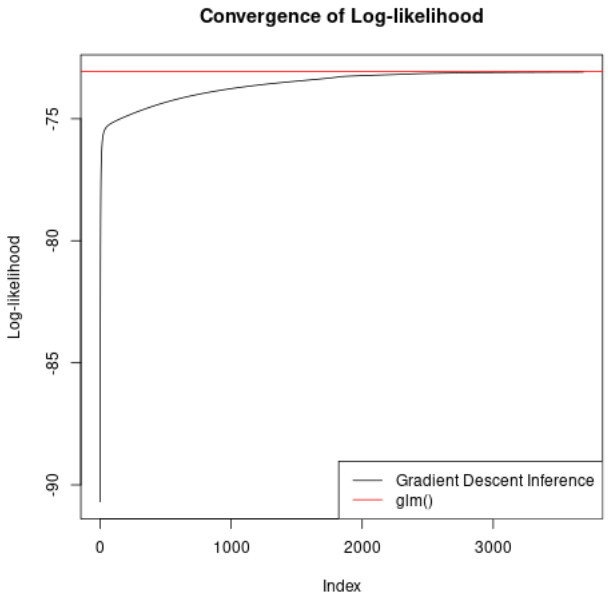

Figure 1: Log-likelihood values from Gradient Descent

D. Consider the Hessian matrix, i.e. the matrix of partial second derivatives of the log-likelihood

$$H(\beta, \phi) = \frac{\partial^2}{\partial \beta \partial \beta^T} \log L(\beta, \phi).$$

Give an expression for the Hessian matrix $H(\beta, \phi)$ of a GLM that is as simple as possible, ideally in matrix form. Assume the canonical link.

By (A) and (B),

$$\frac{\partial}{\partial \beta} \log L(\beta, \phi) = \sum_{i=1}^{n} \frac{w_i(Y_i - \mu_i)x_i}{\phi}$$

From here, we can take the partial derivative of the above with respect to β^T:

$$\begin{aligned}
\frac{\partial}{\partial \beta^T} \sum_{i=1}^{n} \frac{w_i(Y_i - \mu_i)x_i}{\phi} &= \frac{\partial}{\partial \beta^T} \sum_{i=1}^{n} \frac{w_i(Y_i - g^{-1}(x_i^T\beta))x_i}{\phi} \\
&= -\frac{1}{\phi} \sum_{i=1}^{n} w_i \frac{\partial}{\partial \beta^T} \left[g^{-1}(x_i^T\beta) \right] x_i \\
&= -\frac{1}{\phi} \sum_{i=1}^{n} w_i (g^{-1})'(x_i^T\beta) x_i^T x_i \\
&= -\frac{1}{\phi} \sum_{i=1}^{n} w_i b''(x_i^T\beta) x_i^T x_i
\end{aligned}$$

Suppose we want to write the above expression in matrix notation. First, we consider the following lemma.
Lemma: $A_{n\times p}B_{p\times k} = \sum_{i=1}^{n} \mathbf{a}_i \mathbf{b}_i^T$, where $\mathbf{a}_i$ and $\mathbf{b}_i$ denote the columns of A, B.

From the above lemma, we learned that we can write the product of matrices as the sum of column i in A times row i in B; this is known as the *outer product*. Now, we can apply this iteratively to obtain our answer. Note that x_i denotes the ith row of matrix X. Switching to our notation in the problem,

$$AB = \sum_{i=1}^{n} a_i^T b_i,$$

since a_i^T denotes the column of A and b_i denotes the row of B. Suppose we denote $W = \text{diag}\left(\frac{w_i}{\phi} b''(x_i^T\beta)\right)$. Consider X^TWX. Using the above lemma, we

obtain

$$\begin{aligned} X^T W X &= \left(\sum_{i=1}^{n} x_i \frac{w_i}{\phi} b''(x_i^T \beta) \right) X \\ &= \sum_{i=1}^{n} \left(x_i \frac{w_i}{\phi} b''(x_i^T \beta) \right)^T x_i \\ &= \sum_{i=1}^{n} \frac{w_i}{\phi} b''(x_i^T \beta) x_i^T x_i \end{aligned}$$

Therefore,

$$H(\beta, \phi) = -X^T W X$$

E. Now, consider a point $\beta_0 \in \mathbb{R}^p$, which serves as an intermediate guess for our vector of regression coefficients. Show that, for any GLM, the second-order Taylor approximation of $\mathcal{L}(\beta) = \log L(\beta, \phi)$ around the point β_0 can be expressed in the form

$$q(\beta; \beta_0) = -\frac{1}{2}(\tilde{y} - X\beta)^T W(\tilde{y} - X\beta) + c,$$

where $\tilde{y}$ is a vector of "working responses" and W is a diagonal matrix of "working weights". Give explicit expressions for the diagonal elements W_{ii} and for $\tilde{y}$, which will necessarily involve the point β_0 around which you're doing the expansion. Again, we're assuming the canonical link to make the algebra simpler.

Our second-order Taylor approximation is

$$\begin{aligned} q(\beta; \beta_0) &= \mathcal{L}(\beta_0) + \nabla_\beta \mathcal{L}(\beta)^T|_{\beta=\beta_0}(\beta - \beta_0) + \frac{1}{2}(\beta - \beta_0)^T H(\beta_0)(\beta - \beta_0) \\ &= \mathcal{L}(\beta_0) + \nabla_\beta \mathcal{L}(\beta)^T|_{\beta=\beta_0}(\beta - \beta_0) - \frac{1}{2}(\beta - \beta_0)^T X^T W X(\beta - \beta_0), \end{aligned}$$

where we use the Hessian from (D). Now, let's find $\nabla_\beta \mathcal{L}(\beta)^T$:

$$\begin{aligned} \nabla_\beta \mathcal{L}(\beta)^T &= \sum_{i=1}^n \frac{w_i(y_i - \mu_i)x_i^T}{\phi} \\ &= \sum_{i=1}^n \frac{w_i b''(x_i^T\beta)}{\phi} \cdot \frac{(y_i - \mu_i)x_i^T}{b''(x_i^T\beta)} \\ &= \tilde{z}^T W X, \end{aligned}$$

where $\tilde{z} = \left[... \frac{y_i - \mu_i}{b''(x_i^T\beta)} ... \right]$. Now, we revisit the second-order Taylor approximation:

$$\begin{aligned} q(\beta; \beta_0) &= \mathcal{L}(\beta_0) + \nabla_\beta \mathcal{L}(\beta)^T|_{\beta=\beta_0}(\beta - \beta_0) - \frac{1}{2}(\beta - \beta_0)^T X^T W X(\beta - \beta_0) \\ &= \mathcal{L}(\beta_0) + \tilde{z}_{\beta_0}^T W X(\beta - \beta_0) - \frac{1}{2}(\beta - \beta_0)^T X^T W X(\beta - \beta_0) \\ &= \tilde{z}_{\beta_0}^T W X\beta - \frac{1}{2}\beta^T X^T W X\beta + \beta_0^T X^T W X\beta + c^* \\ &= -\frac{1}{2}\beta^T X^T W X\beta + \tilde{y}^T W X\beta + c^* \\ &= -\frac{1}{2}(\tilde{y} - X\beta)^T W(\tilde{y} - X\beta) + c, \end{aligned}$$

where $\tilde{y}^T = \beta_0^T X^T + \tilde{z}_{\beta_0}^T = \beta_0^T X^T + \left[... \frac{y_i - \mu_i|_{\beta=\beta_0}}{b''(x_i^T\beta_0)} ... \right]$ and $W_{ii} = \frac{w_i}{\phi} b''(x_i^T\beta_0)$.

F. Read up on Newton's method for optimizing smooth functions (c.f. Nocedal and Wright, Chapter 2). Implement it for the logistic regression model and test it out on the same data set you just used to test out the gradient descent. Note: while you could do line search, there is a natural step size of 1 in Newton's method. Verify that your solution replicates the β estimates you get when using a program solver.

When using Newton's method, we update β with the following:

$$\beta^* = \beta + H^{-1}_{\mathcal{L}(\hat{\beta})} \nabla \mathcal{L}(\beta),$$

where H is the Hessian matrix we found in (D) (i.e. $H = -X^T W X$) and $\nabla \mathcal{L}(\beta) = s(\beta, \phi)$ is the score function. Our template for Newton's method can be found in Algorithm 2. Figure 2 displays the relatively quick convergence of the log-likelihood to that found using *glm()* in *R*.

Algorithm 2 Newton's Method for Logistic Regression Model

Read in and scale data X, y
Initialize $\log L(\beta) = -100000$, $\beta = \text{rep}(0.1, \text{ncol}(X))$ and conv=FALSE
Set tolerance at $1e-5$
while !conv **do**
  Set $g = 0$
  for i in 1:nrow(X) **do**
    Add $g = g + \left[Y_i - \frac{e^{x_i^T \beta^{(i-1)}}}{1+e^{x_i^T \beta^{(i-1)}}} \right] x_i$
  end for
  Calculate $H = -X^T W X$
  Calculate $\beta^{(i)} = \beta^{(i-1)} + H^{-1} g$
  Compute and store log-likelihood $\mathcal{L}(\beta^{(i)})$
  if $|\mathcal{L}(\beta^{(i)}) - \mathcal{L}(\beta^{(i-1)})| / |\mathcal{L}(\beta^{(i-1)})| <$ tol **then**
    Set conv=TRUE
  end if
end while

Below is a table comparing the results from my implementation of Newton's method and the results from *glm()*. With these minor discrepancies in so few iterations, it is natural to ask, "Why would I use gradient descent over Newton's method?" Let's take a look at the differences between these two methods:

1. Gradient descent is parametric according to the learning rate γ. If we hope to learn about some learning rate, then clearly gradient descent is the right selection. (While a parametric version of Newton's method exists, it is only applicable when we operate with a polynomial function with multiple roots.)

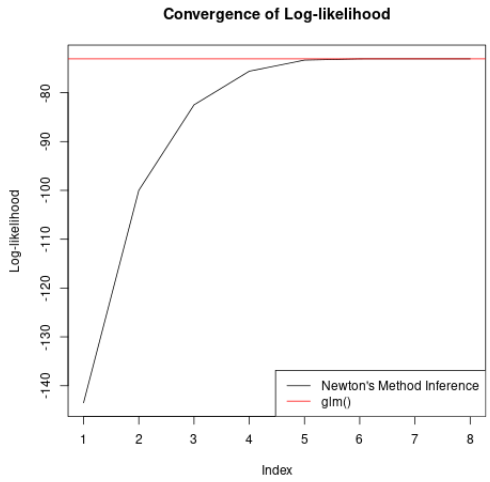

Figure 2: Convergence of Newton's Method

2. Newton's method requires second derivatives, whereas gradient descent can be applied using only the first derivative. Thus, Newton's method has stronger constraints in terms of the differentiability of the function.

3. If gradient descent reaches a stationary point, it continues to run; however, parameters won't update. Conversely, if Newton's method hits a stationary point, then the algorithm will terminate due to division by zero.

Newton's beta	**glm()'s beta**
0.48701671	0.48701675
-7.22184989	-7.22185053
1.65475612	1.65475615
-1.73763049	-1.73763027
14.00484503	14.00484560
1.07495327	1.07495329
-0.07723455	-0.07723455
0.67512312	0.67512313
2.59287422	2.59287426
0.44625630	0.44625631
-0.48248419	-0.48248420

G. Standard asymptotic theory implies that the maximum likelihood estimator is consistent and asymptotically normal around the true value β_0:

$$\hat{\beta} \sim N(\beta_0, \mathcal{I}(\beta_0, \phi)^{-1})),$$

where $\mathcal{I}(\beta_0, \phi)$ is called the *Fisher information matrix* and is defined as the variance of the score equations:

$$\mathcal{I}(\beta_0, \phi) \equiv \mathbf{var}(s(\beta_0, \phi)) = -E[H(\beta_0, \phi)]].$$

The fact that Fisher information is the negative of the expected Hessian motivates the following idea: use the inverse of the negative Hessian matrix at the MLE to approximate the inverse Fisher information, i.e. the covariance matrix of the estimator. Happily, you get this Hessian matrix for free when fitting by Newton's method.
For your logistic regression on the WDBC data fit via Newton's method, compute the square root of each diagonal element of the inverse Hessian matrix, evaluated at the MLE. Compare these to the standard errors you get when using a package solver.

The following table contains the standard errors arising from (1) the logistic regression model fit via Newton's method and (2) fit via *glm()* in *R*. The difference in standard errors is minimal.

Newton's standard errors	glm()'s standard errors
0.5642741	0.5643200
13.0939632	13.0949439
0.2775461	0.2775752
12.2742249	12.2749905
5.8903905	5.8909033
0.4493906	0.4494181
1.0742854	1.0743433
0.6473007	0.6473276
1.1069398	1.1070102
0.2914167	0.2914299
0.6040244	0.6040610

A. By construction, we know that the marginal prior distribution $p(\theta)$ is a gamma mixture of normals. Show that this takes the form of a centered, scaled t distribution:

$$p(\theta) \propto \left(1 + \frac{1}{\nu} \cdot \frac{(x-m)^2}{s^2}\right)^{-\frac{\nu+1}{2}}$$

with center m, scale s, and degrees of freedom ν, where you fill in the blank for m, s^2, and ν in terms of the four parameters of the normal-gamma family. Note: you did a problem just like this on a previous exercise! This shouldn't be a lengthy re-derivation.

We can calculate $p(\theta)$ as the integral of the joint prior $p(\theta, \omega)$ with respect to ω:

$$\begin{aligned}
p(\theta) &= \int_\Omega p(\theta, \omega) d\omega \\
&\propto \underbrace{\int_\Omega w^{\frac{d+1}{2}-1} \exp\left\{-\omega\left(\frac{\kappa(\theta-\mu)^2}{2} + \frac{\eta}{2}\right)\right\} d\omega}_{\text{kernel of Gamma}\left(\frac{d+1}{2}, \frac{\kappa(\theta-\mu)^2}{2} + \frac{\eta}{2}\right)} \\
&\propto \left(\frac{\left[\frac{\kappa(\theta-\mu)^2}{2} + \frac{\eta}{2}\right]^{\frac{d+1}{2}}}{\Gamma(\frac{d+1}{2})}\right)^{-1} \\
&\propto \left[\frac{\kappa(\theta-\mu)^2}{2} + \frac{\eta}{2}\right]^{-\frac{d+1}{2}} \\
&\propto \left[\left(\frac{\kappa(\theta-\mu)^2}{2} + \frac{\eta}{2}\right)\frac{2}{\eta}\right]^{-\frac{d+1}{2}} \frac{\eta}{2}^{-\frac{d+1}{2}} \\
&\propto \left[\frac{1}{d} \cdot \frac{(\theta-\mu)^2}{\eta/(\kappa d)} + 1\right]^{-\frac{d+1}{2}},
\end{aligned}$$

which takes the form of a centered, scaled t distribution with parameters

$$\begin{aligned}
m &= \mu \\
s &= \sqrt{\frac{\eta}{\kappa d}} \\
\nu &= d
\end{aligned}$$

(c.f. "Bayesian Inference in Simple Conjugate Families" (F))

B. Assume the normal sampling model in Equation (1) and the normal-gamma prior in Equation (2). Calculate the joint posterior density $p(\theta, \omega \mid \mathbf{y})$, up to constant factors not depending on ω or θ. Show that this is also a normal/gamma prior in the same form as above:

$$p(\theta, \omega \mid \mathbf{y}) \propto \omega^{(d^\star+1)/2-1} \exp\left\{-\omega \cdot \frac{\kappa^\star(\theta-\mu^\star)^2}{2}\right\} \cdot \exp\left\{-\omega \cdot \frac{\eta^\star}{2}\right\} \quad (1)$$

From this form of the posterior, you should able to read off the new updated parameters, by pattern-matching against the functional form in Equation (2):

- $\mu \longrightarrow \mu^\star = ?$
- $\kappa \longrightarrow \kappa^\star = ?$
- $d \longrightarrow d^\star = ?$
- $\eta \longrightarrow \eta^\star = ?$

We can find the joint posterior by taking the product of the likelihood and the joint prior:

$$\begin{aligned}
p(\theta, \omega|\mathbf{y}) &\propto p(\mathbf{y}|\theta, \omega)p(\theta, \omega) \\
&\propto \prod_{i=1}^{n} \sqrt{\omega} \exp\left\{-\frac{\omega}{2}(y_i-\theta)^2\right\} \omega^{\frac{d+1}{2}-1} \exp\left\{-\omega\left(\frac{\kappa(\theta-\mu)^2}{2}+\frac{\eta}{2}\right)\right\} \\
&\propto \omega^{\frac{d+n+1}{2}-1} \exp\left\{-\frac{\omega}{2}\left[\sum_{i=1}^{n}\left(y_i^2-2y_i\theta+\theta^2\right)+\kappa(\theta^2-2\theta\mu+\mu^2)+\eta\right]\right\} \\
&\propto \omega^{\frac{d+n+1}{2}-1} \exp\left\{-\frac{\omega}{2}\left[\sum_{i=1}^{n} y_i^2-2n\bar{y}\theta+n\theta^2+\kappa\theta^2-2\kappa\theta\mu+\kappa\mu^2+\eta\right]\right\} \\
&\propto \omega^{\frac{d+n+1}{2}-1} \exp\left\{-\frac{\omega}{2}\left[\sum_{i=1}^{n} y_i^2+(\kappa+n)\theta^2-2(n\bar{y}+\kappa\mu)\theta+\kappa\mu^2+\eta\right]\right\} \\
&\propto \omega^{\frac{d+n+1}{2}-1} \exp\left\{-\frac{\omega}{2}\left[(\kappa+n)\theta^2-2(n\bar{y}+\kappa\mu)\theta+\left(\kappa\mu^2+\eta+\sum_{i=1}^{n} y_i^2\right)\right]\right\} \\
&\propto \omega^{\frac{d+n+1}{2}-1} \exp\left\{-\omega\left[\frac{(\kappa+n)\left(\theta-\frac{n\bar{y}+\kappa\mu}{\kappa+n}\right)^2}{2}\right]\right\} \exp\left\{-\omega\left(\frac{\kappa\mu^2+\eta+\sum_{i=1}^{n} y_i^2-\frac{(n\bar{y}+\kappa\mu)^2}{\kappa+n}}{2}\right)\right\} \\
&\equiv \omega^{\frac{d^*+1}{2}-1} \exp\left\{-\omega\frac{\kappa^*(\theta-\mu^*)^2}{2}\right\} \exp\left\{-\omega\frac{\eta^*}{2}\right\},
\end{aligned}$$

where

$$
\begin{aligned}
d^* &= d + n \\
\kappa^* &= \kappa + n \\
\mu^* &= \frac{n\bar{y} + \kappa\mu}{\kappa + n} \\
\eta^* &= \eta + \kappa\mu^2 + \sum_{i=1}^{n} y_i^2 - \frac{(n\bar{y} + \kappa\mu)^2}{\kappa + n} \\
&= \eta + S_y + \frac{n\kappa(\bar{y} - \mu)^2}{n + \kappa}
\end{aligned}
$$

C. From the joint posterior you just derived, what is the conditional posterior distribution $p(\theta \mid \mathbf{y}, \omega)$? Note: this should require no calculation—you should just be able to read it off directly from the joint distribution, since you took care to set up things so that the joint posterior was in the same form as Equation (2).

By pattern matching, we see that

$$\begin{aligned} p(\theta|\mathbf{y},\omega) &\propto \exp\left\{-\frac{\omega\kappa^*}{2}(\theta-\mu^*)^2\right\} \\ &\equiv \text{Normal}(\mu^*, (\omega\kappa^*)^{-1}), \end{aligned}$$

where $\mu^* = \frac{n\bar{y}+\kappa\mu}{\kappa+n}$ and $\kappa^* = \kappa + n$.

D. From the joint posterior you calculated in (A), what is the marginal posterior distribution $p(\omega \mid \mathbf{y})$? Unlike the previous question, this one doesn't come 100% for free—you have to integrate over θ. But it shouldn't be too hard, since you can ignore constants not depending on ω in calculating this integral.

Let's integrate $p(\omega, \theta|\mathbf{y})$ with respect to θ to get our marginal posterior for ω:

$$\begin{aligned} p(\omega|\mathbf{y}) &= \int p(\omega,\theta|\mathbf{y})d\theta \\ &\propto w^{\frac{d^*+1}{2}-1}\exp\left\{-\omega\frac{\eta^*}{2}\right\}\underbrace{\int \exp\left\{-\frac{\omega\kappa^*}{2}(\theta-\mu^*)^2\right\}d\theta}_{\text{kernel of Normal}(\mu^*,(\omega\kappa^*)^{-1})} \\ &\propto \omega^{\frac{d^*+1}{2}-1}\exp\left\{-\omega\frac{\eta^*}{2}\right\}(\sqrt{\omega\kappa^*})^{-1} \\ &\propto \omega^{\frac{d^*}{2}-1}\exp\left\{-\omega\frac{\eta^*}{2}\right\}, \end{aligned}$$

so $p(\omega|\mathbf{y}) \equiv \text{Gamma}\left(\frac{d^*}{2}, \frac{\eta^*}{2}\right)$.

E. Show that the marginal posterior $p(\theta \mid \mathbf{y})$ takes the form of a centered, scaled t distribution and express the parameters of this t distribution in terms of the four parameters of the normal-gamma posterior for (θ, ω). Note: since you've set up the normal-gamma family in this careful conjugate form, this should require no extra work. It's just part (A), except for the posterior rather than the prior.

We can simply pattern-match from (A):

$$p(\theta|\mathbf{y}) \propto \left(1 + \frac{1}{\nu} \cdot \frac{(x-m)^2}{s^2}\right)^{\frac{\nu+1}{2}},$$

where $\nu = d^*, m = \mu^*$, and $s^2 = \frac{\eta^*}{d^*\kappa^*}$.

F. True or false: in the limit as the prior parameters κ, d, and η approach zero, the priors $p(\theta)$ and $p(\omega)$ are valid probability distributions. (Remember that a valid probability distribution must integrate to 1 (or something finite, so that it can normalized to integrate to 1) over its domain.)

FALSE; both priors will be undefined if their hyperparameters are 0.

G. True or false: in the limit as the prior parameters κ, d, and η approach zero, the posteriors $p(\theta \mid \mathbf{y})$ and $p(\omega \mid \mathbf{y})$ are valid probability distributions.

TRUE; if $\kappa, d, \eta \to 0$, then $d^*, \kappa^* \to n, \mu^* \to \bar{y}$, and $\eta^* \to S_y$. Therefore, the parameters in both marginal posteriors (θ and ω) are defined and the probability distributions are valid.

H. Your result in (E) implies that a Bayesian credible interval for θ takes the form

$$\theta \in m \pm t^\star \cdot s\,,$$

where m and s are the posterior center and scale parameters from (E), and $t^\star$ is the appropriate critical value of the t distribution for your coverage level and degrees of freedom (e.g. it would be 1.96 for a 95% interval under the normal distribution).
True or false: In the limit as the prior parameters κ, d, and η approach zero, the Bayesian credible interval for θ becomes identical to the classical (frequentist) confidence interval for θ at the same confidence level.

TRUE; as stated in (G), $d^*, \kappa^* \to n$, $\mu^* \to \bar{y}$, and $\eta^* \to S_y$. So, we have

$$\begin{aligned} s &= \sqrt{\frac{\eta^*}{d^*\kappa^*}} \\ &\to \sqrt{\frac{S_y}{n^2}} \\ &= \frac{\sqrt{S_y}}{n}, \\ m &= \mu^* \\ &\to \bar{y} \end{aligned}$$

Therefore,

$$m \pm t^\star \cdot s \to \bar{y} \pm t^* \frac{\sqrt{S_y}}{n},$$

which is identical to the frequentist confidence intervals for θ.

A. Derive the conditional posterior $p(\beta|\mathbf{y},\omega)$.

First, let's obtain the joint posterior distribution:

$$\begin{aligned}
p(\beta,\omega|\mathbf{y}) &\propto p(\mathbf{y}|\beta,\omega)\cdot p(\beta|\omega)\cdot p(\omega) \\
&\propto |(\omega\Lambda)^{-1}|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{y}-X\beta)^T\omega\Lambda(\mathbf{y}-X\beta)\right\} \\
&\quad\cdot|(\omega K)^{-1}|^{-1/2}\exp\left\{-\frac{1}{2}(\beta-m)^T\omega K(\beta-m)\right\} \\
&\quad\cdot\omega^{d/2-1}\exp\left\{-\frac{\omega\eta}{2}\right\} \\
&\propto \omega^{(n+p+d)/2-1} \\
&\quad\cdot\exp\left\{-\frac{1}{2}(\mathbf{y}-X\beta)^T\omega\Lambda(\mathbf{y}-X\beta)\right\} \\
&\quad\cdot\exp\left\{-\frac{1}{2}(\beta-m)^T\omega K(\beta-m)\right\}\exp\left\{-\frac{\omega\eta}{2}\right\}
\end{aligned}$$

From here, we can obtain the conditional posterior for β by only keeping terms containing β:

$$\begin{aligned}
p(\beta|\omega,\mathbf{y}) &\propto \exp\left\{-\frac{1}{2}\left[-2\mathbf{y}^T\omega\Lambda X\beta+\beta^TX^T\omega\Lambda X\beta-2m^T\omega K\beta+\beta^T\omega K\beta\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[-2\underbrace{(\mathbf{y}^T\omega\Lambda X+m^T\omega K)}_{\mathbf{b}^T}\beta+\beta^T\underbrace{(X^T\omega\Lambda X+\omega K)}_{A}\beta\right]\right\}
\end{aligned}$$

Therefore,

$$p(\beta|\omega,\mathbf{y}) = \text{Normal}\left((X^T\Lambda X+K)^{-1}(X^T\Lambda\mathbf{y}+K^Tm),(X^T\omega\Lambda X+\omega K)^{-1}\right)$$

B. Derive the marginal posterior $p(\omega|\mathbf{y})$.

To obtain the marginal posterior for ω, we integrate the integrate the joint posterior with respect to β:

$$\begin{aligned}
p(\omega|\mathbf{y}) &\propto \int_\beta p(\omega,\beta|\mathbf{y})d\beta \\
&\propto \omega^{(n+p+d)/2-1}\exp\left\{-\frac{\omega}{2}\left[\mathbf{y}^T\Lambda\mathbf{y}+m^TKm+\eta\right]\right\} \\
&\quad\cdot\int_\beta \underbrace{\exp\left\{-\frac{1}{2}\left[-2(\mathbf{y}^T\omega\Lambda X+m^T\omega K)\beta+\beta^T(X^T\omega\Lambda X+\omega K)\beta\right]\right\}}_{\text{kernel of Normal}(A^{-1}b,A^{-1})} d\beta \\
&\propto \omega^{(n+p+d)/2-1}\exp\left\{-\frac{\omega}{2}\left[\mathbf{y}^T\Lambda\mathbf{y}+m^TKm+\eta\right]\right\} \\
&\quad\cdot|\omega^{-1}(X^T\Lambda X+K)^{-1}|^{1/2}\cdot \underbrace{\exp\left\{\frac{1}{2}\left[b^TAb\right]\right\}}_{\text{term that doesn't contain }\beta\text{ in the above integrand}} \\
&\propto \omega^{(n+d)/2-1}\exp\left\{-\frac{\omega}{2}\left[\mathbf{y}^T\Lambda\mathbf{y}+m^TKm+\eta\right]\right\} \\
&\quad\cdot\exp\left\{\frac{\omega}{2}\left[(\mathbf{y}^T\Lambda X+m^TK)^T(X^T\Lambda X+K)^{-1}(\mathbf{y}^T\Lambda X+m^TK)\right]\right\}
\end{aligned}$$

Therefore, the marginal posterior for ω is

$$\begin{aligned}
p(\omega|\mathbf{y}) &= \text{Gamma}\left(\frac{n+d}{2},\frac{\eta^*}{2}\right), \\
\eta^* &= \eta+\mathbf{y}^T\Lambda\mathbf{y}+m^TKm-(\mathbf{y}^T\Lambda X+m^TK)^T(X^T\Lambda X+K)^{-1}(\mathbf{y}^T\Lambda X+m^TK)
\end{aligned}$$

C. Derive the marginal posterior $p(\beta|\mathbf{y})$.

We can derive the marginal posterior for β using Bayes' rule:

$$\begin{aligned}
p(\beta|\mathbf{y}) &\propto \int_\Omega p(\omega, \beta|\mathbf{y}) d\omega \\
&\propto \int_\Omega p(\beta|\omega, \mathbf{y}) p(\omega|\mathbf{y}) d\omega \\
&\propto \int_\Omega \omega^{(n+p+d)/2-1} \\
&\quad \cdot \exp\left\{-\frac{1}{2}(\mathbf{y} - X\beta)^T \omega \Lambda (\mathbf{y} - X\beta)\right\} \\
&\quad \cdot \exp\left\{-\frac{1}{2}(\beta - m)^T \omega K (\beta - m)\right\} \exp\left\{-\frac{\omega\eta}{2}\right\} d\omega,
\end{aligned}$$

which resembles the kernel of a Gamma distribution. Thus, we see that

$$\begin{aligned}
p(\beta|\mathbf{y}) &\propto \left(\frac{(\beta^*)^{a^*}}{\Gamma(a^*)}\right)^{-1} \\
&\propto \frac{\Gamma\left(\frac{n+p+d}{2}\right)}{\left(\frac{1}{2}\left[(\mathbf{y} - X\beta)^T \Lambda (\mathbf{y} - X\beta) + (\beta - m)^T K (\beta - m) + \eta\right]\right)^{\frac{n+p+d}{2}}} \\
&\propto \left(\frac{1}{2}\left[(\mathbf{y} - X\beta)^T \Lambda (\mathbf{y} - X\beta) + (\beta - m)^T K (\beta - m) + \eta\right]\right)^{-\frac{n+p+d}{2}} \\
&\propto \left(\frac{1}{2}\left[-2(\mathbf{y}^T \Lambda X + m^T K)\beta + \beta^T (X^T \Lambda X + K)\beta\right]\right)^{-\frac{\nu^*+p}{2}} \\
&\propto \left(\frac{1}{2}\left[-2(\mathbf{y}^T \Lambda X + m^T K)\beta + \beta^T \Lambda^* \beta\right]\right)^{-\frac{\nu^*+p}{2}} \\
&\propto \left((\beta - \mu^*)^T \Lambda^* (\beta - \mu^*) + \eta^*\right)^{-\frac{\nu^*+p}{2}} \\
&\propto \left(\frac{1}{n+d}(\beta - \mu^*)^T \frac{n+d}{\eta^*} \Lambda^* (\beta - \mu^*) + 1\right)^{-\frac{\nu^*+p}{2}} \\
&\propto \left(\frac{1}{\nu^*}(\beta - \mu^*)^T \Sigma^* (\beta - \mu^*) + 1\right)^{-\frac{\nu^*+p}{2}}
\end{aligned}$$

where $\nu^* = n + d$, $\Lambda^* = X^T \Lambda X + K$, $\mu^* = (\Lambda^*)^{-1}(X^T \Lambda \mathbf{y} + K^T m)$ and $\Sigma^* = \frac{\nu^*}{\eta^*}\Lambda^*$. From the above, we see that the marginal posterior for β is a t-distribution with ν^* degrees of freedom, location μ^*, and covariance Σ^*.

D. Using the "greenbuildings.csv" data set, what is the 95% Bayesian credible interval for the coefficient on the green rating variable? How does your result compare to the classical 95% confidence interval? What does a histogram of the model residuals reveal? Are you happy with your model?

Using the derivation for the marginal posterior $p(\beta|\mathbf{y})$, we are able to sample β without using a Gibbs sampling scheme. From our derivation in (c), we get the following 95% Bayesian credible intervals for the coefficient on the green rating variable:

$$[-0.03, 1.72] \quad \text{(using HDI estimation)}$$
$$[-0.04, 1.70] \quad \text{(using ETI estimation)}$$

In comparison, our 95% confidence interval using *lm()* in *R* for the same covariate is

$$[-0.087, 1.69]$$

It appears that the 95% confidence interval is wider than our 95% credible intervals with a notably lower lower-bound. In Figure 1, we report a histogram of the model residuals. It's important to note that there is a long right tail, implying that our model is not ideal. This is most likely the case because of two reasons: (1) our model should not assume normality and (2) our assumptions on the covariance appear to be flawed. I'm not happy with this model because the histogram of these residuals should not display any skewness.

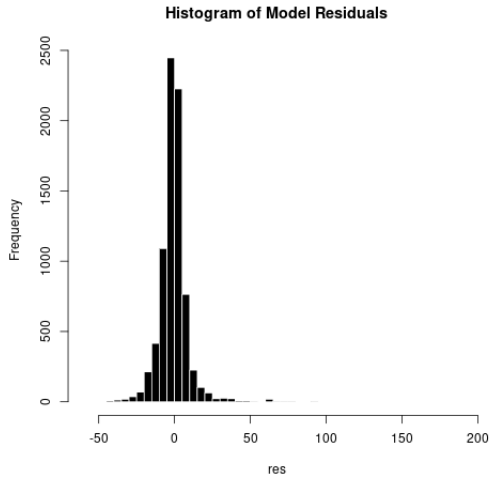

Figure 1: Histogram of Model Residuals

A. Under this model, what is the implied conditional distribution $p(y_i|\beta,\omega)$?

We can obtain the desired conditional distribution by marginalizing out λ_i:

$$\begin{aligned}
p(y_i|\beta,\omega) &= \int_\Lambda p(y_i|\lambda_i,\beta,\omega)\cdot p(\lambda_i)d\lambda_i \\
&\propto \int_\Lambda \sqrt{\lambda_i}\exp\left\{-\frac{\omega\lambda_i}{2}\left(y_i - X_i^T\beta\right)^2\right\}\cdot \lambda_i^{\frac{h}{2}-1}e^{-\frac{\lambda_i h}{2}}d\lambda_i \\
&\propto \int_\Lambda \underbrace{\lambda_i^{\frac{h+1}{2}-1}\exp\left\{-\lambda_i\left[\frac{\omega}{2}\left(y_i - X_i^T\beta\right)^2 + \frac{h}{2}\right]\right\}}_{\text{kernel of Gamma}\left(\frac{h+1}{2},\frac{\omega}{2}\left(y_i - X_i^T\beta\right)^2+\frac{h}{2}\right)} d\lambda_i \\
&\propto \left[\frac{\omega}{2}(y_i - X_i^T\beta)^2 + \frac{h}{2}\right]^{-\frac{h+1}{2}} \\
&\propto \left[\left(\frac{\omega}{2}(y_i - X_i^T\beta)^2 + \frac{h}{2}\right)\frac{2}{h}\right]^{-\frac{h+1}{2}}\left(\frac{2}{h}\right)^{\frac{h+1}{2}} \\
&\propto \left[\frac{\omega}{h}(y_i - X_i^T\beta)^2 + 1\right]^{-\frac{h+1}{2}},
\end{aligned}$$

which resembles a t-distribution. In fact,

$$p(y_i|\beta,\omega) = t\left(\nu = h, m = X_i^T\beta, s^2 = \frac{1}{\omega}\right)$$

B. What is the conditional posterior distribution $p(\lambda_i|\beta,\omega)$?

This is the integrand of the above problem, so we can easily see that

$$p(\lambda_i|\beta,\omega) = \text{Gamma}\left(\frac{h+1}{2}, \frac{\omega}{2}\left(y_i - X_i^T\beta\right)^2 + \frac{h}{2}\right)$$

C. Code a Gibbs sampler using the full-conditional for β (in the previous section), the marginal posterior for ω (in the previous section), and the conditional posterior distribution for λ_i. Apply it to the green buildings data set. Are you happier with the fit? How do the 95% credible intervals on each model term compare to our previous model? Are there certain regions of predictor space that seem to be associated with higher variance residuals?

In Figure 2, we see a side-by-side comparison of the histogram of residuals between our heteroskedastic (blue) and homoskedastic (red) models. It appears that with my choice of $h = 2$, both models demonstrate quite heavy tails; this means that my choice in $h = 2$ does not greatly improve our fit to the data. It seems that the heteroskedastic model has residuals that are more concentrated at 0, so that's at least a bit of improvement!

In Figure 3, we see the plot that Dr. Scott hinted at. Here, we see that for data points with smaller values, the relative variance could be significantly greater than the variance for data points of higher values. We see a somewhat-funnel shape to this plot, implying that greater values in data points have relatively less variance.

In Figure 4, we see the 95% credible intervals from our old (homoskedastic) and new (heteroskedastic) model. They appear to match well for $\beta_1, \beta_3, \beta_4$; however, there are noticeable changes in estimation for $\beta_2, \beta_5, \beta_6$.

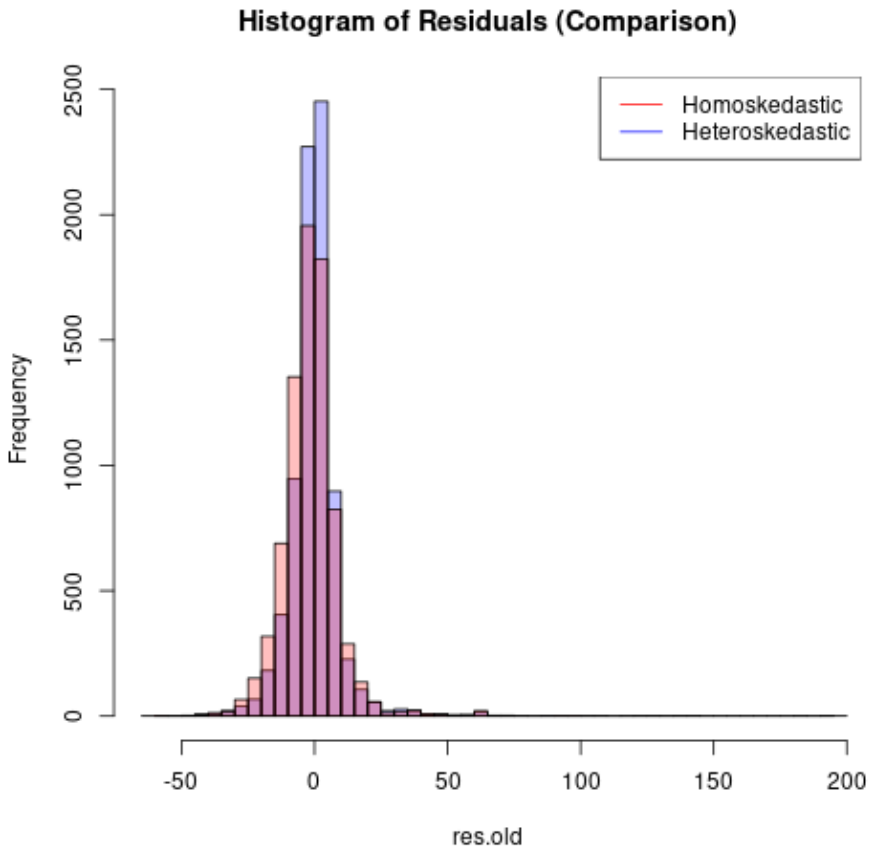

Figure 2: Compare Histograms of Residuals.

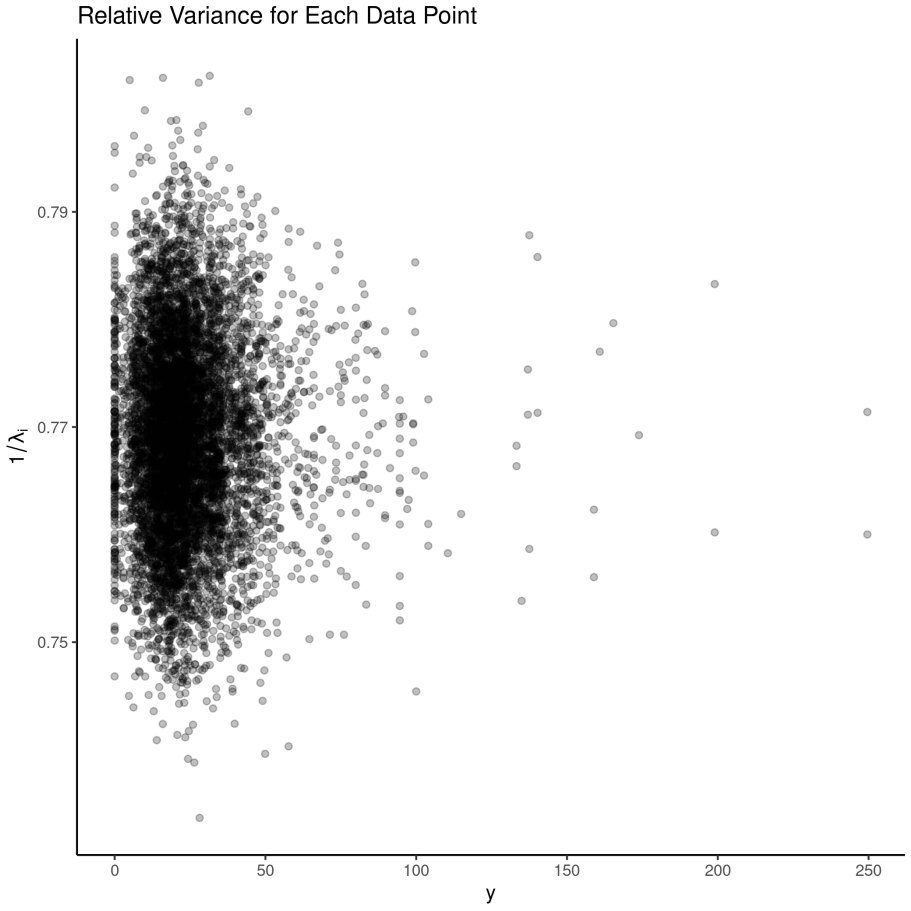

Figure 3: Plot of $\frac{1}{\lambda_i}$ v. y_i.

```
Old Model with  beta 1 : 95% HDI: [-10.00, -7.94]
NewModel with  beta 1 : 95% HDI: [-10.62, -7.29]

Old Model with  beta 2 : 95% HDI: [0.60, 2.32]
NewModel with  beta 2 : 95% HDI: [0.33, 2.40]

Old Model with  beta 3 : 95% HDI: [0.98, 1.02]
NewModel with  beta 3 : 95% HDI: [0.96, 1.05]

Old Model with  beta 4 : 95% HDI: [-0.01, 0.01]
NewModel with  beta 4 : 95% HDI: [-0.01, 0.01]

Old Model with  beta 5 : 95% HDI: [7.63, 9.34]
NewModel with  beta 5 : 95% HDI: [7.28, 9.61]

Old Model with  beta 6 : 95% HDI: [3.28, 4.69]
NewModel with  beta 6 : 95% HDI: [3.02, 5.05]
```

Figure 4: Compare the 95% Credible Intervals Between Models for Each Parameter Term.

A. Show (somewhat trivially) that the maximum likelihood estimate for θ is just the vector of sample means $\hat{\theta}_{MLE} = (\bar{y}_1, ..., \bar{y}_P)$.

First, we get the likelihood and log-likelihood:

$$L(\theta) = \prod_{i=1}^{P} \prod_{j=1}^{N_i} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_{ij} - \theta_i)^2\right\}$$

$$\begin{aligned}\mathcal{L}(\theta) &= \log\left[\prod_{i=1}^{P} \prod_{j=1}^{N_i} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_{ij} - \theta_i)^2\right\}\right] \\ &= \sum_{i=1}^{P} \sum_{j=1}^{N_i} \left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_{ij} - \theta_i)^2\right]\end{aligned}$$

From here, we can maximize the log-likelihood to obtain the MLE for $\theta = (\theta_1, ..., \theta_P)$:

$$\begin{aligned}\frac{\partial}{\partial\theta_i}\mathcal{L}(\theta_i) &= \sum_{j=1}^{N_i} \frac{1}{\sigma^2}(y_{ij} - \theta_i) \\ &\stackrel{set}{=} 0 \\ \hat{\theta}_{MLE} &= \frac{\sum_{j=1}^{N_i} y_{ij}}{N_i} = \bar{y}_i\end{aligned}$$

B. Make a plot that illustrates the following fact: extreme school-level averages $\bar{y}_i$ (both high and low) tend to be at schools where fewer students were sampled. Explain briefly why this would be.

In Figure 1, we see that at schools with a smaller student population (i.e., smaller sample size), the in-school test scores have a large variance; consequently, the school-level averages are more variable. However, as school population increases, we see a regression towards mediocrity in the sense that school-level averages display smaller variance.

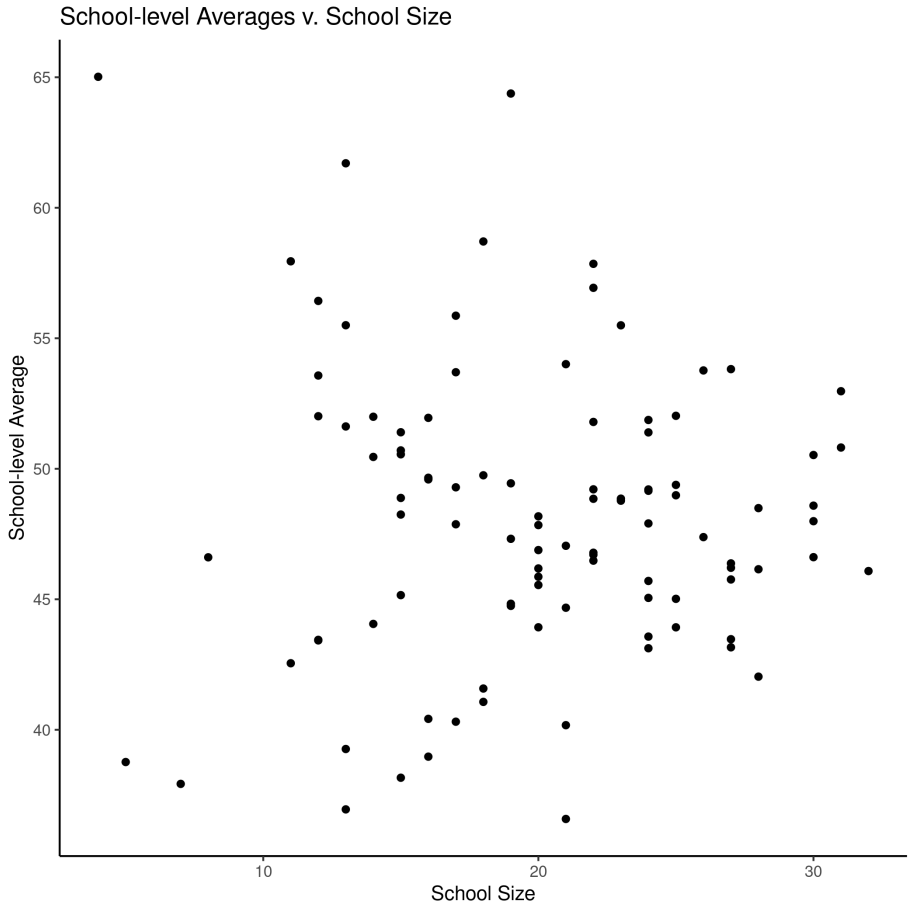

Figure 1: School-level Averages v. Sample Size

C. Fit the following two-level hierarchical model to these data via Gibbs sampling:

$$(y_{ij}|\theta_i, \sigma^2) \sim N(\theta_i, \sigma^2)$$
$$(\theta_i|\tau^2, \sigma^2) \sim N(\mu, \tau^2\sigma^2).$$

As a staring point, use a flat prior on μ, Jeffreys' prior on σ^2, and an IG$(1/2, 1/2)$ prior on τ^2.

Before we fit the model via Gibbs sampler, we need to obtain the full-conditionals for $\mu, \tau^2, \sigma^2, \theta$:

$$\begin{aligned}
p(\mu|\cdot) &\propto \prod_{i=1}^{P} p(\theta_i|\mu, \tau^2\sigma^2) \cdot p(\mu) \\
&\propto \prod_{i=1}^{P} \exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta_i - \mu)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\tau^2\sigma^2}\sum_{i=1}^{P}(\theta_i^2 - 2\mu\theta_i + \mu^2)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[-2\left(\frac{\sum_{i=1}^{P}\theta_i}{\tau^2\sigma^2}\right)\mu + \mu^2\left(\frac{P}{\tau^2\sigma^2}\right)\right]\right\}, \\
p(\theta_i|\cdot) &\propto \prod_{j=1}^{N_i}\left[\exp\left\{-\frac{1}{2\sigma^2}(y_{ij} - \theta_i)^2\right\}\right]\exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta_i - \mu)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{N_i}(-2y_{ij}\theta_i + \theta_i^2)\right\}\exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta_i^2 - 2\mu\theta_i)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[-2\left(\frac{\sum_{j=1}^{N_i}y_{ij}}{\sigma^2} + \frac{\mu}{\tau^2\sigma^2}\right)\theta_i + \theta_i^2\left(\frac{N_i}{\sigma^2} + \frac{1}{\tau^2\sigma^2}\right)\right]\right\}
\end{aligned}$$

so, we see that

$$\begin{aligned}
p(\mu|\cdot) &\equiv N\left(\frac{\sum_{i=1}^{P}\theta_i}{P}, \frac{\tau^2\sigma^2}{P}\right) \\
p(\theta_i|\cdot) &\equiv N\left(\left[\frac{N_i}{\sigma^2} + \frac{1}{\tau^2\sigma^2}\right]^{-1}\left(\frac{\sum_{j=1}^{N_i}y_{ij}}{\sigma^2} + \frac{\mu}{\tau^2\sigma^2}\right), \left[\frac{N_i}{\sigma^2} + \frac{1}{\tau^2\sigma^2}\right]^{-1}\right)
\end{aligned}$$

Next, let's get those full-conditionals for σ^2 and τ^2:

$$\begin{aligned}
p(\sigma^2|\cdot) &\propto \prod_{i=1}^{P}\left[\prod_{j=1}^{N_i}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(y_{ij}-\theta_i)^2\right\}\cdot\frac{1}{\sqrt{2\pi\tau^2\sigma^2}}\exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta_i-\mu)^2\right\}\right]\cdot\frac{1}{\sigma^2} \\
&\propto \sigma^{-2}\prod_{i=1}^{P}\left[\sigma^{-N_i}\exp\left\{-\frac{1}{\sigma^2}\sum_{j=1}^{N_i}\left(\frac{y_{ij}-\theta_i}{2}\right)^2\right\}\cdot\sigma^{-1}\exp\left\{-\frac{1}{\sigma^2}\left(\frac{(\theta_i-\mu)^2}{2\tau^2}\right)\right\}\right] \\
&\propto (\sigma^2)^{-\left(\frac{\sum_{i=1}^{P}\frac{N_i+P}{2}}{2}\right)-1}\exp\left\{-\frac{1}{\sigma^2}\left[\sum_{i=1}^{P}\sum_{j=1}^{N_i}\left(\frac{(y_{ij}-\theta_i)^2}{2}\right)+\sum_{i=1}^{P}\left(\frac{(\theta_i-\mu)^2}{2\tau^2}\right)\right]\right\}, \\
p(\tau^2|\cdot) &\propto \prod_{i=1}^{P}\left[\frac{1}{\sqrt{2\pi\tau^2\sigma^2}}\exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta_i-\mu)^2\right\}\right]\cdot(\tau^2)^{-\frac{1}{2}-1}\exp\left\{-\frac{1}{2\tau^2}\right\} \\
&\propto (\tau^2)^{-\left(\frac{P+1}{2}\right)-1}\exp\left\{-\frac{1}{\tau^2}\left[\frac{\sum_{i=1}^{P}(\theta_i-\mu)^2}{2\sigma^2}+\frac{1}{2}\right]\right\},
\end{aligned}$$

which means that we have the following full-conditionals for σ^2 and τ^2:

$$\begin{aligned}
p(\sigma^2|\cdot) &\equiv \text{IG}\left(\frac{\sum_{i=1}^{P}N_i+P}{2}, \sum_{i=1}^{P}\sum_{j=1}^{N_i}\left(\frac{(y_{ij}-\theta_i)^2}{2}\right)+\sum_{i=1}^{P}\left(\frac{(\theta_i-\mu)^2}{2\tau^2}\right)\right), \\
p(\tau^2|\cdot) &\equiv \text{IG}\left(\frac{P+1}{2}, \frac{\sum_{i=1}^{P}(\theta_i-\mu)^2}{2\sigma^2}+\frac{1}{2}\right)
\end{aligned}$$

Although this portion of the question does not directly ask for any graphics, I figured that I'd post my trace plots here, which can be seen below.

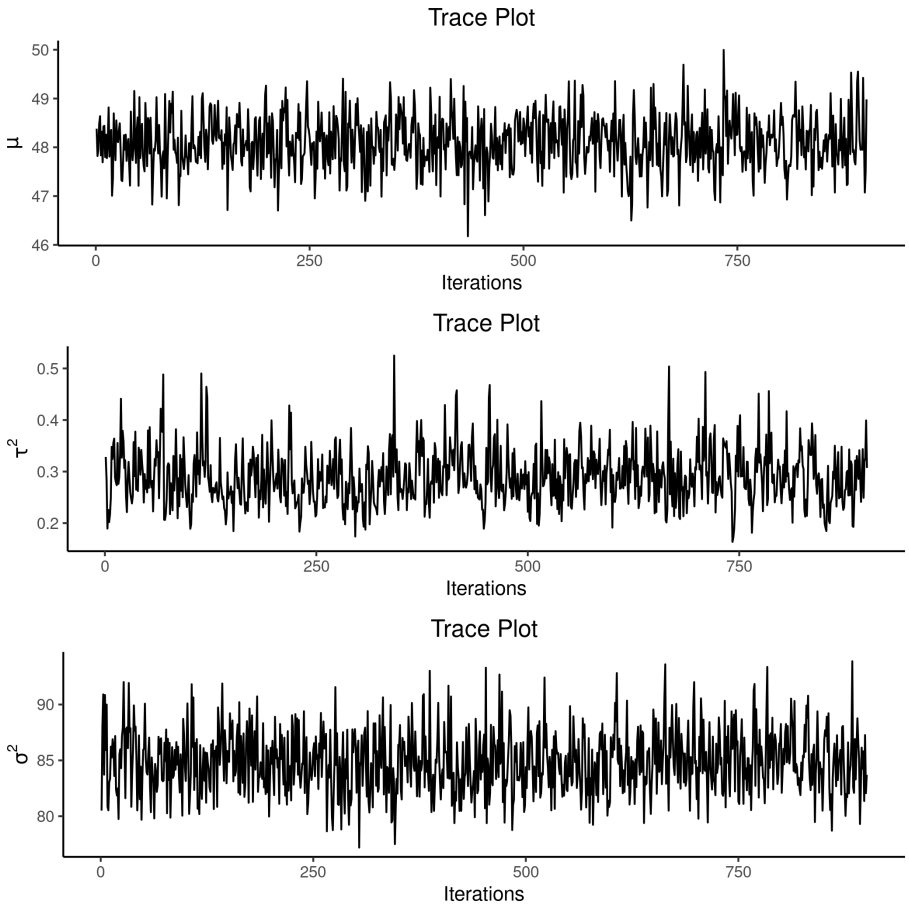

Figure 2: Trace Plots for μ, τ^2, σ^2.

D. Express the conditional posterior mean for θ_i in the following form:

$$E(\theta_i|y, \tau^2, \sigma^2, \mu) = \kappa_i \mu + (1 - \kappa_i)\bar{y}_i,$$

i.e. a convex combination of prior mean and data mean. Here κ_i is a *shrinkage coefficient* whose form you should express in terms of the model hyperparameters. Compute κ_i for each school in your MCMC algorithm.

From (c), we know that the mean in the full-conditional for θ_i is

$$\begin{aligned} E(\theta_i|y, \tau^2, \sigma^2, \mu) &= \left[\frac{N_i}{\sigma^2} + \frac{1}{\tau^2\sigma^2}\right]^{-1} \left(\frac{\sum_{j=1}^{N_i} y_{ij}}{\sigma^2} + \frac{\mu}{\tau^2\sigma^2}\right) \\ &= \left[\frac{N_i\tau^2 + 1}{\tau^2\sigma^2}\right]^{-1} \left(\frac{\sum_{j=1}^{N_i} y_{ij}}{\sigma^2} + \frac{\mu}{\tau^2\sigma^2}\right) \\ &= \frac{\tau^2 \sum_{j=1}^{N_i} y_{ij}}{N_i\tau^2 + 1} + \frac{\mu}{N_i\tau^2 + 1} \\ &= (1 - \kappa_i)\bar{y}_i + \kappa_i\mu, \end{aligned}$$

where $\kappa_i = \frac{1}{1+\tau^2 N_i}$.

E. Observe that an equivalent way to write your model involves the following decomposition:

$$y_{ij} = \mu + \delta_i + e_{ij}$$

where $\delta_i \sim N(0, \tau^2\sigma^2)$ and $e_{ij} \sim N(0, \sigma^2)$. (In the paper by Gelman that I've asked you to read, he writes it this way, where the school-level "offsets" are centered at zero, although he doesn't scale these offsets by σ the way I prefer to do.) To translate between the two parameterizations, just observe that in the previous version, $\theta_i = \mu + \delta_i$. Conditional on the "grand mean" μ, but *marginally* over both δ_i and e_{ij}, compute the following two covariances:

$$\mathbf{cov}(y_{i,j}, y_{i,k}),\ j \neq k$$
$$\mathbf{cov}(y_{i,j}, y_{i',k}),\ i \neq i'\ j \neq k$$

Does this make sense to you? Why or why not?

We can easily derive the desired expressions as follows:

$$\begin{aligned}
\text{cov}(y_{ij}, y_{ik}) &= \text{cov}(\mu + \delta_i + e_{ij}, \mu + \delta_i + e_{ik}) \\
&= E\left[(\delta_i + e_{ij})(\delta_i + e_{ik})\right] \\
&= E(\delta_i^2) \\
&= \tau^2\sigma^2, \\
\text{cov}(y_{ij}, y_{i'k}) &= \text{cov}(\mu + \delta_i + e_{ij}, \mu + \delta_{i'} + e_{i'k}) \\
&= E\left[(\delta_i + e_{ij})(\delta_{i'} + e_{i'k})\right] \\
&= 0,
\end{aligned}$$

where we assume independence between i and i' to obtain the penultimate equality. This makes sense to me since test scores for students in the same school are expected to be correlated while test scores for students at different schools should not be correlated.

F. Does the assumption that σ^2 is common to all schools look justified in light of the data?

It appears that this assumption is appropriate. We account for in-school variability with the scaling τ^2, while variability between schools is modeled with σ^2. This matches our intuition that the variability between schools and the variability within schools should be closely tied together.

A. Is the experimental medication effective at reducing blood pressure? Do the naive thing and perform a t-test for a difference of means, pooling all the data from treatment 1 into group 1, and all the data from treatment 2 into group 2. What does this t-test say about the difference between these two group means and the standard error for the difference? Why is the t-test (badly) wrong?

R output from a t-test for a difference in means can be found in Figure 3; from it, we can see that the difference between these two group means is 9.469 with a standard error of 1.004414. However, we should not trust this inference because we violate an assumption when using the t-test.

The t-test assumes that observations within groups are independent, which is clearly not the case here since we have repeated measurements of the same individuals. Perhaps if we had a single measurement for each individual within the groups, the t-test would be more valid. However, as our data stands, we should not perform a t-test for a difference in means.

```
        Welch Two Sample t-test

data:  group1 and group2
t = 9.4273, df = 391.66, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  7.494212 11.443648
sample estimates:
mean of x mean of y
  142.455   132.986

> test$stderr
[1] 1.004414
```

Figure 3: R Output from a t-test with Pooled Observations.

B. Now, do something better but less than ideal. Calculate $\bar{y}_i$, the mean blood pressure measurement for each patient. Then, treat each person-level mean as if it were just a single data point, and conduct a different t-test for mean blood pressure between treatment 1 and treatment 2. What does this t-test say about the difference between these two group means and the standard error for the difference? Why is the standard error so much bigger, and why is this appropriate? Even so, why is this approach (subtly) wrong?

R output from a t-test for a difference in means using average blood pressure measurements for each patient can be found in Figure 4. In it, we can see an estimated difference between group means of 7.416 and a standard error of 4.511762.

The standard error here is bigger than it was in (A) because we are using less observations for our t-test, which more closely reflects reality. Although we have a total of 426 observations in this data set, we only have 20 individuals with repeated samples. Therefore, our sample size is actually much smaller than it first seems. For that reason, our standard error will be larger than it was when we told our t-test that we had a sample size of 426.

While a better approach to the t-test, this testing methodology is still not that great because we are not obtaining standard errors between individuals and between groups. To fully understand the data, we should account for differences within-subjects and within-groups. Here, we are just accounting for difference between groups.

```
        Welch Two Sample t-test

data:  group1 and group2
t = 1.6437, df = 17.09, p-value = 0.1185
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.099195 16.931139
sample estimates:
mean of x mean of y
 141.5435  134.1275

> testb$stderr
[1] 4.511762
```

Figure 4: R Output from a t-test with Mean Observations.

C. Now fit a two-level hierarchical model to this data, of the following form:

$$(y_{ij} \mid \theta_i, \sigma^2) \sim \mathbf{N}(\theta_i, \sigma^2)$$
$$(\theta_i \mid \tau^2, \sigma^2) \sim \mathbf{N}(\mu + \beta x_i, \tau^2\sigma^2)$$

where y_{ij} is blood pressure measurement j on person i, and x_i is a dummy (0/1) variable indicating whether a patient received treatment 2, the experimental medication. Apply what you learned on the previous problem about sampling, hyperparameters, etc, but account for the extra wrinkle here, i.e. the presence of the βx_i term that shifts the mean between the treatment and control groups.

Write out your model's complete conditional distributions, and fit it. Make a histogram of the posterior distribution for β, which represents the treatment effect here. In particular, what are the posterior mean and standard deviation of β? How do these compare to the estimates and standard errors from the approaches in (A) and (B)?

Before we fit the model via Gibbs sampler, we need to obtain the full-conditionals for $\mu, \tau^2, \sigma^2, \theta, \beta$:

$$\begin{aligned}
p(\mu|\cdot) &\propto \prod_{i=1}^{P} p(\theta_i|\mu, \tau^2\sigma^2) \cdot p(\mu) \\
&\propto \prod_{i=1}^{P} \exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta_i - \mu - \beta x_i)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\tau^2\sigma^2}\sum_{i=1}^{P}(-2\theta_i\mu + \mu^2 + 2\mu\beta x_i)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[-2\left(\frac{\sum_{i=1}^{P}\theta_i - \beta\sum_{i=1}^{P}x_i}{\tau^2\sigma^2}\right)\mu + \mu^2\left(\frac{P}{\tau^2\sigma^2}\right)\right]\right\}, \\
p(\theta_i|\cdot) &\propto \prod_{j=1}^{N_i}\left[\exp\left\{-\frac{1}{2\sigma^2}(y_{ij} - \theta_i)^2\right\}\right]\exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta_i - \mu - \beta x_i)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{N_i}(-2y_{ij}\theta_i + \theta_i^2)\right\}\exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta_i^2 - 2\mu\theta_i - 2\beta x_i\theta_i)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[-2\left(\frac{\sum_{j=1}^{N_i}y_{ij}}{\sigma^2} + \frac{\mu + \beta x_i}{\tau^2\sigma^2}\right)\theta_i + \theta_i^2\left(\frac{N_i}{\sigma^2} + \frac{1}{\tau^2\sigma^2}\right)\right]\right\}
\end{aligned}$$

so, we see that

$$p(\mu|\cdot) \equiv N\left(\frac{\sum_{i=1}^{P}\theta_i - \beta\sum_{i=1}^{P}x_i}{P}, \frac{\tau^2\sigma^2}{P}\right)$$
$$p(\theta_i|\cdot) \equiv N\left(\left[\frac{N_i}{\sigma^2} + \frac{1}{\tau^2\sigma^2}\right]^{-1}\left(\frac{\sum_{j=1}^{N_i}y_{ij}}{\sigma^2} + \frac{\mu+\beta x_i}{\tau^2\sigma^2}\right), \left[\frac{N_i}{\sigma^2} + \frac{1}{\tau^2\sigma^2}\right]^{-1}\right)$$

Next, let's get those full-conditionals for σ^2 and τ^2:

$$\begin{aligned}
p(\sigma^2|\cdot) &\propto \prod_{i=1}^{P}\left[\prod_{j=1}^{N_i}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(y_{ij}-\theta_i)^2\right\}\cdot\frac{1}{\sqrt{2\pi\tau^2\sigma^2}}\exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta_i-\mu-\beta x_i)^2\right\}\right]\cdot\frac{1}{\sigma^2} \\
&\propto \sigma^{-2}\prod_{i=1}^{P}\left[\sigma^{-N_i}\exp\left\{-\frac{1}{\sigma^2}\sum_{j=1}^{N_i}\left(\frac{y_{ij}-\theta_i}{2}\right)^2\right\}\cdot\sigma^{-1}\exp\left\{-\frac{1}{\sigma^2}\left(\frac{(\theta_i-\mu-\beta x_i)^2}{2\tau^2}\right)\right\}\right] \\
&\propto (\sigma^2)^{-\left(\frac{\sum_{i=1}^{P}N_i+P}{2}\right)-1}\exp\left\{-\frac{1}{\sigma^2}\left[\sum_{i=1}^{P}\sum_{j=1}^{N_i}\left(\frac{(y_{ij}-\theta_i)^2}{2}\right)+\sum_{i=1}^{P}\left(\frac{(\theta_i-\mu-\beta x_i)^2}{2\tau^2}\right)\right]\right\}, \\
p(\tau^2|\cdot) &\propto \prod_{i=1}^{P}\left[\frac{1}{\sqrt{2\pi\tau^2\sigma^2}}\exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta_i-\mu-\beta x_i)^2\right\}\right]\cdot(\tau^2)^{-\frac{1}{2}-1}\exp\left\{-\frac{1}{2\tau^2}\right\} \\
&\propto (\tau^2)^{-\left(\frac{P+1}{2}\right)-1}\exp\left\{-\frac{1}{\tau^2}\left[\frac{\sum_{i=1}^{P}(\theta_i-\mu-\beta x_i)^2}{2\sigma^2}+\frac{1}{2}\right]\right\},
\end{aligned}$$

which means that we have the following full-conditionals for σ^2 and τ^2:

$$p(\sigma^2|\cdot) \equiv \text{IG}\left(\frac{\sum_{i=1}^{P}N_i+P}{2}, \sum_{i=1}^{P}\sum_{j=1}^{N_i}\left(\frac{(y_{ij}-\theta_i)^2}{2}\right)+\sum_{i=1}^{P}\left(\frac{(\theta_i-\mu-\beta x_i)^2}{2\tau^2}\right)\right),$$
$$p(\tau^2|\cdot) \equiv \text{IG}\left(\frac{P+1}{2}, \frac{\sum_{i=1}^{P}(\theta_i-\mu-\beta x_i)^2}{2\sigma^2}+\frac{1}{2}\right)$$

Lastly, we need to obtain the full-conditional for our β parameter, which we assume has a normal prior with mean μ_β and variance σ_β^2:

$$\begin{aligned}
p(\beta|\cdot) &\propto \prod_{i=1}^{P}\left[\exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta_i-\mu-\beta x_i)^2\right\}\right]\cdot\exp\left\{-\frac{1}{2\sigma_\beta^2}(\beta-\mu_\beta)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\tau^2\sigma^2}\sum_{i=1}^{P}(-2\theta_i x_i\beta+2\mu x_i\beta+\beta^2x_i^2)\right\}\cdot\exp\left\{-\frac{1}{2\sigma_\beta^2}(\beta^2-2\mu_\beta\beta)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[-2\left(\frac{\sum_{i=1}^{P}(\theta_i-\mu)x_i}{\tau^2\sigma^2}+\frac{\mu_\beta}{\sigma_\beta^2}\right)\beta+\beta^2\left(\frac{\sum_{i=1}^{P}x_i^2}{\tau^2\sigma^2}+\frac{1}{\sigma_\beta^2}\right)\right]\right\},
\end{aligned}$$

which is the kernel of a Normal distribution, so

$$p(\beta|\cdot) \equiv \text{Normal}\left(\left(\frac{\sum_{i=1}^{P} x_i^2}{\tau^2\sigma^2} + \frac{1}{\sigma_\beta^2}\right)^{-1}\left(\frac{\sum_{i=1}^{P}(\theta_i - \mu)x_i}{\tau^2\sigma^2} + \frac{\mu_\beta}{\sigma_\beta^2}\right), \left(\frac{\sum_{i=1}^{P} x_i^2}{\tau^2\sigma^2} + \frac{1}{\sigma_\beta^2}\right)^{-1}\right),$$

The posterior mean for β is -7.8 and the posterior standard deviation for β is 4.78 with $M = 1000$ iterations, seed 702, $\mu_\beta = 0$, $\sigma_\beta^2 = 10^9$, and a burn-in period of 100 iterations. A histogram of the posterior distribution for β can be found in Figure 5. Additionally, trace plots for the parameters in our hierarchical model can be found in Figure 6 for diagnostics.

In comparison to (A) and (B), where the difference in means was estimated to be roughly 8, our posterior estimate for β is -7.8; this means that if someone takes the treatment, they will experience a drop of 7 units in systolic measurements on average. This is relatively similar to the inference we obtain from (A) and (B). Additionally, our standard error is nearly identical to that in (A) and (B).

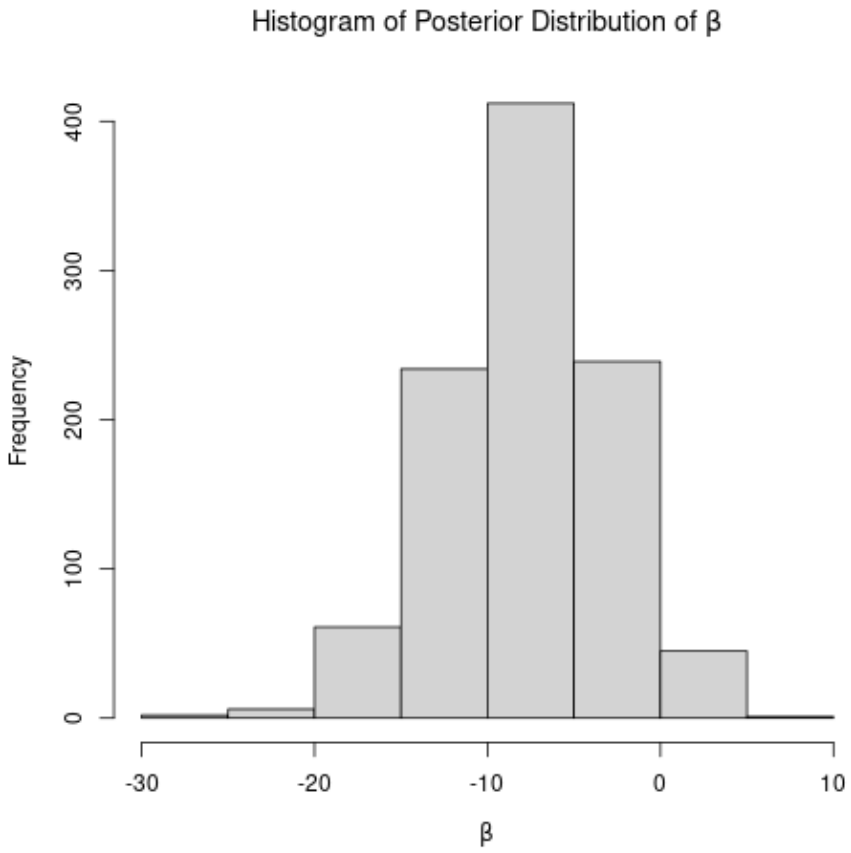

Figure 5: Histogram of Posterior for β.

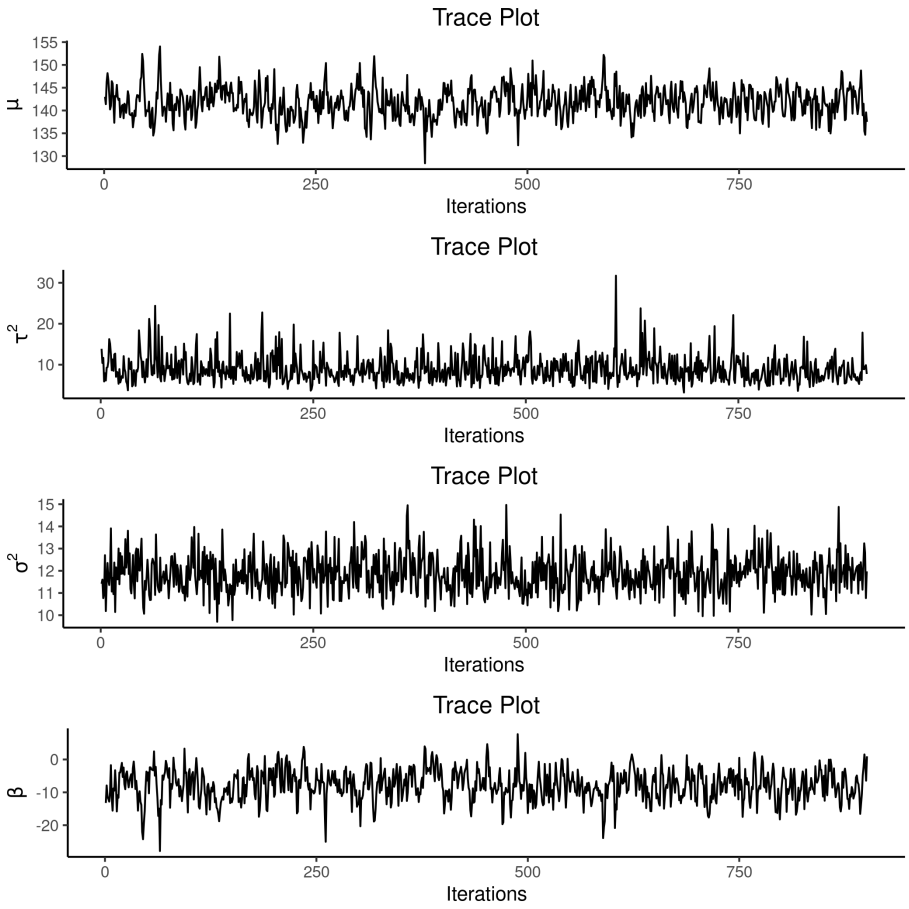

Figure 6: TRACES OF BLOOD!

D. Your two-level model assumes that, conditional on θ_i, the y_{ij} are independent. Written concisely: $(y_{ij} \perp y_{ik} \mid \theta_i)$ for $j \neq k$. There are many ways this assumption could break down. So check! Does this assumption look (approximately) sensible in light of the data? Provide evidence one way or another.

I believe this assumption to be sensible since blood pressure measurements appear to not be autocorrelated within individuals. For example, subject 2 in Figure 7 appears to have the most autocorrelation in their blood measurements; however, when looking at their autocorrelation plot in Figure 8. This indicates to me that, conditional on θ_i, the y_{ij} are relatively independent.

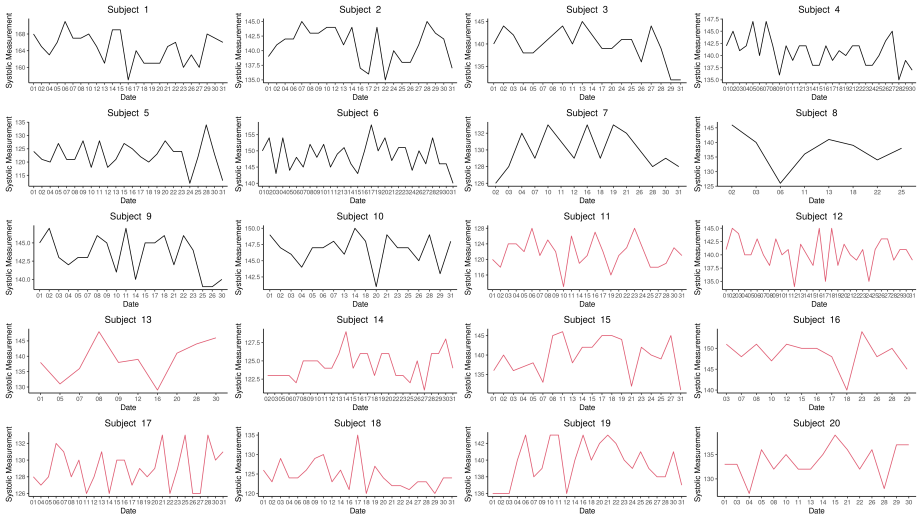

Figure 7: Subject-level blood measurements.

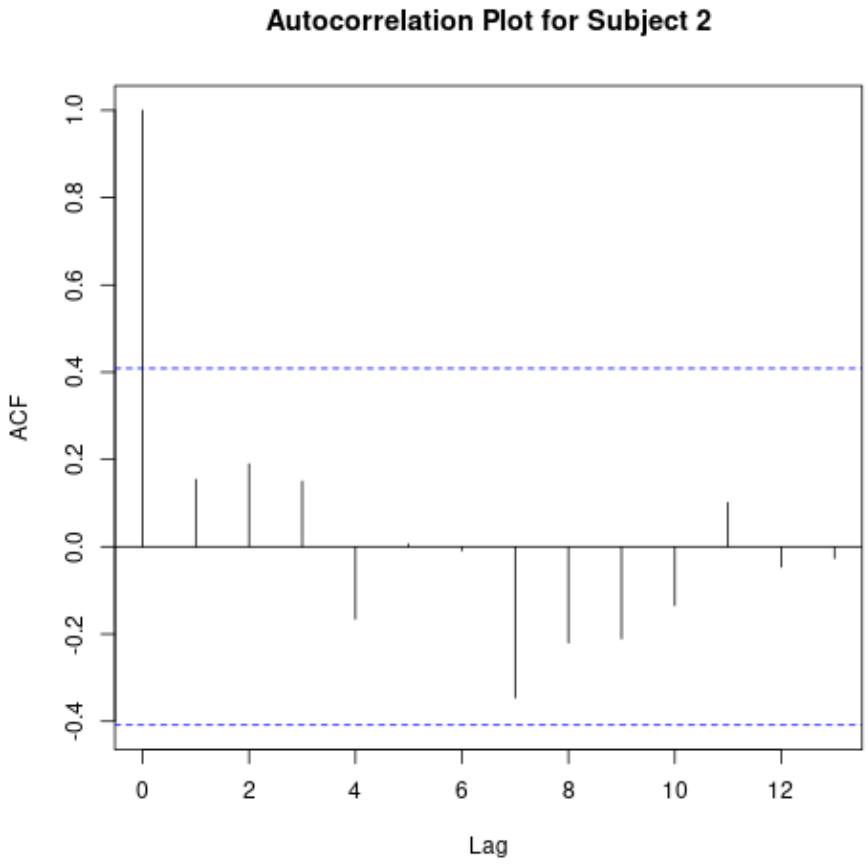

Figure 8: Autocorrelation Plot for Subject 2.

A. The data in "cheese.csv" are about sales volume, price, and advertisting display activity for packages of Borden sliced "cheese." The data are taken from Rossi, Allenby, and McCulloch's textbook on *Bayesian Statistics and Marketing*. For each of 88 stores (store) in different US cities, we have repeated observations of the weekly sales volume (vol, in terms of packages sold), unit price (price), and whether the product was advertised with an in-store display during that week (disp = 1 for display). Your goal is to estimate, on a store-by-store basis, the effect of display ads on the demand curve for cheese. A standard form of a demand curve in economics is of the form $Q = \alpha P^{\beta}$, where Q is quantity demanded (i.e. sales volume), P is price, and α and β are parameters to be estimated. You'll notice that this is linear on a log-log scale,

$$\log Q = \log \alpha + \beta \log P$$

which you should feel free to assume here. Economists would refer to β as the price elasticity of demand (PED). Notice that on a log-log scale, the errors enter multiplicatively. There are several things for you to consider in analyzing this data set.

1. **The demand curve might shift (different α) and also change shape (different β) depending on whether there is a display ad or not in the store.**
2. **Different stores will have very different typical volumes, and your model should account for this.**
3. **Do different stores have different PEDs? If so, do you really want to estimate a separate, unrelated β for each store?**
4. **If there is an effect on the demand curve due to showing a display ad, does this effect differ store by store, or does it look relatively stable across stores?**
5. **Once you build the best model you can using the log-log specification, do see you any evidence of major model mis-fit?**

Propose an appropriate hierarchical model that allows you to address these issues, and use Gibbs sampling to fit your model.

Let's index stores by $i = 1, ..., n$ and observations per store $j = 1, ..., N_i$. Note that there may be a different number of observations per store. The quantity demanded at store i and observation j is denoted Q_{ij}, and similar use of indexes applies for price P and the display covariate. Using the standard form of the economic demand curve given above, we can denote our sampling model as

$$\log Q_{ij} = (\beta_0)_i + (\beta_1)_i \log P_{ij} + (\beta_2)_i \mathbb{1}(\text{disp}_{ij} = 1) + (\beta_3)_i \log P_{ij} \cdot \mathbb{1}(\text{disp}_{ij} = 1)$$

We can rewrite the above as a linear regression model:

$$y_{ij} = X_{ij}^T \beta_i + \epsilon_{ij},$$

where

$$\begin{aligned} y_{ij} &= \log Q_{ij} \\ X_{ij} &= \left(1,\ \log P_{ij},\ \mathbb{1}(\text{disp}_{ij}),\ \log P_{ij} \cdot \mathbb{1}(\text{disp}_{ij})\right)^T \\ \beta_i &= ((\beta_0)_i, (\beta_1)_i, (\beta_2)_i, (\beta_3)_i)^T \end{aligned}$$

This linear regression model can be extended to a hierarchical model with the following specification, where priors were selected to primarily induce conjugate as well as allow the data to speak for itself:

$$\begin{aligned} [y_{ij} \mid \beta_i, \sigma_i^2] &\equiv \text{Normal}\left(X_{ij}^T \beta_i, \sigma_i^2\right), \\ [\beta_i \mid \mu_\beta, \Sigma] &\equiv \text{Normal}\left(\mu_\beta, \Sigma = \text{diag}(s_1^2, \ldots, s_4^2)\right), \\ [\mu_\beta] &\propto 1, \\ [s_p^2] &\equiv \text{Inv-Ga}\left(\frac{1}{2}, \frac{1}{2}\right), \\ [\sigma_i^2] &\equiv \text{Inv-Ga}\left(\frac{a}{2}, \frac{b}{2}\right), \\ [a] &\equiv \text{Gamma}(3, 1), \\ [b] &\equiv \text{Gamma}(3, 1), \end{aligned}$$

where $p = 1, ..., P$. Here, we have four covariates, so $P = 4$. We use an improper prior for μ_β to avoid providing *a priori* information. With this model specification, we are allowing the demand curve to shift (different α) and change shape (different β) depending on whether a display ad was in effect. Inference on β_2 and β_3 will inform us regarding the marginal effect of advertising for α and β, respectively. Additionally, by varying β_0 and β_1 by store, we assume that different stores have different typical volumes and different PEDs.

To implement this hierarchical regression model in an MCMC algorithm, we need to obtain those tasty full-conditionals. The full-conditional for β_i is

$$\begin{aligned}
[\beta_i|\cdot] &\equiv \text{Normal}\left(\mu_i^*, \Sigma_i^*\right), \\
\Sigma_i^* &= \left(\frac{X_i^T X_i}{\sigma_i^2} + \Sigma^{-1}\right)^{-1}, \\
\mu_i^* &= \Sigma_i^* \left(\frac{X_i^T Y_i}{\sigma_i^2} + \Sigma^{-1}\mu_\beta\right),
\end{aligned}$$

where $X_i = [X_{i,1}^T, ..., X_{i,N_i}^T]^T$ and $Y_i = [Y_{i,1}, ..., Y_{i,N_i}]^T$.

The full-conditional for μ_β is

$$\begin{aligned}
[\mu_\beta|\cdot] &\equiv \text{Normal}\left(\bar{\beta}, \frac{1}{n}\Sigma\right), \\
\bar{\beta} &= \frac{1}{n}\sum_{i=1}^{n}\beta_i
\end{aligned}$$

The full-conditional for s_p^2 is

$$[s_p^2|\cdot] \equiv \text{Inverse-Gamma}\left(\frac{n}{2} + \frac{1}{2}, \frac{1}{2}\left(1 + \sum_{i=1}^{n}(\beta_{ip} - \mu_{\beta_p})^2\right)\right)$$

The full-conditional for σ_i^2 is

$$[\sigma_i^2|\cdot] \equiv \text{Inverse-Gamma}\left(\frac{a}{2} + \frac{N_i}{2}, \frac{1}{2}\left(b + (Y_i - X_i\beta_i)^T(Y_i - X_i\beta_i)\right)\right)$$

Note that the full-conditionals for a and b are not conjugate. Therefore, we will update these parameters with Metropolis-Hastings ratios. Consider using a Uniform$(1, 10)$ proposal for both a and b. Then, their Metropolis-Hastings ratios would look like

$$\begin{aligned}
mh_a &= \frac{\prod_{i=1}^{n}\left([\sigma_i^2|a^{(*)}, b]\right)[a^{(*)}][a^{(k-1)}|a^{(*)}]}{\prod_{i=1}^{n}\left([\sigma_i^2|a^{(k-1)}, b]\right)[a^{(k-1)}][a^{(*)}|a^{(k-1)}]}, \\
mh_b &= \frac{\prod_{i=1}^{n}\left([\sigma_i^2|b^{(*)}, a]\right)[b^{(*)}][b^{(k-1)}|b^{(*)}]}{\prod_{i=1}^{n}\left([\sigma_i^2|b^{(k-1)}, a]\right)[b^{(k-1)}][b^{(*)}|b^{(k-1)}]}
\end{aligned}$$

With these updates in hand, we can implement our MCMC algorithm and obtain the estimates for σ_i^2 and β_i.

First, let's examine the trace plots for $\mu_{\beta,p}$ and s_p^2 in Figure 1. We appear to learn very well and see good mixing. This is to be expected since we have plenty of data to learn about these eight parameters across the many sites. The trace plots are the result of a burn-in period of 3000 iterations and thinning every other iteration.

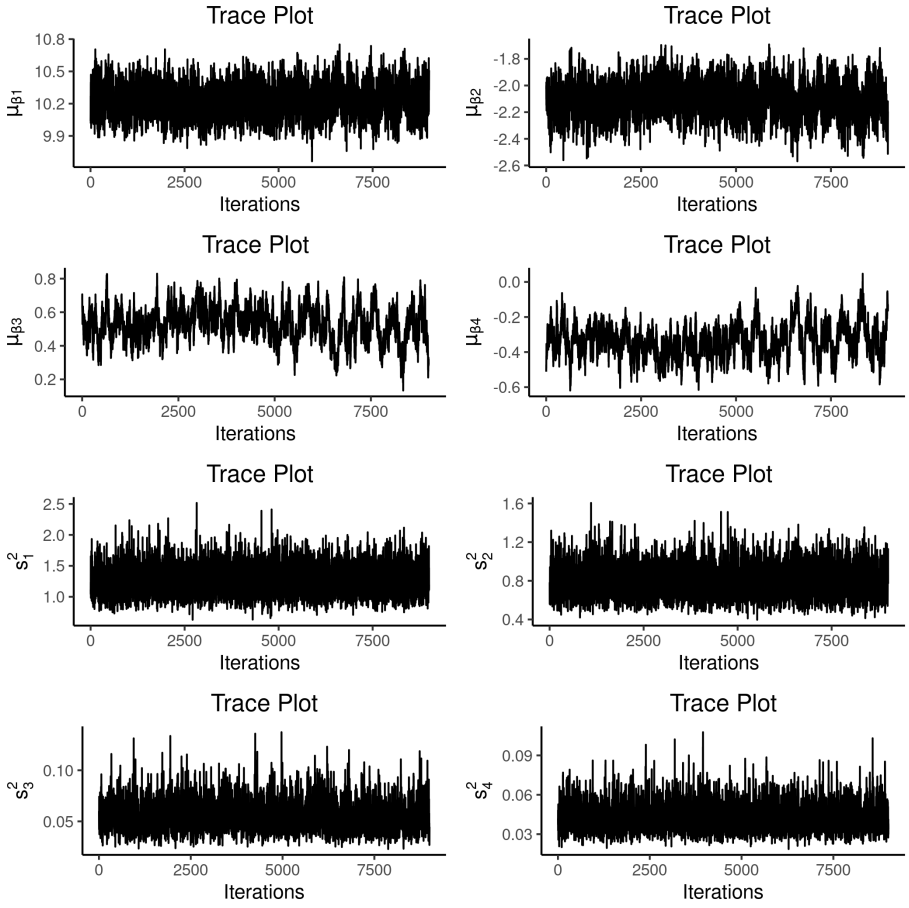

Figure 1: Trace Plots for s_p^2 and μ_β.

Now, let's take a look at a plot of σ_i^2 per store; note that the ordering does not mean anything in Figure 2. We can see that the store-level variance for volume sold varies only slightly; estimates range from 0.01 to 0.41. The relatively low variance for each store is likely due to the relatively large number of observations per store. Note that different numbers of observations (i.e. N_i) per store may lead to different values of σ_i^2. Therefore, it appears that modeling per-store variance was an appropriate move, as having a global variance parameter would not enable us to see the different in per-store variance, albeit the difference is quite small.

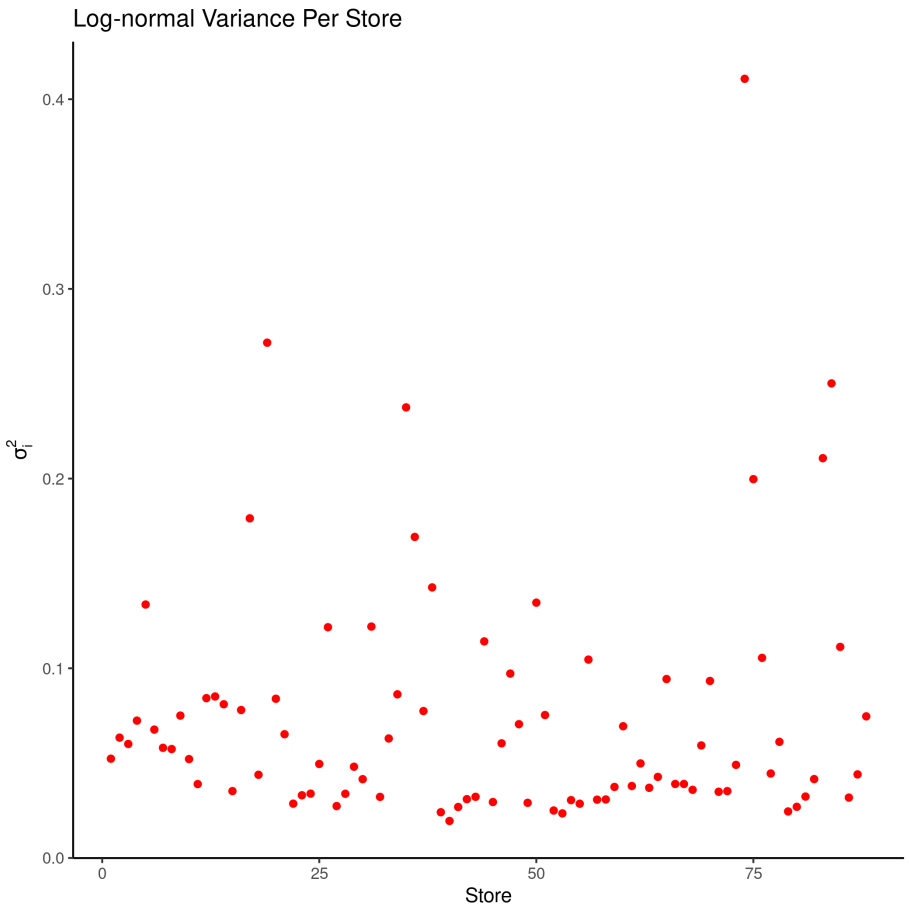

Figure 2: σ_i^2 v. Store.

Now, let's examine the histograms of the β values. In Figure 3, we obtain histograms for the β coefficients for the four parameters across all stores. We can interpret each of these parameters in the log-quantity sold space. It appears that the typical intercept for the demand curve is roughly 10 for each store. While there is some variability, the standard deviation is quite low. The marginal shift in the demand curve due to advertising is roughly 0.5 for each store, indicating that advertising leads to an increase in volume sold. The change in shape for the demand curve given a unit change in price appears to be -2 for each store, indicating that a unit increase in log-price will decrease log-volume sold by 2 units. Finally, the marginal change in shape when advertising appears to be -0.5 per store, indicating that an advertisement can not offset an increase in price on the volume sold.

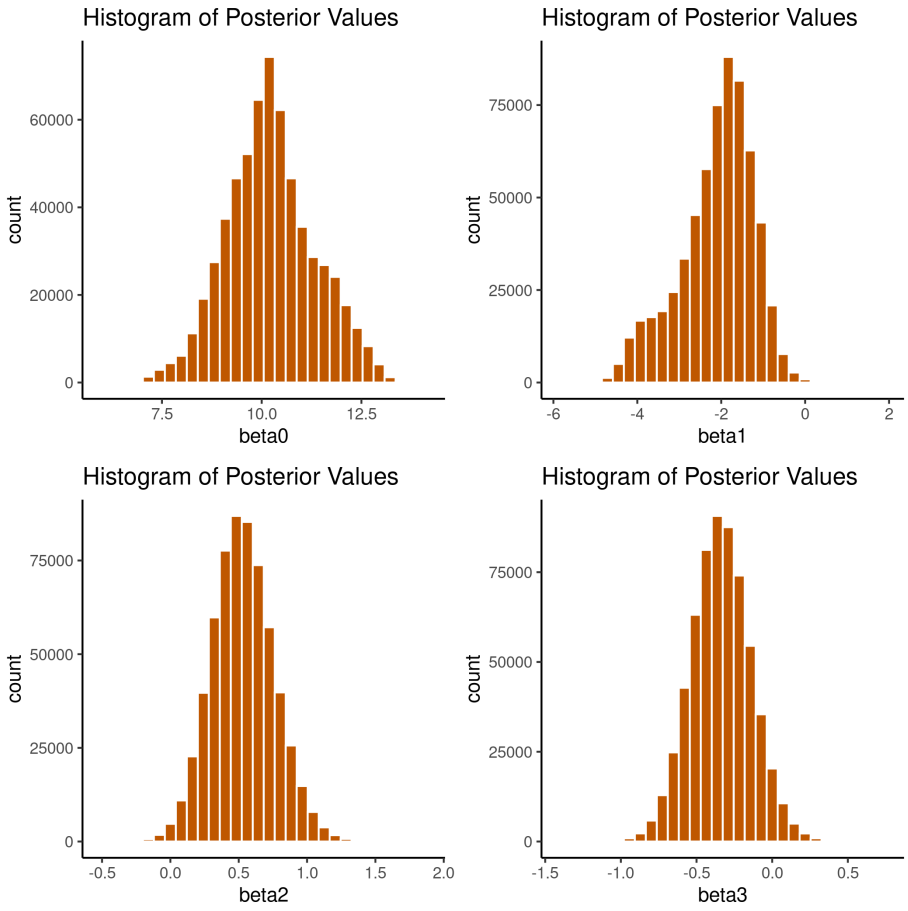

Figure 3: Histogram for β Coefficients.

Finally, let's examine the posterior means for β_2 and β_3 for each store in Figures 4 and 5, respectively. Once again, the ordering of these points does not imply any relationship. Rather, we can tell from the difference in heights of the points that each store has a different posterior estimate for both β_2 and β_3. It appears that the posterior estimates for β_2 vary slightly more than those for β_3. This indicates that the effect on the demand curve when advertising is different across each store. However, one may argue that the difference is not entirely significant. For example, we found that the intercept $\beta_0 \approx 10$, so a change of 0.4 does not greatly affect the overall intercept of the model when advertising.

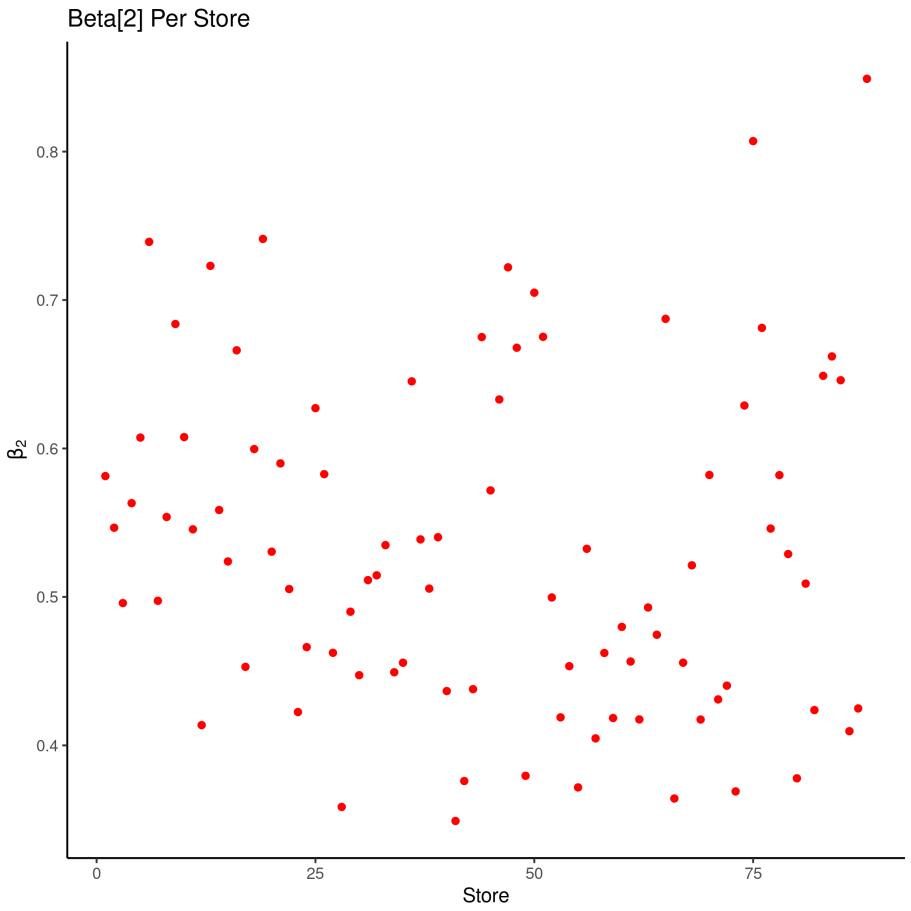

Figure 4: Estimates of β_2 per store.

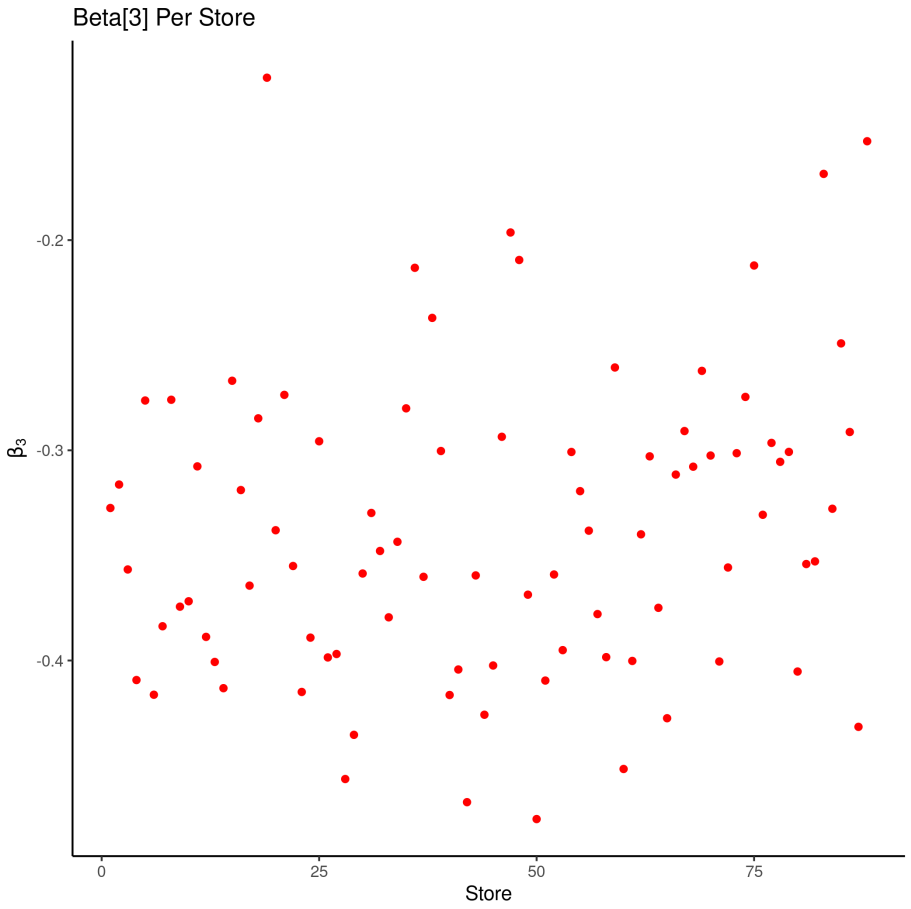

Figure 5: Estimates of β_3 per store.

Now, I address each of the five considerations given in the problem statement.

1. By the model specification, we allow the demand curve to shift and change shape depending on whether or not a display was present in the store. We do this by including $(\beta_2)_i$ and $(\beta_3)_i$ in our model.
2. We account for the difference in typical volumes between stores by having the coefficient $(\beta_0)_i$ vary depending on the store.
3. We account for the difference in PEDs by including the coefficient $(\beta_1)_i$ per store.
4. It appeared that the effect on the demand curve due to showing a display ad was very similar across stores relative to the quantities of interest.
5. When fitting this model, I saw a case of major model mis-fit in the quite terrible updating of a and b; see Figure 6. Evidently, we can obtain much better mixing and exploration than our current Metropolis-Hastings scheme. In the future, I would try different proposal distributions for a and b, as I do not suspect the prior for a or b to be the main culprit.

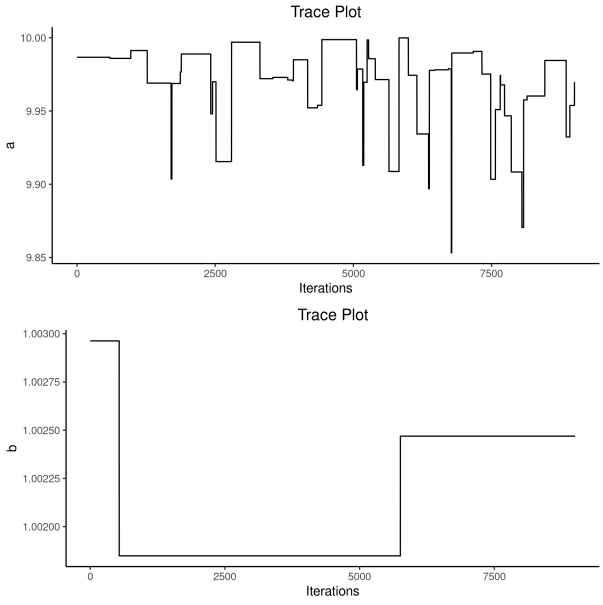

Figure 6: Trace plots for a and b.

A. In "polls.csv" you will find the results of several political polls from the 1988 U.S. presidential election. The outcome of interest is whether someone plans to vote for George Bush (senior, not junior). There are several potentially relevant demographic predictors here, including the respondent's state of residence. The goal is to understand how these relate to the probability that someone will support Bush in the election. You can imagine this information would help a great deal in poll re-weighting and aggregation (ala Nate Silver). Use Gibbs sampling, together with the Albert and Chib trick, to fit a hierarchical probit model of the following form:

$$\begin{aligned} \mathbf{Pr}(y_{ij} = 1) &= \Phi(z_{ij}) \\ z_{ij} &= \mu_i + x_{ij}^T \beta_i \,. \end{aligned}$$

Here y_{ij} is the response (Bush=1, other=0) for respondent j in state i; $\Phi(\cdot)$ is the probit link function, i.e. the CDF of the standard normal distribution; μ_i is a state-level intercept term; x_{ij} is a vector of respondent-level demographic predictors; and β_i is a vector of regression coefficients for state i.

Of the provided covariates, I decided to use age, education, race, and sex. I converted the categorical data of age and education into three indicator variables each. This decision was made so that we are able to see the marginal effect of more education or age on voting outcome. My hierarchical model is an extension of that found in Albert & Chib (1993):

$$\begin{aligned} P(Y_{ij} = 1) = \Phi(X_{ij}^T \alpha_i), \\ \alpha_i \sim \text{Normal}(\beta_i^*, B^*), \\ \beta_i^* \sim \text{Normal}(0, 10^4 \cdot I_{P+1}), \\ B^* \sim \text{Inverse-Wishart}(P+2, I_{P+1}), \end{aligned}$$

where $i = 1, ..., n$ indexes the n states in the data, $j = 1, ..., N_i$ indexes the individuals in each state, and there are P covariates and an intercept in the model. I used a multivariate normal prior on β_i^* and an inverse-Wishart on B^* to exploit conjugacy. With this hierarchical model, we get the following

conjugate posteriors:

$$
\begin{aligned}
[\alpha_i|\cdot] &\equiv \text{Normal}(\tilde{\beta}, \tilde{B}), \\
\tilde{B} &= \left[(B^*)^{-1} + X_i^T X_i\right]^{-1}, \\
\tilde{\beta} &= \tilde{B}\left[(B^*)^{-1}\beta_i^* + X_i^T Z_i\right], \\
[Z_{ij}|\cdot] &\equiv \begin{cases} \text{Normal}(X_{ij}^T\alpha_1, 1)\Big|_{[0,\infty)} & , Y_{ij=1} \\ \text{Normal}(X_{ij}^T\alpha_1, 1)\Big|_{(-\infty,0]} & , Y_{ij=0} \end{cases}, \\
[\beta_i^*|\cdot] &\equiv \text{Normal}(A^{-1}b, A^{-1}), \\
A &= (B^*)^{-1} + (10^4 \cdot I_{P+1})^{-1}, \\
b^T &= \alpha_i^T (B^*)^{-1}, \\
[B^*|\cdot] &\equiv \text{Inverse-Wishart}\left(n + P + 2, I_{P+1} + \sum_{i=1}^{n}(\alpha_i - \beta_i^*)(\alpha_i - \beta_i^*)^T\right)
\end{aligned}
$$

After removing the NA values in the provided data set, I was able to fit this model and obtain the posterior mean estimates for the coefficients in Figure 7. In this figure, we can see the intercept μ and coefficients β_2 through β_9.

Based off the way I set up my indicator functions μ_i is the mean of a non-Black, 65+ year-old, no-high-school male's propensity to vote for Bush in state i. The marginal change in voting propensity for Bush as a woman in state i is measured by β_2. The marginal change for a black American is measured by β_3. The marginal change if someone were to have a Bachelor's degree, high school, or some college is measured with β_4, β_5 and β_6, respectively. The marginal change if someone were to be 18-29, 30-44, or 45-64 are measured by β_7, β_8, and β_9, respectively.

Note that not every state is in each of the subplots in Figure 7. This is because some states did not have enough data to accurately estimate some parameters. For example, Maine did not have a complete observation for a black American; therefore, "ME" does not show up in the subplot for β_3. As one would expect, black voters tend to vote more democratic (blue) than non-black voters. This is why most of the points in the subplot for β_3 are blue.

Just for fun, I included trace plots for the beta estimates for the states of Texas (Figure 8) and New York (Figure 9). With enough data, we are able to accurately estimate the coefficients. For example, New York was a swing state with a margin of just 4.10% for Michael Dukakis. Despite the relatively tight race, we have enough points to robustly estimate the β's.

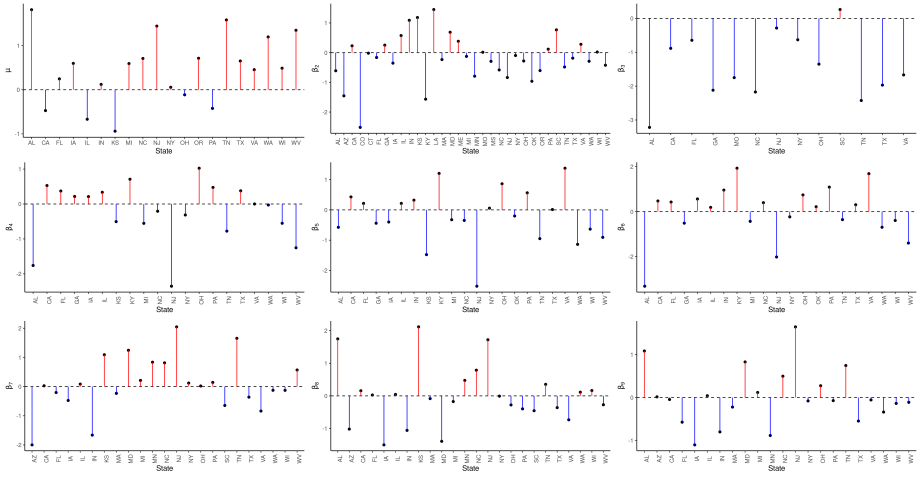

Figure 7: Posterior Mean Estimates for Coefficients.

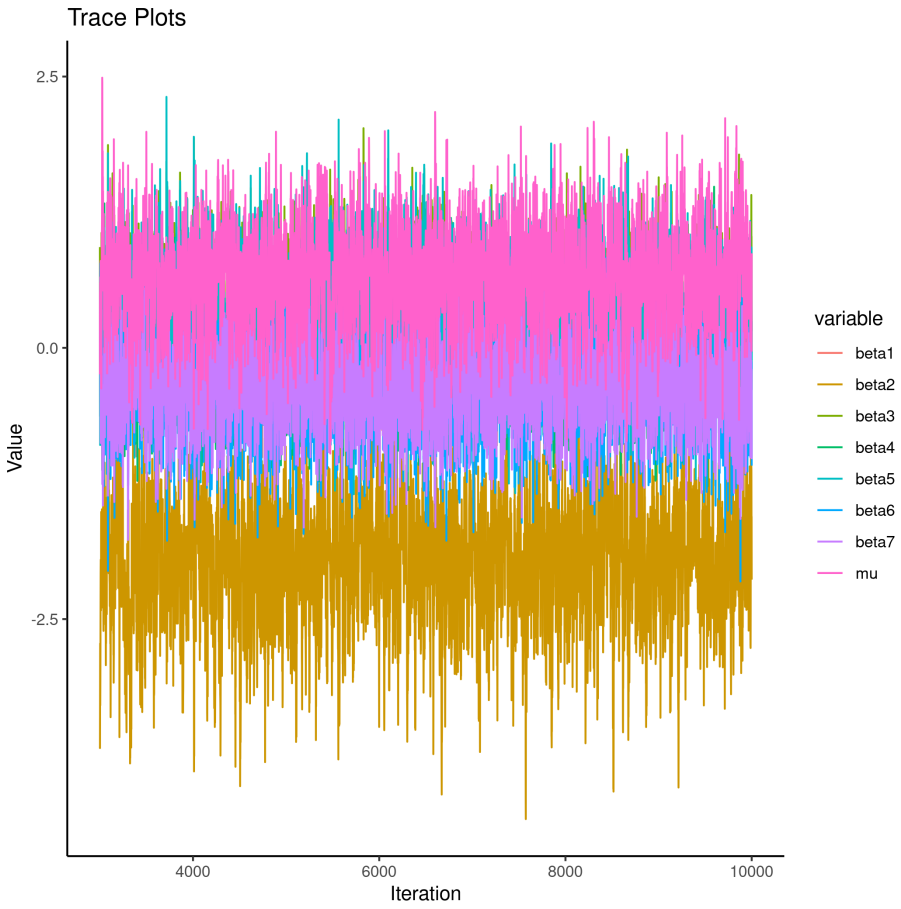

Figure 8: Estimates for Texas.

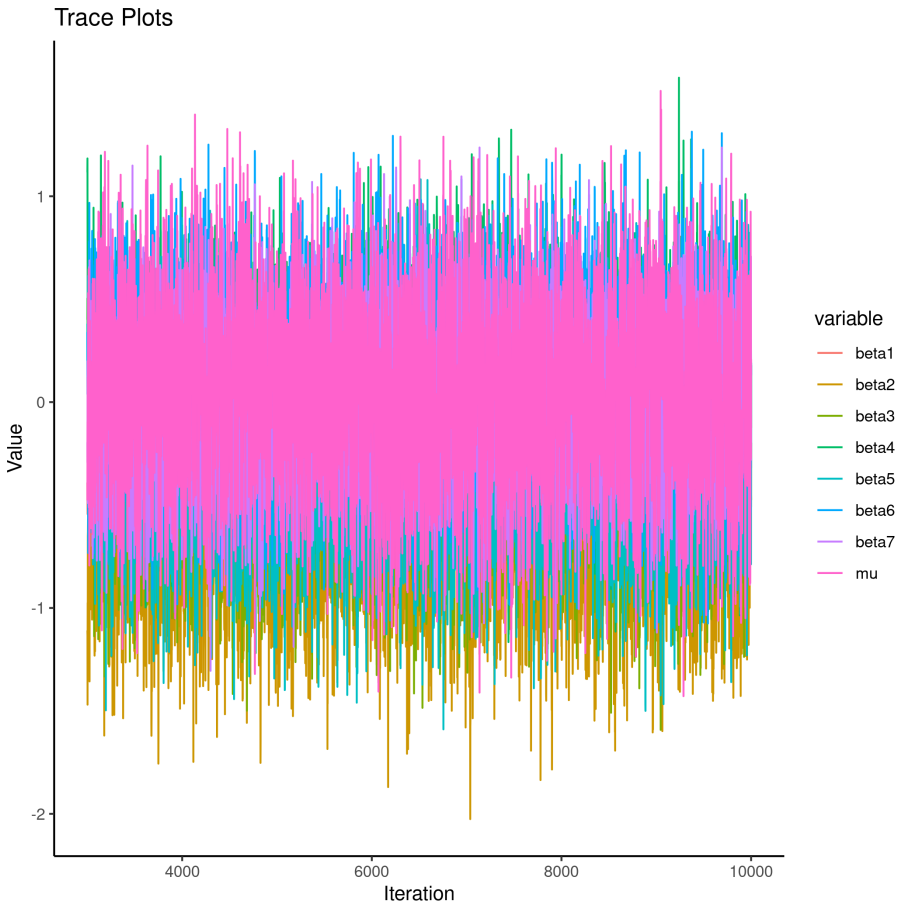

Figure 9: Estimates for New York.

A. Suppose we want to estimate the value of the regression function y^* at some new point x^*, denoted $\hat{f}(x^*)$. Assume for the moment that $f(x)$ is linear, and that y and x have already had their means subtracted, in which case $y_i = \beta x_i + \epsilon_i$.

Return to your least-squares estimator for multiple regression. Show that for the one-predictor case, your prediction $\hat{y}^* = f(x^*) = \hat{\beta}x^*$ may be expressed as a *linear smoother* of the following form:

$$\hat{f}(x^*) = \sum_{i=1}^{n} w(x_i, x^*) y_i$$

for any x^*. Inspect the weighting function you derived. Briefly describe your understanding of how the resulting smoother behaves, compared with the smoother that arises from an alternate form of the weight function $w(x_i, x^*)$:

$$w_K(x_i, x^*) = \begin{cases} 1/K, & x_i \text{ one of the } K \text{ closest sample points to } x^*, \\ 0, & \text{otherwise.} \end{cases}$$

This is referred to as *K-nearest-neighbor smoothing*.

Recall the two following things we need to solve this problem:

$$\hat{\beta} = (X^T X)^{-1} X^T Y,$$
$$X^T Y = \sum_{i=1}^{n} x_i^T y_i,$$

where the first is a known result from multiple linear regression and the second is the expression of a product of matrices as the sum of outer products. With these two, we obtain the following:

$$\begin{aligned} \hat{\beta}x^* &= (X^T X)^{-1} X^T Y x^* \\ &= \left(\sum_{i=1}^{n} x_i^T x_i \right)^{-1} \sum_{i=1}^{n} x_i^T y_i x^* \\ &= \sum_{i=1}^{n} \frac{x_i^T x^*}{\sum_{i=1}^{n} x_i^T x_i} y_i \end{aligned}$$

From the above, we see that

$$w(x_i, x^*) = \frac{x_i^T x^*}{\sum_{i=1}^{n} x_i^T x_i},$$

which smooths the new x^* by scaling it with the ratio $\frac{x_i^T}{x_i^T x_i}$; this effectively takes the proximity of x^* to each x_i into account when summing over all x_i. In comparison, the K-nearest-neighbor smoothing simply scales x^* uniformly across all x_i that are "close" to x_i. We can do this by defining a neighborhood of x_i near x^*, then applying the uniform scale depending on how many x_i belong to this neighborhood.

B. A *kernel function* $K(x)$ is a smooth function satisfying

$$\int_{\mathbb{R}} K(x)dx = 1\,, \quad \int_{\mathbb{R}} xK(x)dx = 0\,, \quad \int_{\mathbb{R}} x^2K(x)dx > 0\,.$$

A very simple example is the uniform kernel,

$$K(x) = \frac{1}{2}I(x) \quad \textbf{where} \quad I(x) = \begin{cases} 1, & |x| \leq 1 \\ 0, & \textbf{otherwise}\,. \end{cases}$$

Another common example is the Gaussian kernel:

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,.$$

Kernels are used as weighting functions for taking local averages. Specifically, define the weighting function

$$w(x_i, x^*) = \frac{1}{h} K\left(\frac{x_i - x^*}{h}\right)\,,$$

where h is the bandwidth. Using this weighting function in a linear smoother is called *kernel regression*. (The weighting function gives the unnormalized weights; you should normalize the weights so that they sum to 1.)

Write your own R function that will fit a kernel smoother for an arbitrary set of x-y pairs and arbitrary choice of (positive real) bandwidth h. You choose the kernel. Set up an R script that will simulate noisy data from some nonlinear function, $y = f(x) + \epsilon$; subtract the sample means from the simulated x and y; and use your function to fit the kernel smoother for some choice of h. Plot the estimated functions for a range of bandwidths wide enough to yield noticeable differences in the qualitative behavior of the prediction functions.

I chose to work with the Gaussian kernel since I have read papers that mention it. For the bandwidth, I used $h = 1, 2, 5, 10$ to yield noticeable differences in the

behavior of the estimated function. Thus, with these bandwidths, the Gaussian kernel, $f_1(x) = 4x^3 + x^2 - 12$, and $x^* \in [-10, 10]$, I obtain Figure 1. Notice that as h increases, our estimated function grows increasingly linear, which is the behavior we would expect to see from our linear smoothing. This behavior can also be seen in Figure 2, where I change the original function to $f_2(x) = x\sin(x)$.

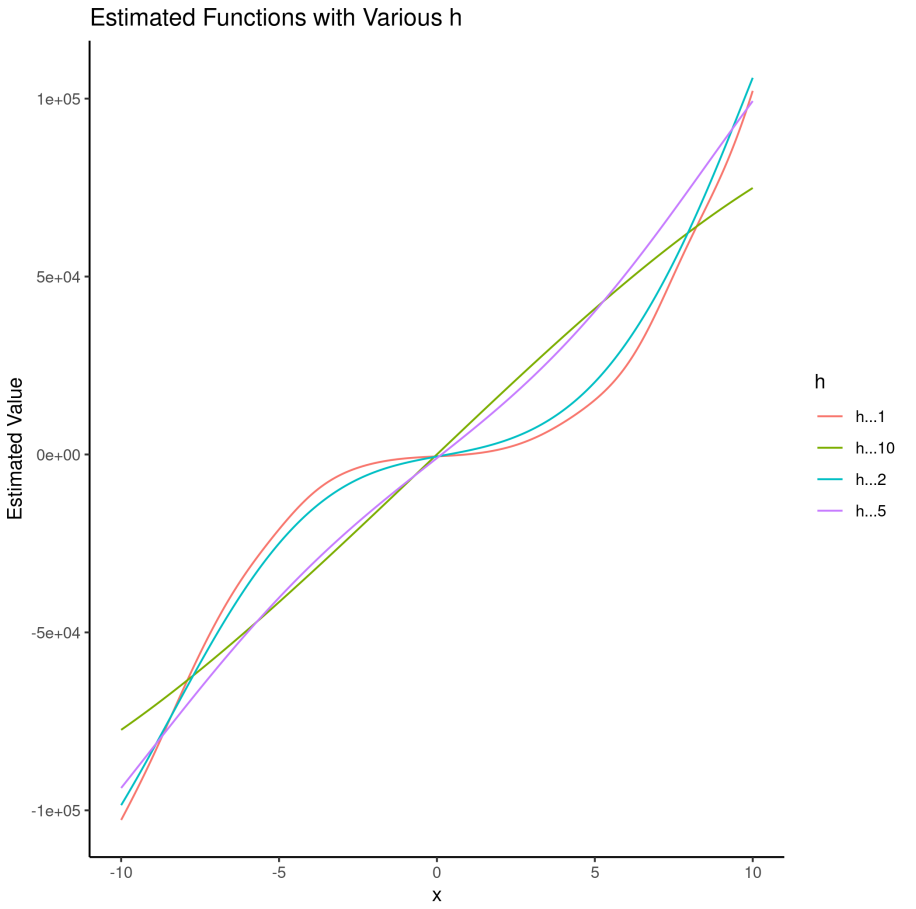

Figure 1: Estimated Functions Under Gaussian Kernel with Various Bandwidths and $f_1(x)$.

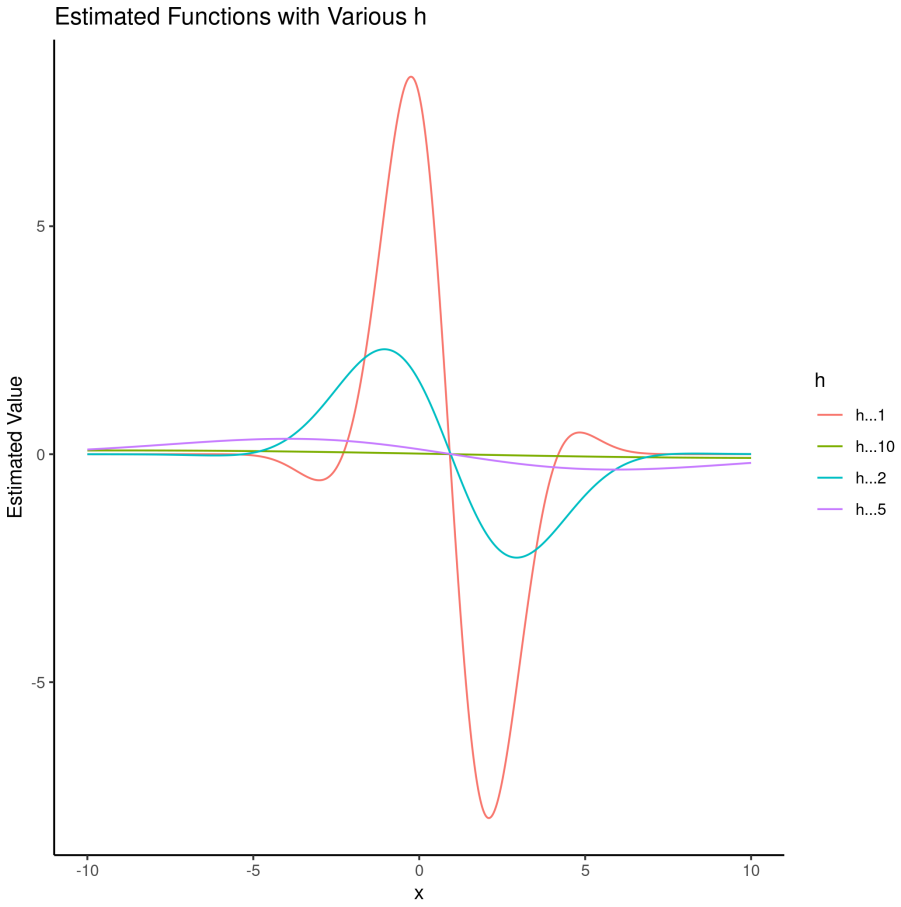

Figure 2: Estimated Functions Under Gaussian Kernel with Various Bandwidths and $f_2(x)$.

A. Presumably a good choice of h would be one that led to smaller predictive errors on fresh data. Write a function or script that will: (1) accept an old ("training") data set and a new ("testing") data set as inputs; (2) fit the kernel-regression estimator to the training data for specified choices of h; and (3) return the estimated functions and the realized prediction error on the testing data for each value of h. This should involve a fairly straightforward "wrapper" of the function you've already written.

Using $f_2(x)$ and $h = 1, 2, 5, 10$ from the previous section, I obtain Figure 3, where the black dots are data in the testing data set. Visually, we can see that the testing data best matches with the kernel-regression estimator when $h = 2$. This conclusion is *validated* when looking at the mean square prediction errors (MSPE) for $h = 1, 2, 5, 10$ which were $MSPE_h = 2871.7169, 504.5653, 777.8502,$ and 791.7469, respectively. Since we want to choose the estimated function with the lowest MSPE, our choice for bandwidth is surely $h = 2$.

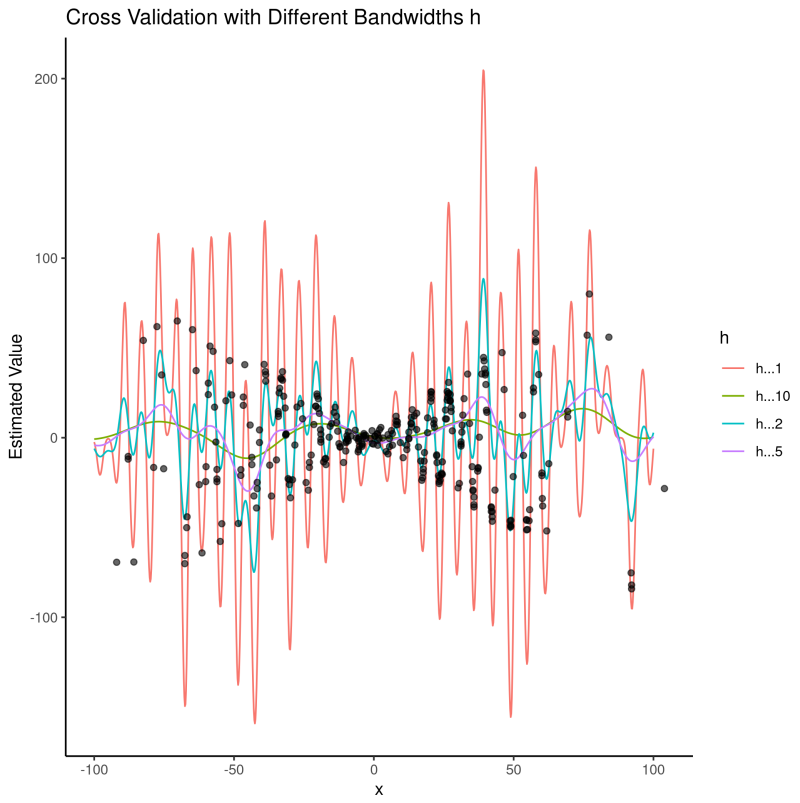

Figure 3: Cross Validation with Different Bandwidths h.

B. Imagine a conceptual two-by-two table for the unknown, true state of affairs. The rows of the table are "wiggly function" and "smooth function," and the columns are "highly noisy observations" and "not so noisy observations." Simulate one data set (say, 500 points) for each of the four cells of this table, where the x's take values in the unit interval. Then split each data set into training and testing subsets. You choose the functions. Apply your method to each case, using the testing data to select a bandwidth parameter. Choose the estimate that minimizes the average squared error in prediction, which estimates the mean-squared error:

$$L_n(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n^*}(y_i^* - \hat{y}_i^*)^2 \,,$$

where (y_i^*, x_i^*) are the points in the test set, and $\hat{y}_i^*$ is your predicted value arising from the model you fit using only the training data. Does your out-of-sample predictive validation method lead to reasonable choices of h for each case?

I used the following functions for this problem:

	High Noise	**Low Noise**
Wiggly Function	$f(x) = \cos(10x) + N(0, 0.8)$	$f(x) = \cos(10x) + N(0, 0.25)$
Smooth Function	$f(x) = x^{5/3} + N(0, 0.8)$	$f(x) = x^{5/3} + N(0, 0.25)$

Once again, I used the Gaussian kernel with bandwidths $h = 0.1, 0.2, 0.5, 1$. I split my 500 observations into 375 for training and 125 for testing. Note that, in the unit interval, the "high noise" significantly scales the observations, which is evident in the left column of Figure 4.

Under these four scenarios, the optimal h for the "wiggly" function and the "smooth" function under low noise was the lowest bandwidth $h = 0.1$. This makes sense since the wiggly function is best "matched" by a linear smoother that does the least amount of smoothing. Visually, we can see in the top row of Figure 4 that the less we smooth, the closer the testing data is to our estimated function. Meanwhile, when starting with a relatively smoother function such as in the bottom row, our optimal smoother may come in the form of a small value for h. In fact, for the "smooth" function with high noise, we find that $h = 0.2$ is the optimal bandwidth under seed *702*. Therefore, it seems that our out-of-sample predictive validation does lead to reasonable choices of h for each case since we have a decent amount of data.

	Wiggly/High	Wiggly/Low	Smooth/High	Smooth/Low
h=0.1	**0.638**	**0.139**	0.616	**0.066**
h=0.2	0.861	0.400	**0.614**	0.074
h=0.5	0.975	0.532	0.641	0.114
h=1	0.979	0.536	0.665	0.139

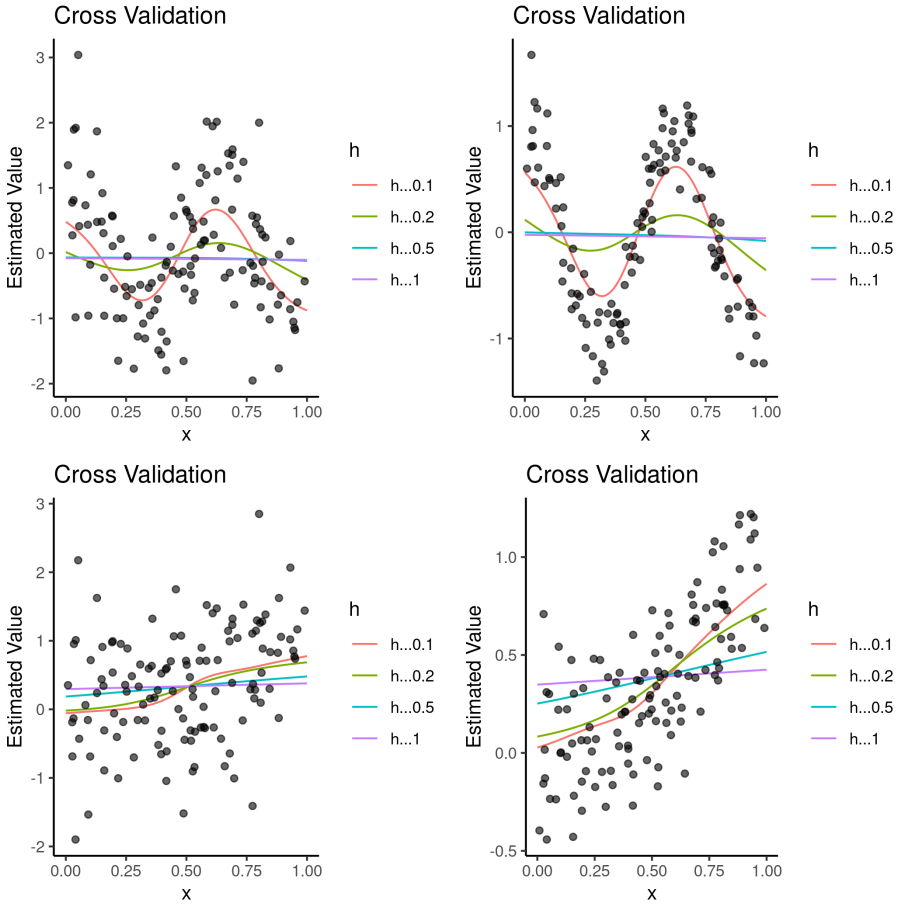

Figure 4: Comparing Optimal Bandwidths Across Different Scenarios in Out-of-Sample Cross Validation.

C. Use the leave-one-out lemma to revisit the examples you simulated in Part B, using leave-one-out cross validation to select h in each case. Because of the leave-one-out lemma, you won't need to actually refit the model N times!

Under linear smoothers, computation of LOOCV simplifies greatly to

$$\text{LOOCV} = \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - H_{ii}} \right)^2$$

This is convenient, since we can easily obtain some metric to determine the optimal bandwidth without having to split our data set into two. Note that with the out-of-sample approach, we are only able to use 75% of the data to fit an estimated function, whereas with this method, we are able to use all of the data!

In Figure 5, we can see the same smoothed estimated functions from Figure 4 with all of the data points overlaid. Note that the estimated functions may differ slightly in shape because we were able to use all of the data to fit them. Using the LOOCV criterion, the optimal bandwidth for both the "wiggly" and "smooth" function under both high and low noise was $h = 0.1$, indicating that when we use all of the data to obtain our estimated function in a variety of conditions, it's most optimal not to smooth the data given a sufficient amount of data.

	Wiggly/High	**Wiggly/Low**	**Smooth/High**	**Smooth/Low**
h=0.1	**0.684**	**0.150**	**0.594**	**0.068**
h=0.2	0.988	0.426	0.601	0.074
h=0.5	1.139	0.564	0.644	0.114
h=1	1.143	0.569	0.672	0.139

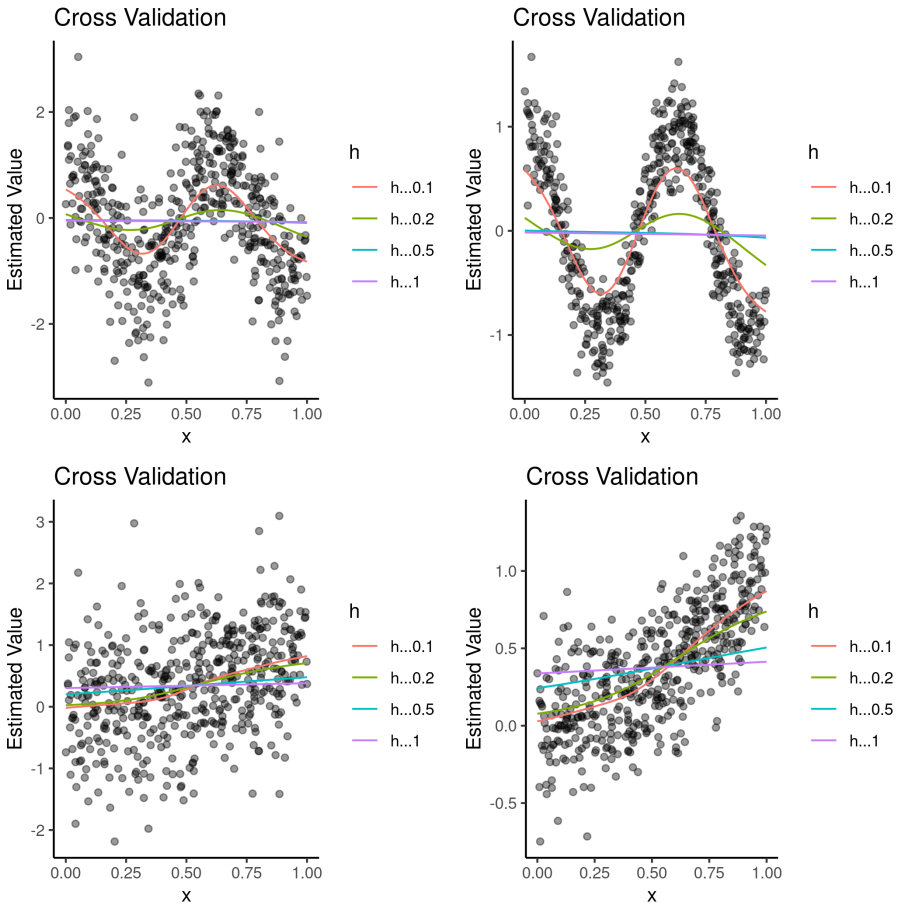

Figure 5: Comparing Optimal Bandwidths Across Different Scenarios in Leave-One-Out Cross Validation.

A. A natural generalization of locally constant regression is local polynomial regression. For points u in a neighborhood of the target point x, define the polynomial

$$g_x(u;a) = a_0 + \sum_{k=1}^{D} a_k(u-x)^k$$

for some vector of coefficients $a = (a_0, \ldots, a_D)$. As above, we will estimate the coefficients a in $g_x(u;a)$ at some target point x using weighted least squares:

$$\hat{a} = \arg\min_{R^{D+1}} \sum_{i=1}^{n} w_i \left\{ y_i - g_x(x_i;a) \right\}^2 ,$$

where $w_i \equiv w(x_i, x)$ are the kernel weights defined just above, normalized to sum to one. Derive a concise (matrix) form of the weight vector $\hat{a}$, and by extension, the local function estimate $\hat{f}(x)$ at the target value x. Life will be easier if you define the matrix R_x whose (i,j) entry is $(x_i - x)^{j-1}$, and remember that (weighted) polynomial regression is the same thing as (weighted) linear regression with a polynomial basis.

First, let's express $\sum_{i=1}^{n} w_i \left\{ y_i - g_x(x_i;a) \right\}^2$ in matrix form. This is not such an arduous task once we realize that the summation contains a quadratic. However, before we get to that, let's use the given hint:

$$\begin{aligned} R_x\mathbf{a} &= \begin{bmatrix} a_0 + a_1(x_1 - x) + \ldots + a_D(x_1 - x)^D \\ \vdots \\ a_0 + a_1(x_n - x) + \ldots + a_D(x_n - x)^D \end{bmatrix} \\ &= \begin{bmatrix} g_x(x_1|\mathbf{a}) \\ \vdots \\ g_x(x_n|\mathbf{a}) \end{bmatrix} \end{aligned}$$

Now we can write the summation in matrix notation:

$$\sum_{i=1}^{n} w_i \left\{ y_i - g_x(x_i;a) \right\}^2 = (\mathbf{y} - R_x\mathbf{a})^T \mathrm{diag}(\tilde{\mathbf{w}})(\mathbf{y} - R_x\mathbf{a})$$

Now, let's derive the matrix expression with respect to $\mathbf{a}$ to obtain $\hat{\mathbf{a}}$:

$$\begin{aligned} \frac{\partial}{\partial \mathbf{a}} \left[(\mathbf{y} - R_x\mathbf{a})^T \mathrm{diag}(\tilde{\mathbf{w}})(\mathbf{y} - R_x\mathbf{a}) \right] &= -\left(\mathbf{y}^T \mathrm{diag}(\tilde{\mathbf{w}}) R_x \right)^T - R_x^T \mathrm{diag}(\tilde{\mathbf{w}})\mathbf{y} + 2\left(R_x^T \mathrm{diag}(\tilde{\mathbf{w}}) R_x \right)\mathbf{a} \\ &= -2R_x^T \mathrm{diag}(\tilde{\mathbf{w}})\mathbf{y} + 2\left(R_x^T \mathrm{diag}(\tilde{\mathbf{w}}) R_x \right)\mathbf{a} \stackrel{set}{=} 0 \\ R_x^T \mathrm{diag}(\tilde{\mathbf{w}}) R_x \mathbf{a} &= R_x^T \mathrm{diag}(\tilde{\mathbf{w}})\mathbf{y} \\ \hat{\mathbf{a}} &= \left(R_x^T \mathrm{diag}(\tilde{\mathbf{w}}) R_x \right)^{-1} R_x^T \mathrm{diag}(\tilde{\mathbf{w}})\mathbf{y} \end{aligned}$$

By extension, $\hat{f}(x) = \mathbf{e}^T\hat{\mathbf{a}}$, where $e^T = [1\ \mathbf{0}]$.

B. From this, conclude that for the special case of the local linear estimator ($D = 1$), we can write $\hat{f}(x)$ as a linear smoother of the form

$$\hat{f}(x) = \frac{\sum_{i=1}^n w_i(x) y_i}{\sum_{i=1}^n w_i(x)},$$

where the unnormalized weights are

$$\begin{aligned} w_i(x) &= K\left(\frac{x - x_i}{h}\right)\{s_2(x) - (x_i - x)s_1(x)\} \\ s_j(x) &= \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)(x_i - x)^j. \end{aligned}$$

With $D = 1$, note that $g_x(x_i|a) = a_0 + a_1(x_i - x)$. Further, we can write

$$\begin{aligned} \hat{f}(x) &= \mathbf{e}^T\hat{\mathbf{a}} \\ &= \begin{bmatrix} 1 & 0 \end{bmatrix}(R_x^T \text{diag}(\tilde{\mathbf{w}})R_x)^{-1}R_x^T\text{diag}(\tilde{\mathbf{w}})\mathbf{y} \end{aligned}$$

With hopes to keep this computation as organized as possible, let's take a modular approach and obtain the following:

$$\begin{aligned} R_x^T\text{diag}(\tilde{\mathbf{w}}) &= \begin{bmatrix} 1 & \dots & 1 \\ x_1 - x & \dots & x_n - x \end{bmatrix}\begin{bmatrix} \tilde{w}_1 & \dots & 0 \\ \vdots & \tilde{w}_i & \vdots \\ 0 & \dots & \tilde{w}_n \end{bmatrix} \\ &= \begin{bmatrix} \tilde{w}_1 & \dots & \tilde{w}_n \\ \tilde{w}_1(x_1 - x) & \dots & \tilde{w}_n(x_n - x) \end{bmatrix}, \\ R_x^T\text{diag}(\tilde{\mathbf{w}})R_x &= \begin{bmatrix} \tilde{w}_1 & \dots & \tilde{w}_n \\ \tilde{w}_1(x_1 - x) & \dots & \tilde{w}_n(x_n - x) \end{bmatrix}\begin{bmatrix} 1 & x_1 - x \\ \vdots & \vdots \\ 1 & x_n - x \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^n \tilde{w}_i & \sum_{i=1}^n \tilde{w}_i(x_i - x) \\ \sum_{i=1}^n \tilde{w}_i(x_i - x) & \sum_{i=1}^n \tilde{w}_i(x_i - x)^2 \end{bmatrix}, \\ \left(R_x^T\text{diag}(\tilde{\mathbf{w}})R_x\right)^{-1} &= \frac{1}{\mathcal{D}}\begin{bmatrix} \sum_{i=1}^n \tilde{w}_i(x_i - x)^2 & -\sum_{i=1}^n \tilde{w}_i(x_i - x) \\ -\sum_{i=1}^n \tilde{w}_i(x_i - x) & \sum_{i=1}^n \tilde{w}_i \end{bmatrix}, \end{aligned}$$

where $\mathcal{D}$ is the determinant of the matrix:

$$\begin{aligned}\mathcal{D} &= \sum_{i=1}^{n} \tilde{w}_i (x_i - x)^2 - \left(\sum_{i=1}^{n} \tilde{w}_i (x_i - x) \right)^2 \\ &= \sum_{i=1}^{n} K(\cdot)(x_i - x)^2 - \left(\sum_{i=1}^{n} K(\cdot)(x_i - x) \right)^2 \\ &= s_2(x) - s_1^2(x),\end{aligned}$$

where I'm getting lazy so I'm using shorthand notation like $S_k^{-1} = 1/\sum_{i=1}^{n} K(\cdot)$ and $K(\cdot) = K((x_i - x)/h)$. Now, let's continue with the matrix multiplication:

$$\begin{aligned}&\begin{bmatrix}1 & 0\end{bmatrix} \frac{1}{\mathcal{D}} \begin{bmatrix} \sum_{i=1}^{n} \tilde{w}_i (x_i - x)^2 & -\sum_{i=1}^{n} \tilde{w}_i (x_i - x) \\ -\sum_{i=1}^{n} \tilde{w}_i (x_i - x) & \sum_{i=1}^{n} \tilde{w}_i \end{bmatrix} \\ &\quad = \begin{bmatrix} \frac{\sum_{i=1}^{n} K(\cdot)(x_i - x)^2}{\mathcal{D}} & \frac{-\sum_{i=1}^{n} K(\cdot)(x_i - x)}{\mathcal{D}} \end{bmatrix}, \\ \hat{f}(x) &= \begin{bmatrix} \frac{\sum_{i=1}^{n} K(\cdot)(x_i - x)^2}{\mathcal{D}} & \frac{-\sum_{i=1}^{n} K(\cdot)(x_i - x)}{\mathcal{D}} \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{n} K(\cdot) y_i \\ \sum_{i=1}^{n} K(\cdot)(x_i - x) y_i \end{bmatrix} \\ &\quad = \frac{s_2(x) \sum_{i=1}^{n} K(\cdot) y_i - s_1(x) \sum_{i=1}^{n} K(\cdot)(x_i - x) y_i}{s_2(x) - s_1^2(x)} \\ &\quad = \frac{\sum_{i=1}^{n} K(\cdot) \left[s_2(x) - (x_i - x) s_1(x) \right] y_i}{s_2(x) \sum_{i=1}^{n} K(\cdot) - s_1(x) \sum_{i=1}^{n} K(\cdot)(x_i - x)} \\ &\quad = \frac{\sum_{i=1}^{n} w_i(x) y_i}{\sum_{i=1}^{n} w_i(x)}\end{aligned}$$

C. Suppose that the residuals have constant variance σ^2 (that is, the spread of the residuals does not depend on x). Derive the mean and variance of the sampling distribution for the local polynomial estimate $\hat{f}(x)$ at some arbitrary point x. Note: the random variable $\hat{f}(x)$ is just a scalar quantity at x, not the whole function.

Note that we can write our local polynomial regression as

$$\begin{aligned}\hat{f}(x_i) &= e^T \hat{a} \\ &= e^T (X^T W X)^{-1} X^T W Y\end{aligned}$$

Now, we can take the expectation and variance of the above. Generally,

$$\begin{aligned}E[\hat{f}(x)] &= e^T (X^T W X)^{-1} X^T W f(x), \\ \text{var}[\hat{f}(x)] &= \sigma^2 e^T (X^T W X)^{-1} X^T W W^T X (X^T W X)^{-1} e\end{aligned}$$

In the case for (B) (i.e., $\hat{f}(x) = \tilde{w}^T y$), we can simplify the above expression and obtain:

$$E[\hat{f}(x)] = \tilde{w}^T f(x),$$
$$\text{var}[\hat{f}(x)] = \sigma^2 \tilde{w}^T \tilde{w},$$

where $\tilde{w} = \left[\frac{w_1(x)}{\sum_i w_i(x)}, ..., \frac{w_n(x)}{\sum_i w_i(x)}\right]^T$ comes from (B). Note that more smoothing implies more bias but less variance since $\tilde{w}^T \tilde{w} \leq 1$. This is another example of the bias-variance trade-off.

D. We don't know the residual variance, but we can estimate it. A basic fact is that if x is a random vector with mean μ and covariance matrix Σ, then for any symmetric matrix Q of appropriate dimension, the quadratic form $x^T Q x$ has expectation

$$E(x^T Q x) = \mathbf{tr}(Q\Sigma) + \mu^T Q \mu \,.$$

Write the vector of residuals as $r = y - \hat{y} = y - Hy$, where H is the smoothing matrix. Compute the expected value of the estimator

$$\hat{\sigma}^2 = \frac{\|r\|_2^2}{n - 2\mathbf{tr}(H) + \mathbf{tr}(H^T H)} \,,$$

and simplify things as much as possible. Roughly under what circumstances will this estimator be nearly unbiased for large n? Note: the quantity $2\mathbf{tr}(H) - \mathbf{tr}(H^T H)$ is often referred to as the "effective degrees of freedom" in such problems.

We can write $\hat{\sigma}^2$ as

$$\hat{\sigma}^2 = \frac{(y - Hy)^T (y - Hy)}{n - 2tr(H) + tr(H^T H)}$$

The denominator can be rewritten as

$$n - 2tr(H) + tr(H^T H) = tr\left[(I - H)^T (I - H)\right]$$

If we take the expectation of the numerator, we obtain

$$\begin{aligned} E\left[||r||_2^2\right] &= tr\left((I - H)^T (I - H)\sigma^2\right) + \mu^T (I - H)^T (I - H)\mu \\ &= tr\left((I - H)^T (I - H)\sigma^2\right) + ||f(x) - Hf(x)||_2^2, \end{aligned}$$

which is unbiased for σ^2 only when the second term is 0; this would only happen if the squared Euclidean norm of the bias was 0, implying bias is 0. Since this is unlikely, it's unlikely that $\hat{\sigma}^2$ is unbiased for σ^2.

E. Write code that fits the local linear estimator using a Gaussian kernel for a specified choice of bandwidth h. Then load the data in "utilities.csv" into R. This data set shows the monthly gas bill (in dollars) for a single-family home in Minnesota, along with the average temperature in that month (in degrees F), and the number of billing days in that month. Let y be the average daily gas bill in a given month (i.e. dollars divided by billing days), and let x be the average temperature. Fit y versus x using local linear regression and some choice of kernel. Choose a bandwidth by leave-one-out cross-validation.

Before fitting the local linear regression model, I tested 100 bandwidths between 0 and 15 to find the optimal bandwidth, which was $h = 6.873$; this was done using leave-one-out validation under a Gaussian kernel and the specified weighting scheme. The local linear estimates under these various bandwidths can be seen in Figure 6. Then, I obtained my estimates $\hat{\mathbf{y}}$ under this optimal bandwidth, which are reported in Figure 7, where the red dots are the actual observations and the black line are fitted values.

F. Inspect the residuals from the model you just fit. Does the assumption of constant variance (homoskedasticity) look reasonable? If not, do you have any suggestion for fixing it?

The residuals from the fitted model can be seen in Figure 8, which clearly depicts homoskedasticity. I do not believe the assumption of constant variance to be reasonable since variance is larger for smaller values of X. This may be remedied by admitting heterogeneous error in the model specification or with weighted local polynomial regression! The variance appears to be a function of X, so suppose we do the following to fix the heteroskedasticity:

$$\begin{aligned}
\log \text{var}(e_i) &= g(x_i), \\
\text{var}(e_i) &= E\left[(e_i - E(e_i))^2\right] \\
&= E[e_i^2], \\
\log E(e_i^2) &= g(x_i)
\end{aligned}$$

This should fix the heteroskedasticity since we are accounting for the values of X when determining the variance and, consequently, the residuals.

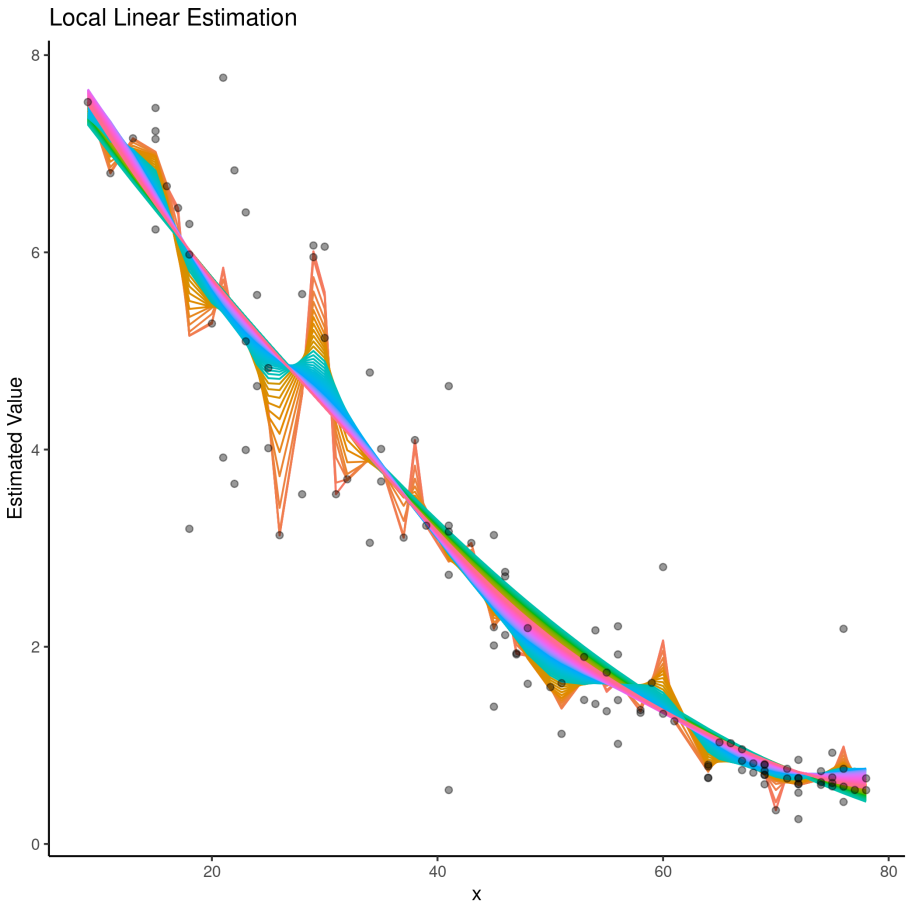

Figure 6: Local Linear Estimation at Various Bandwidths h.

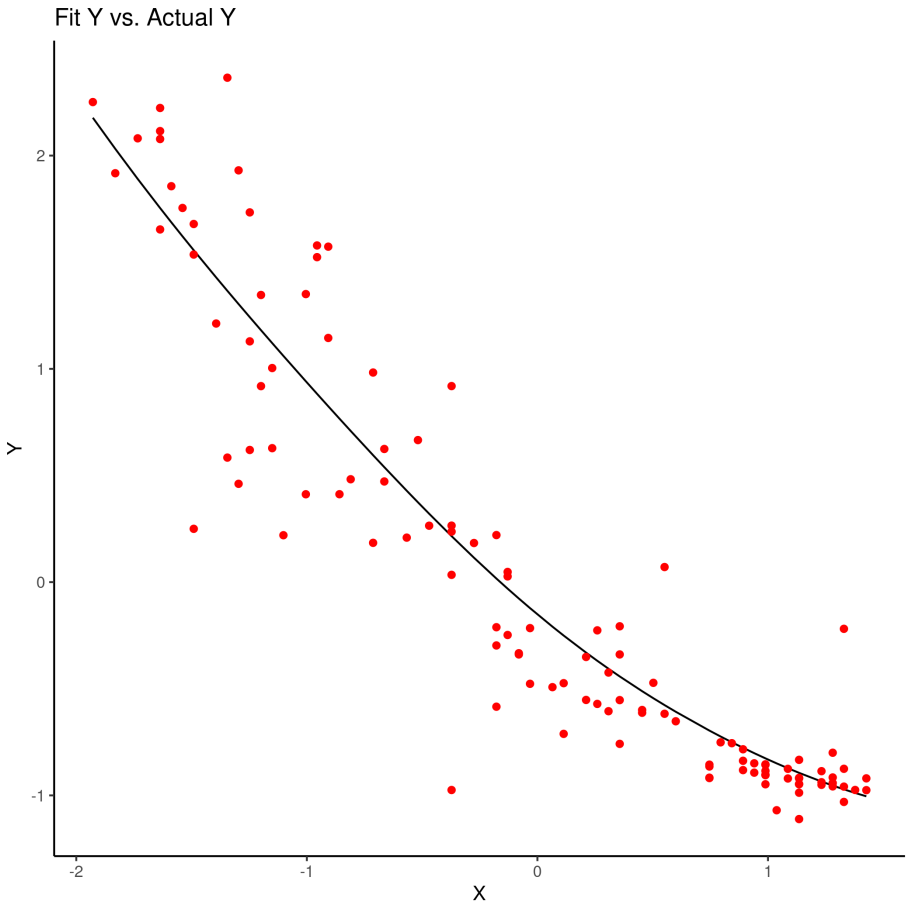

Figure 7: $\hat{y}$ vs. y Under Optimal Bandwidth.

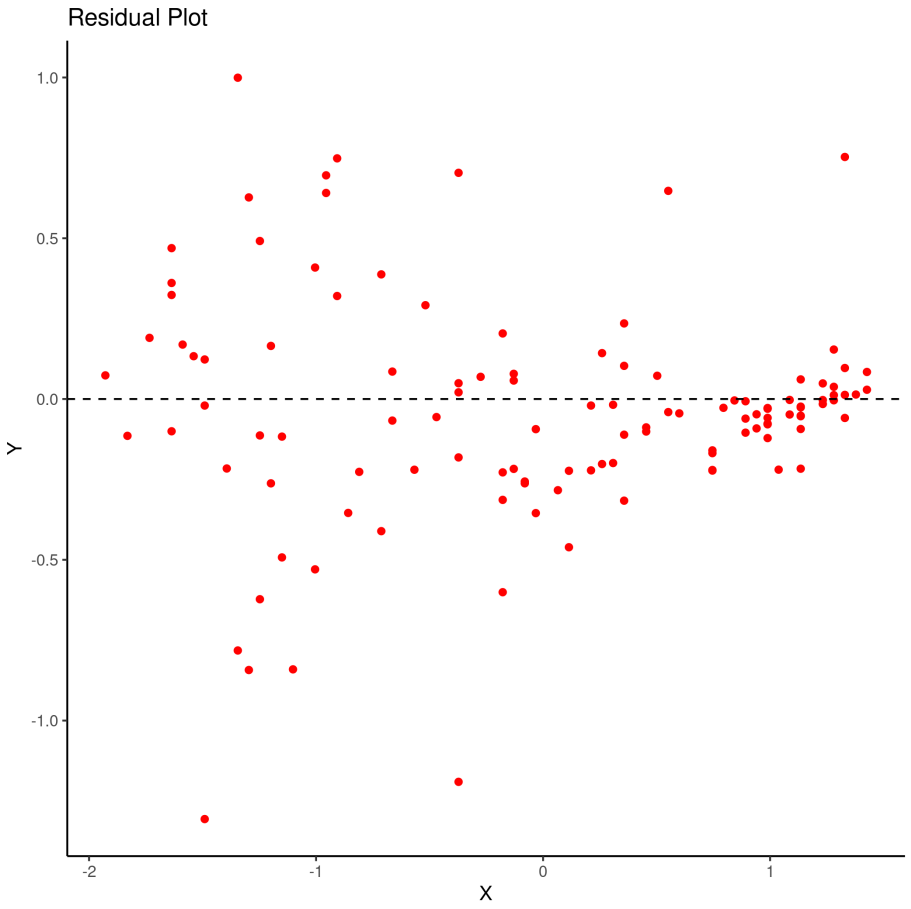

Figure 8: Residuals from Fit Model Under Optimal Bandwidth.

G. Put everything together to construct an approximate point-wise 95% confidence interval for the local linear model (using your chosen bandwidth) for the value of the function at each of the observed points x_i for the utilities data. Plot these confidence bands, along with the estimated function, on top of a scatter plot of the data.

The confidence bands (UT orange) overlaid on the estimated function (black line) and observations (red dots) can be found in Figure 9. It appears that, for the most part, our confidence bands slightly decrease in height as X increases. These bands were obtained from the following:

$$\hat{f}(x) \pm 1.96 \cdot \sqrt{\hat{\sigma}^2 ||h||_2^2},$$

where h is the row of the smoothing matrix and $\hat{\sigma}^2$ is obtained using (D).

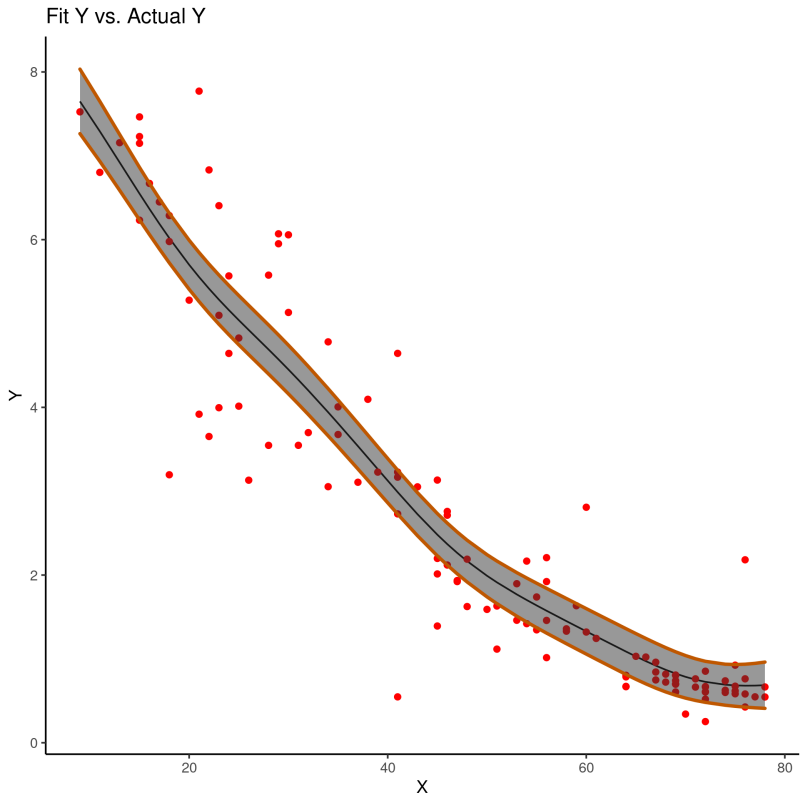

Figure 9: Point-wise 95% Confidence Intervals.

A. Read up on the Matern class of covariance functions. The Matern class has the *squared exponential* covariance function as a special case:

$$C_{SE}(x_1, x_2) = \tau_1^2 \exp\left\{-\frac{1}{2}\left(\frac{d(x_1, x_2)}{b}\right)^2\right\} + \tau_2^2 \delta(x_1, x_2)\,,$$

where $d(x_1, x_2) = \|x_1 - x_2\|_2$ is Euclidean distance (or just $|x - y|$ for scalars). The constants (b, τ_1^2, τ_2^2) are often called *hyperparameters*, and $\delta(a, b)$ is the Kronecker delta function that takes the value 1 if $a = b$, and 0 otherwise. But usually this covariance function generates functions that are "too smooth," and so we use other covariance functions in the Matern class as a default.

Let's start with the simple case where $\mathcal{X} = [0, 1]$, the unit interval. Write a function that simulates a mean-zero Gaussian process on $[0, 1]$ under the squared exponential covariance function. The function will accept as arguments: (1) finite set of points $x_1, \ldots, x_N$ on the unit interval; and (2) a triplet (b, τ_1^2, τ_2^2). It will return the value of the random process at each point: $f(x_1), \ldots, f(x_N)$.

Use your function to simulate (and plot) Gaussian processes across a range of values for b, τ_1^2, and τ_2^2. Try starting with a very small value of τ_2^2 (say, 10^{-6}) and playing around with the other two first. On the basis of your experiments, describe the role of these three hyperparameters in controlling the overall behavior of the random functions that result. What happens when you try $\tau_2^2 = 0$? Why? If you can fix this, do—remember our earlier discussion on different ways to simulate the MVN.

Now simulating a few functions with a different covariance function, the Matérn with parameter $5/2$:

$$C_{M52}(x_1, x_2) = \tau_1^2 \left\{1 + \frac{\sqrt{5}d}{b} + \frac{5d^2}{3b^2}\right\} \exp\left(\frac{-\sqrt{5}d}{b}\right) + \tau_2^2 \delta(x_1, x_2)\,,$$

where $d = \|x_1 - x_2\|_2$ is the distance between the two points x_1 and x_2. Comment on the differences between the functions generated from the two covariance kernels.

First, I test the two given covariance functions (the squared exponential and Matern $\frac{5}{2}$ covariance functions) across a grid of 100 different points in the parameter space for (b, τ_1^2). These points for b and τ_1^2 range from 0.0001 to 1 and 0

to 2, respectively. The resulting simulations for the squared exponential covariance function and the Matern $\frac{5}{2}$ covariance function can be found in Figures 10 and 11, respectively. Note how the squared exponential covariance function generates much smoother simulations at every point in the parameter space for (b, τ_1^2).

In the aforementioned simulations, we fixed $\tau_2^2 = 10^{-6}$. However, once $\tau_2^2 = 0$, we run into identifiability issues, because the nugget in the covariance functions (the Kronecker delta) coincides with the variance imposed by the Gaussian specification. Additionally, when $\tau_2^2 = 0$, our results are often numerically unstable. We can get around this by implementing a singular matrix specification of the normal distribution.

For the squared exponential covariance function, the b hyperparameter acts as our bandwidth; as b increases, our functions become smoother, which is evident in each of the discussed figures. In fact, in the top left subplot for each figure, you can see the lack of smoothness for each value τ_1^2. As τ_1^2 increases, the covariance increasingly scales. Note how when $\tau_1^2 \approx 0$, each subplot in Figure 10 depicts a horizontal line at 0. However, as τ_1^2 increases, the scale (amplitude) increases.

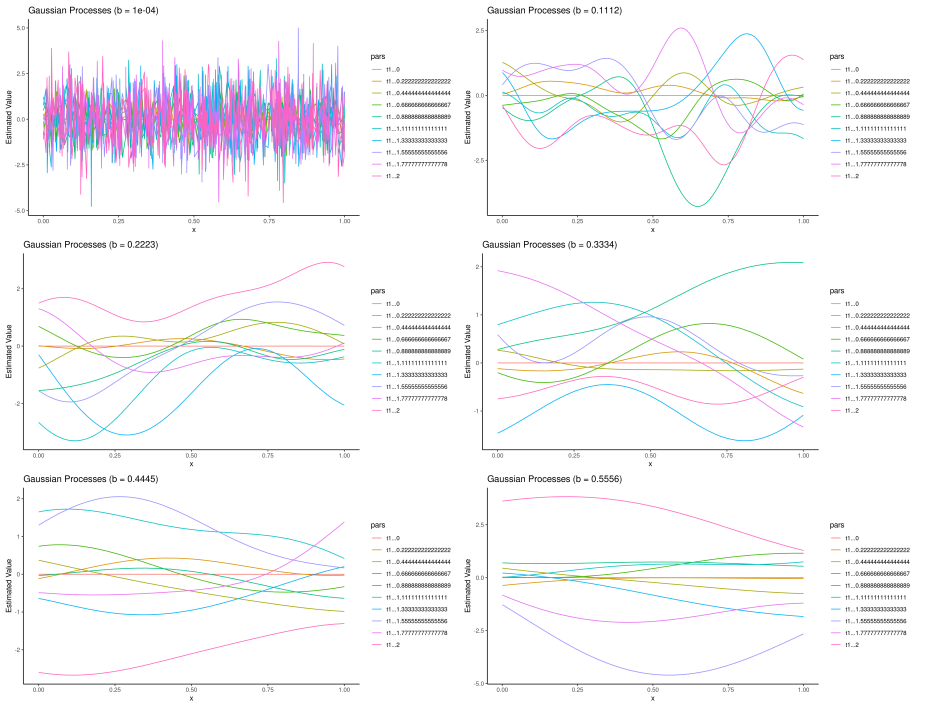

Figure 10: Gaussian Process Simulations with Squared Exponential Covariance Function.

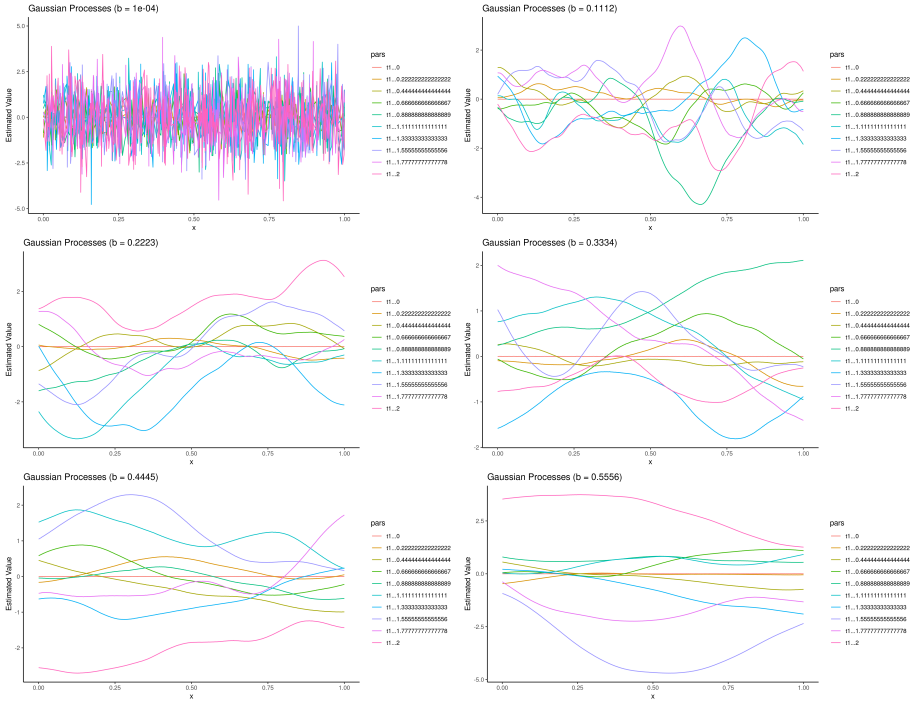

Figure 11: Gaussian Process Simulations with Matern $\frac{5}{2}$ Covariance Function.

B. Suppose you observe the value of a Gaussian process $f \sim \mathbf{GP}(m, C)$ at points $x_1, \ldots, x_N$. What is the conditional distribution of the value of the process at some new point $x^\star$? For the sake of notational ease simply write the value of the (i, j) element of the covariance matrix as $C_{i,j}$, rather than expanding it in terms of a specific covariance function.

Recall that we can obtain the conditional distribution of multivariate Gaussian distributions in the following way. Suppose that

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

Then, the posterior distribution of y_1 conditional on y_2 is given by

$$\begin{aligned} p(y_1|y_2) &\sim \mathcal{N}\left(\mu^*, \Sigma^*\right), \\ \mu^* &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2), \\ \Sigma^* &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \end{aligned}$$

Here, we are given a *new* data point x^*, and we want to predict $f(x^*)$. The corresponding joint distribution of $\mathbf{f} = [f(x_1), ..., f(x_n)]$ and $f(x^*)$ is

$$\begin{bmatrix} f(x^*) \\ \mathbf{f} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(x^*) \\ m(\mathbf{x}) \end{bmatrix}, \begin{bmatrix} C^* & \tilde{C}^T \\ \tilde{C} & C \end{bmatrix} \right),$$

where

$$\begin{aligned} C &= C(\mathbf{x}, \mathbf{x}), \\ \tilde{C} &= C(\mathbf{x}, x^*), \\ C^* &= C(x^*, x^*) \end{aligned}$$

The corresponding conditional distribution of $f(x^*)$ given $\mathbf{f}$ and $\mathbf{x}$ is

$$f(x^*) \Big| \mathbf{f}, \mathbf{x}, x^* \sim \mathcal{N}\left(m(x^*) + \tilde{C}^T C^{-1}(\mathbf{y} - m(\mathbf{x})),\ C^* - \tilde{C}^T C^{-1} \tilde{C} \right)$$

C. Prove the following lemma.

Lemma 1 ***Suppose that the joint distribution of two vectors y and θ has the following properties: (1) the conditional distribution for y given θ is multivariate normal, $(y \mid \theta) \sim N(R\theta, \Sigma)$; and (2) the marginal distribution of θ is multivariate normal, $\theta \sim N(m, V)$. Assume that R, Σ, m, and V are all constants. Then, the joint distribution of y and θ is multivariate normal.***

We know that we can obtain the joint distribution for $(\boldsymbol{\theta}, \mathbf{y})$ with the following:

$$\begin{aligned} p(\boldsymbol{\theta}, \mathbf{y}) &= p(\mathbf{y}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}) \\ &\propto \exp\left\{ -\frac{1}{2} \left[\underbrace{(\mathbf{y} - R\boldsymbol{\theta})^T \Sigma^{-1} (\mathbf{y} - R\boldsymbol{\theta}) + (\boldsymbol{\theta} - \mathbf{m})^T V^{-1} (\boldsymbol{\theta} - \mathbf{m})}_{A} \right] \right\} \end{aligned}$$

Focusing on the A term, we can expand it to obtain the following:

$$A \propto \mathbf{y}^T \Sigma^{-1} \mathbf{y} - 2\mathbf{y}^T \Sigma^{-1} R\boldsymbol{\theta} + \boldsymbol{\theta}^T R^T \Sigma^{-1} R\boldsymbol{\theta} + \boldsymbol{\theta}^T V^{-1} \boldsymbol{\theta} - 2m^T V^{-1} \boldsymbol{\theta}$$

From here, we can complete the square to rewrite A in the following matrix form:

$$A \propto \begin{bmatrix} \boldsymbol{\theta} - \mathbf{m} \\ \mathbf{y} - R\mathbf{m} \end{bmatrix}^T \begin{bmatrix} V^{-1} + R^T \Sigma^{-1} R & -R^T \Sigma^{-1} \\ -\Sigma^{-1} R & \Sigma^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta} - \mathbf{m} \\ \mathbf{y} - R\mathbf{m} \end{bmatrix}$$

We can multiply this out to verify the matrix form:

$$\begin{aligned} A &\propto (\theta - m)^T (V^{-1} + R^T \Sigma^{-1} R)(\theta - m) - (y - Rm)^T \Sigma^{-1} R(\theta - m) \\ &\quad - (\theta - m)^T R^T \Sigma^{-1} (y - Rm) + (y - Rm)^T \Sigma^{-1} (y - Rm) \\ &\propto \theta^T (V^{-1} + R^T \Sigma^{-1} R)\theta - 2m^T (V^{-1} + R^T \Sigma^{-1} R)\theta \\ &\quad - y^T \Sigma^{-1} R\theta + y^T \Sigma^{-1} Rm + m^T R^T \Sigma^{-1} R\theta \\ &\quad - \theta^T R^T \Sigma^{-1} y + \theta^T R^T \Sigma^{-1} Rm + m^T R^T \Sigma^{-1} y \\ &\quad + y^T \Sigma^{-1} y^T - 2m^T R^T \Sigma^{-1} y \\ &= \theta^T V^{-1} \theta + \theta^T R^T \Sigma^{-1} R\theta - 2m^T V^{-1} \theta - 2m^T R^T \Sigma^{-1} R\theta \\ &\quad - y^T \Sigma^{-1} R\theta + y^T \Sigma^{-1} Rm + m^T R^T \Sigma^{-1} R\theta \\ &\quad - \theta^T R^T \Sigma^{-1} y + \theta^T R^T \Sigma^{-1} Rm + m^T R^T \Sigma^{-1} y \\ &\quad + y^T \Sigma^{-1} y^T - 2m^T R^T \Sigma^{-1} y \end{aligned}$$

Then, the above green and red colored expressions cancel to obtain the original form of A. The blue expressions combine together nicely, as well. While we

found what A is proportional to, we can neglect all those additional summed terms that don't contain $\mathbf{y}$ or $\boldsymbol{\theta}$ since A is in an exponent. Thus, we have that

$$p(\boldsymbol{\theta}, \mathbf{y}) \propto \exp\left\{-\frac{1}{2}A\right\},$$

which indicates that $p(\boldsymbol{\theta}, \mathbf{y})$ is a multivariate normal distribution with

$$\begin{aligned} \text{mean} &= \begin{bmatrix} \mathbf{m} \\ R\mathbf{m} \end{bmatrix}, \\ \text{precision matrix} &= \begin{bmatrix} V^{-1} + R^T\Sigma^{-1}R & -R^T\Sigma^{-1} \\ -\Sigma^{-1}R & \Sigma^{-1} \end{bmatrix} \end{aligned}$$

By the given marginal distribution and conditional distribution, we knew the mean vector would take the above form. The trickier part is knowing the precision matrix. However, once we identify the quadratic terms in the original expression for A, we know the main diagonal elements. The only thing left to find is the off-diagonal elements in the precision matrix. We know that they are tranposes of each other. Additionally, with a keen eye, we may notice that the original expression for A contains the term $-2\mathbf{y}^T\Sigma^{-1}R\boldsymbol{\theta}$. This indicates that we need to have a Σ^{-1} and an R^T in the off-diagonal terms because we otherwise wouldn't be able to get a term with a single $\boldsymbol{\theta}$ and a single R^T. Completing the term specifications, we can note that they must be negative to cancel with the existing terms.

A. Suppose we observe data $y_i = f(x_i) + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, for some unknown function f. Suppose that the prior distribution for the unknown function is a mean-zero Gaussian process: $f \sim \mathbf{GP}(0, C)$ for some covariance function C. Let $x_1, \ldots, x_N$ denote the previously observed x points. Derive the posterior distribution for the random vector $[f(x_1), \ldots, f(x_N)]^T$, given the corresponding outcomes $y_1, \ldots, y_N$, assuming that you know σ^2.

We can denote

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{f}, \sigma^2 I\right)$$

and obtain the full-conditional distribution for $\mathbf{f}$ just as we have in the past:

$$\begin{aligned}
[\mathbf{f}|\cdot] &\propto [\mathbf{y}|\mathbf{f}, \sigma^2] \cdot [\mathbf{f}] \\
&\propto \exp\left\{-\frac{1}{2}\left[(\mathbf{y}-\mathbf{f})^T(\sigma^2 I)^{-1}(\mathbf{y}-\mathbf{f}) + \mathbf{f}^T C^{-1}\mathbf{f}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[-2\underbrace{\mathbf{y}^T(\sigma^2 I)^{-1}}_{b^T}\mathbf{f} + \mathbf{f}^T\left(\underbrace{(\sigma^2 I)^{-1} + C^{-1}}_{A}\right)\mathbf{f}\right]\right\}
\end{aligned}$$

Therefore, we can see that

$$[\mathbf{f}|\cdot] \equiv \mathcal{N}\left((I + \sigma^2 C^{-1})^{-1}\mathbf{y}, (\sigma^{-2} I + C^{-1})^{-1}\right)$$

B. As before, suppose we observe data $y_i = f(x_i) + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, for $i = 1, \ldots, N$. Now we wish to predict the value of the function $f(x^\star)$ at some new point $x^\star$ where we haven't seen previous data. Suppose that f has a mean-zero Gaussian process prior, $f \sim GP(0, C)$. Show that the posterior mean $E\{f(x^\star) \mid y_1, \ldots, y_N\}$ is a linear smoother, and derive expressions both for the smoothing weights and the posterior variance of $f(x^\star)$.

Recall that we can obtain the joint distribution and, consequently, the conditional distribution of $f(x^*)$ given $\mathbf{y}, \mathbf{f}, \sigma^2$ using parts (c) and (b) in the previous section, respectively. Therefore, the expected posterior mean is

$$\begin{aligned} E\left[f(x^*)|\mathbf{y}\right] &= m(x^*) + \tilde{C}^T(C - \sigma^2 I)^{-1}\left(\mathbf{y} - m(\mathbf{x})\right) \\ &= \tilde{C}^T(C - \sigma^2 I)^{-1}\mathbf{y} \\ &= \sum_{i=1}^{n} \tilde{C}(x_i, x^*)(C - \sigma^2 I)^{-1}(x_i, x_i) y_i \\ &= \sum_{i=1}^{n} w_i y_i, \end{aligned}$$

with smoothing weights $w_i = C(x^*, x_i) - C(x^*, \mathbf{x}_{-i})C(x_i, \mathbf{x}_{-i})$.

Additionally, we can write the posterior mean as a linear combination of the kernel function values, which may be interpretted as a correction to the prior mean consisting of a weighted combination of kernel functions (one for each training data point):

$$\begin{aligned} E\left[f(x^*)|\mathbf{y}\right] &= \sum_{i=1}^{n} \alpha_i C(x_i, x^*), \\ \alpha_i &= C^{-1}(\mathbf{x}, \mathbf{x}) y_i \end{aligned}$$

Recall from the aforementioned (b) that the posterior variance of $f(x^*)$ is

$$C^* - \tilde{C}^T C^{-1} \tilde{C}$$

where

$$\begin{aligned} C &= C(\mathbf{x}, \mathbf{x}), \\ \tilde{C} &= C(\mathbf{x}, x^*), \\ C^* &= C(x^*, x^*) \end{aligned}$$

C. Go back to the utilities data, and plot the point-wise posterior mean and 95% posterior confidence interval for the value of the function at each of the observed points x_i (again, superimposed on top of the scatter plot of the data itself). Choose τ_2^2 to be very small, say 10^{-6}, and choose (b, τ_1^2) that give a sensible-looking answer.

To obtain Figure 12, I set $\tau_2^2 = 10^{-6}$, used the MAP for σ^2 (i.e. $\hat{\sigma}^2 = 0.61$), and used a range of values for the other hyperparameters: $b \in \{3, 10, 15\}$ and $\tau_1^2 \in \{1, 5, 10\}$. Note that as b increases, the posterior means and intervals become more smooth; as τ_1^2 increases, we scale the information of weights more.

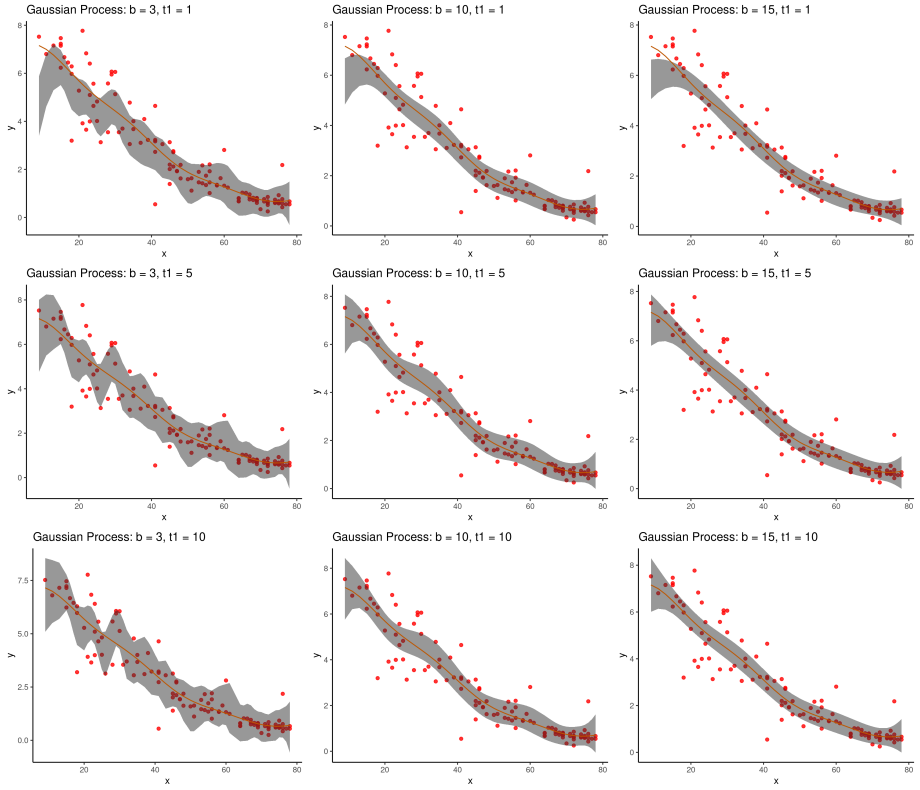

Figure 12: Gaussian Processes for Analyzing Utilities Under Various Hyperparameters.

D. Let $y_i = f(x_i) + \epsilon_i$, and suppose that f has a Gaussian-process prior under the Matern(5/2) covariance function C with scale τ_2^1, range b, and nugget τ_2^2. Derive an expression for the marginal distribution of $y = (y_1 \ldots, y_N)$ in terms of (τ_1^2, b, τ_2^2), integrating out the random function f. This is called a marginal likelihood.

Recall that, if $p(a|b) = \mathcal{N}(a|Ab, S)$ and $p(b) = \mathcal{N}(b|\mu, \Sigma)$, then

$$p(a) = \int p(a|b)p(b)db = \mathcal{N}(a|A\mu, A\Sigma A^T + S)$$

This is due to the convenient properties of multivariate normal distributions. Through applying this result, we see that

$$p(\mathbf{y}) = \mathcal{N}(0, C + \sigma^2 I)$$

E. Return to the utilities or ethanol data sets. Fix $\tau_2^2 = 0$, and evaluate the log of the marginal likelihood function $p(y \mid \tau_1^2, b)$ over a discrete 2-d grid of points. If you're getting errors in your code with $\tau_2^2 = 0$, use something very small instead. Use this plot to choose a set of values $(\hat{\tau_1^2}, \hat{b})$ for the hyperparameters. Then use these hyperparameters to compute the posterior mean for f, given y. Comment on any lingering concerns you have with your fitted model.

We can simply take the logarithm of the distribution in (D):

$$\begin{aligned}
\log p(\mathbf{y}) &= \log \left[|2\pi(C + \sigma^2 I)|^{-1/2} \exp \left\{ -\frac{1}{2} \mathbf{y}^T (C + \sigma^2 I)^{-1} \mathbf{y} \right\} \right] \\
&= -\frac{1}{2} \log |2\pi(C + \sigma^2 I)| - \frac{1}{2} \mathbf{y}^T (C + \sigma^2 I)^{-1} \mathbf{y} \\
&= -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |C + \sigma^2 I| - \frac{1}{2} \mathbf{y}^T (C + \sigma^2 I)^{-1} \mathbf{y}
\end{aligned}$$

On a grid of 250,000 points for (τ_1^2, b) for $0.001 \leq b \leq 100$ and $0.001 \leq \tau_1^2 \leq 100$, the point that maximizes the log-marginal likelihood function is $(\hat{\tau}_1^2, \hat{b}) = (39.67996, 61.52308)$. The posterior mean for $\mathbf{f}|\mathbf{y}$ is depicted in Figure 13.

Although I have not worked with Gaussian processes extensively before, it seems that the optimal b and τ_1^2 are larger than we would want. Perhaps they are so large because the data is *roughly* linear; however, it seems like we may be over-smoothing.

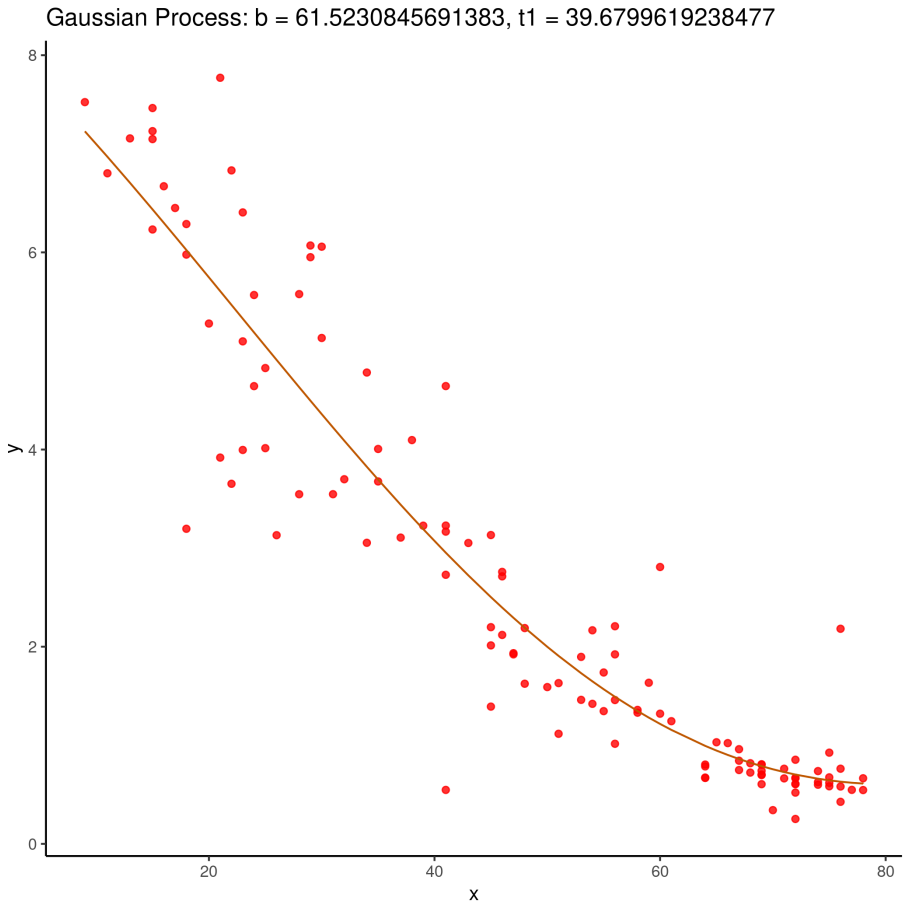

Figure 13: GP's Posterior Mean Under Optimal Hyperparameters.

F. In *weather.csv* you will find data on two variables from 147 weather stations in the American Pacific northwest: (1) pressure is the difference between the forecasted pressure and the actual pressure reading at that station in Pascals; (2) temperature is the difference between the forecasted temperature and the actual temperature reading at that station in Celsius. There are also latitude and longitude coordinates of each station. Fit a Gaussian process model for each of the temperature and pressure variables. Choose hyperparameters appropriately. Visualize your fitted functions (both the posterior mean and posterior standard deviation) on a regular grid using something like a contour plot or color image. Read up on the *image*, *filled.contour*, or *contourplot* functions in R. An important consideration: is Euclidean distance the appropriate measure to go into the covariance function? Or do we need separate length scales for the two dimensions, i.e.

$$d^2(x,z) = \frac{(x_1 - z_1)^2}{b_1^2} + \frac{(x_2 - z_2)^2}{b_2^2}.$$

Justify your reasoning for using Euclidean distance or this "nonisotropic" distance.

I, as well as several others,[1] believe that Euclidean distance is appropriate for one or two dimensions. Once our parameter space gets to higher dimensions, such as $d \geq 3$, Euclidean distance becomes less favorable. Therefore, I stand by our Euclidean distance for this application.

For pressure, Figures 14 and 15 depict the posterior predictive mean and standard deviation and the perspective of the posterior predictive mean at the optimal values $\hat{b}$ and $\hat{\tau}_1^2$, respectively. We obtained optimal values for the hyperparameters in the same way that did in (E), namely evaluating the log-marginal likelihood at a grid of points within the support of the data. The posterior predictive mean and standard deviation surfaces and perspective surface for temperature are depicted in Figures 16 and 17, respectively.

In the heat plots, white indicates highest values, then yellow, then red. For both temperature and pressure, we have very few observations in the left of the subplots. Therefore, the grid values in that area are too far from data to adequately learn. The posterior predictive standard deviation is incredibly high there, indicated by the bright white.

[1](1): Domingos, Pedro. "A few useful things to know about machine learning." Communications of the ACM 55.10 (2012): 78-87.
(2): Aggarwal, Charu C., Alexander Hinneburg, and Daniel A. Keim. "On the surprising behavior of distance metrics in high dimensional space." International conference on database theory. Springer, Berlin, Heidelberg, 2001.
(3): Mirkes, Evgeny M., Jeza Allohibi, and Alexander Gorban. "Fractional norms and quasinorms do not help to overcome the curse of dimensionality." Entropy 22.10 (2020): 1105.

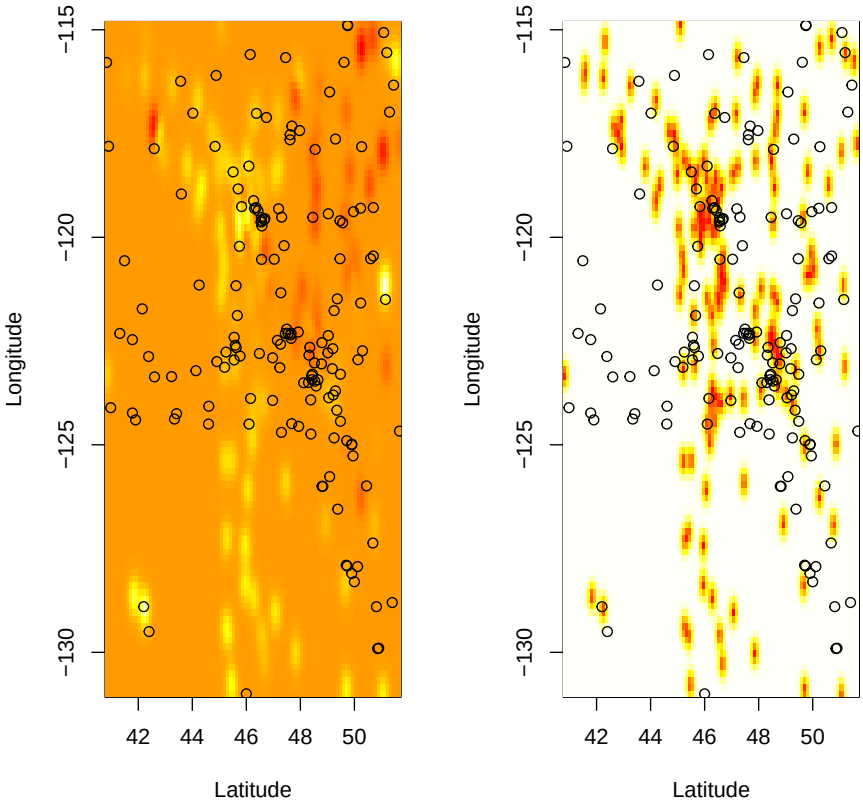

Figure 14: Posterior predictive mean (left) and standard deviation (right) surfaces for pressure. Data points are indicated by open circles.

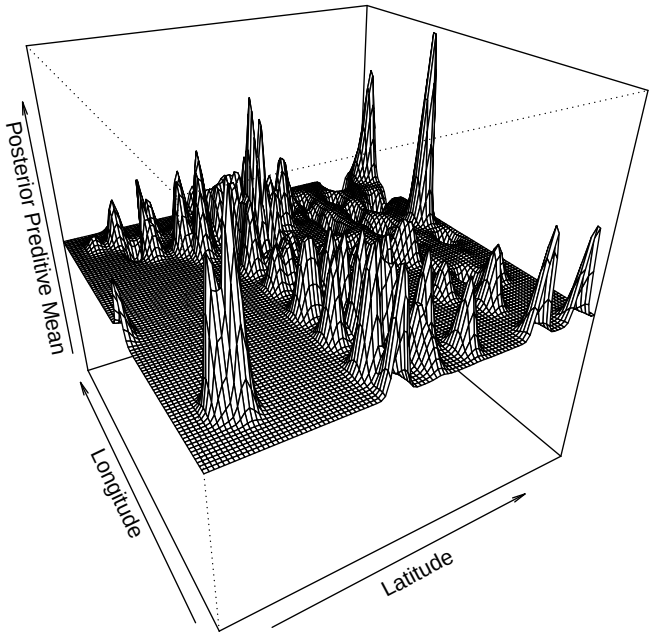

Figure 15: Perspective view on the posterior mean surface for pressure.

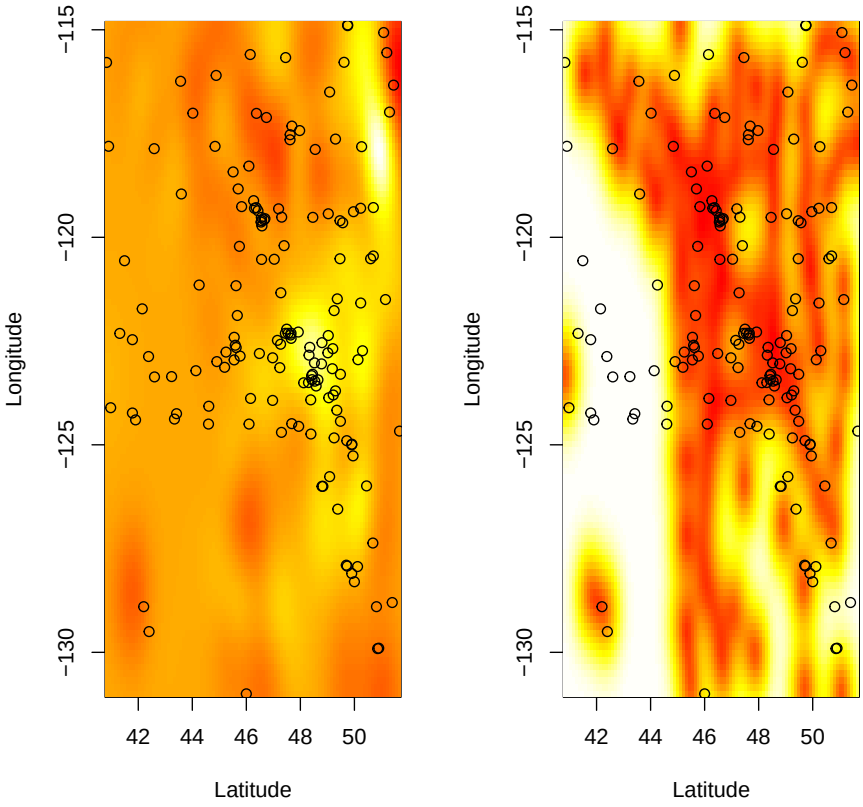

Figure 16: Posterior predictive mean (left) and standard deviation (right) surfaces for temperature. Data points are indicated by open circles.

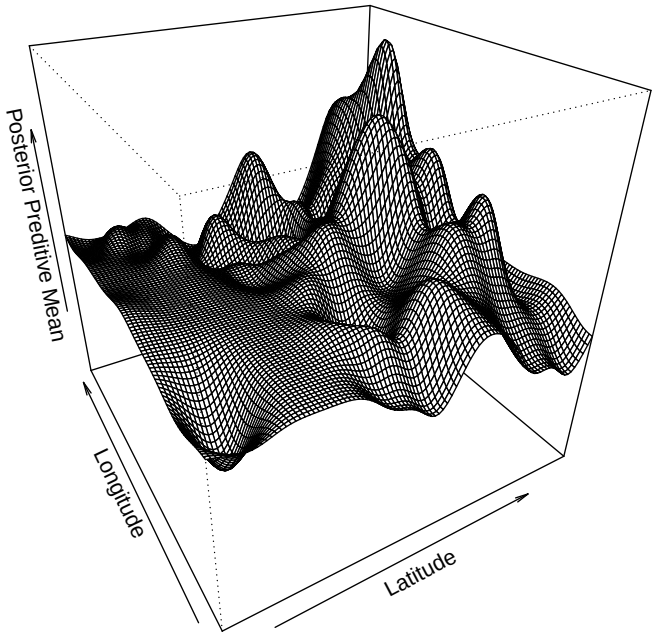

Figure 17: Perspective view on the posterior mean surface for temperature.