Problem 1

a)

Given $X_1, X_2, ..., X_n \sim N(\theta, \sigma^2)$ — σ is known.

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $se^2 = \sigma^2/n$

Prior distribution for $\theta \sim N(a, b^2)$. We have,

f(θ) $= (2\pi b^2)^{\frac{-1}{2}}.exp(\frac{-(\theta-a)^2}{2b^2})$ $--$ **1**

$f(\mathbf{x}|\theta) = (2\pi\sigma^2)^{\frac{-1}{2}}.\Pi_{i=1}^{n} exp(\frac{-(x_i-\theta)^2}{2\sigma^2})$ $--$ **2**

$f(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta).f(\theta)$

$Using\ 1\ and\ 2;$

$f(\theta|\mathbf{x}) = (2\pi b^2)^{\frac{-1}{2}}.exp(\frac{-(\theta-a)^2}{2b^2}).(2\pi\sigma^2)^{\frac{-1}{2}}.\Pi_{i=1}^{n} exp(\frac{-(x_i-\theta)^2}{2\sigma^2})$

$= exp(\frac{-1}{2}\{\frac{\sum_{i=1}^{n}(x_i-\theta)^2}{\sigma^2} + \frac{(\theta-a)^2}{b^2}\})$

$= exp(\frac{-1}{2}\{\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i^2 + \theta^2 - 2x_i\theta) + \frac{(\theta-a)^2}{b^2}\})$

$Ignoring\ constants;$

$= exp(-\frac{\theta^2 n}{2\sigma^2} + \frac{2\theta\sum_{i=1}^{n} x_i}{\sigma^2} - \frac{\theta^2}{2b^2} - \frac{a^2}{2b^2} + \frac{\theta a}{b^2})$

$= exp(\theta^2(-\frac{n}{2\sigma^2} - \frac{1}{2b^2}) + \theta(\frac{n\overline{X}}{\sigma^2} + \frac{a}{b^2}) + constant)$ $--$ **3**

$For\ a\ Normal\ distribution\ with\ response\ y\ with\ mean\ x\ and\ variance\ y^2\ we\ have$

$g(r) = (2\pi y^2)^{\frac{-1}{2}} exp\{(r-x)^2/2y^2\}$

$\propto exp\{\frac{-1}{2}r^2y^{-1} + rx/y + constant\}$ $--$ **4**

$Comparing\ equations$ **3** and **4**

$x = y^2(\frac{a}{b^2} + \frac{n\overline{X}}{\sigma^2})$ $--$ **5**;

$y^2 = (\frac{1}{b^2} + \frac{n}{\sigma^2})^{-1}$ $--$ **6**

$Solving\ for\ y$

$y^2 = (\frac{1}{b^2} + \frac{1}{se})^{-1}$

$y^2 = \frac{b^2.se^2}{b^2+se^2}$ $--$ **7**

$Putting$ **7** in **5**;

$x = \frac{b^2.se^2}{b^2+se^2}.\frac{b^2.\overline{X}+a.se^2}{b^2.se^2}$

$Thus,\ we\ have:\ x = \frac{b^2.\overline{X}+a.se^2}{b^2+se^2};\ y^2 = \frac{b^2.se^2}{b^2+se^2}$

Hence Proved!

Finding an interval $C = (c, d)$ such that $P(\theta \in C|\mathbf{x}) = (1 - \alpha)$.

Choose c and d such that: $P(\theta < c|\mathbf{x}) = 0.025$ and $P(\theta > d|\mathbf{x}) = 0.025$

$$
\begin{aligned}
P(d < \theta < c|\mathbf{x}) &= P(\frac{(d-x)}{y} < \frac{(\theta - x)}{y} < \frac{(c-x)}{y}|\mathbf{x}) \\
&= P(\frac{(d-x)}{y} < Z < \frac{(c-x)}{y}) = (1-\alpha) \quad --\mathbf{I} \\
&\quad From\ definition\ of\ (1-\alpha)\ C.I; \\
&\quad P(-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}) = (1-\alpha) \quad --\mathbf{II} \\
&\quad Comparing \quad --\mathbf{I}\ and \quad --\mathbf{II} \\
&\quad c = x + y.z_{\frac{\alpha}{2}}; \qquad d = x - y.z_{\frac{\alpha}{2}} \\
&\quad Posterior\ interval = (x - y.z_{\frac{\alpha}{2}}, x + y.z_{\frac{\alpha}{2}})
\end{aligned}
$$

Since $x \to \overline{X}$ *and* $y \to se$ *as* $n \ \to \infty$

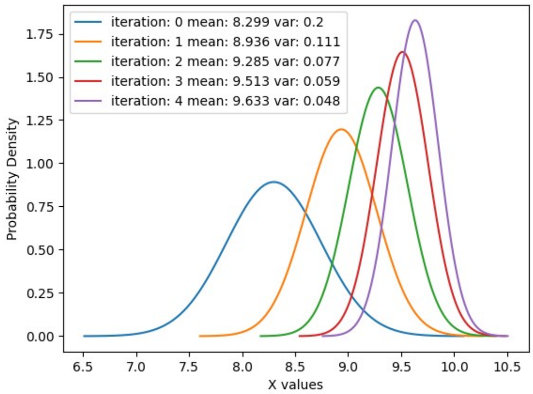
$Posterior\ interval = (\overline{X} \pm z_{\frac{\alpha}{2}}.se)$

This is the frequentist confidence interval.

Problem 2

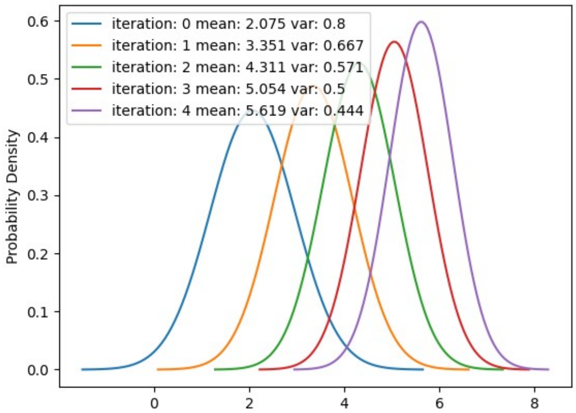
a)
mean variance
8.299 0.2
8.936 0.111
9.285 0.077
9.513 0.059
9.633 0.048

The posterior is trying to converge i.e. move away from prior.



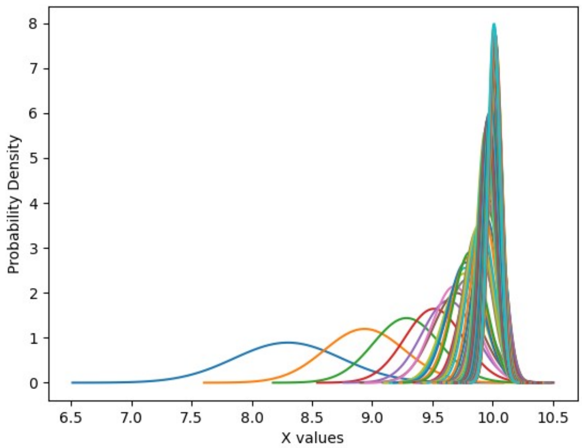
b)
mean variance
2.075 0.8
3.351 0.667
4.311 0.571
5.054 0.5
5.619 0.444

When σ is large, the likelihood gets wider, so the posterior probabilities move slower.



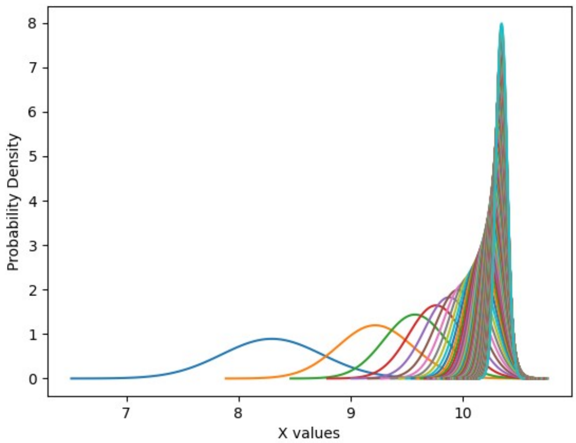
c)
Final mean and variance:
10.011 0.002

The posterior is trying to converge with mean = 10.011



d)
Final mean and variance:
10.348 0.002

The posterior is trying to converge with mean = 10.348 which is likely a wrong value since we are trying to overfit on the same smaller dataset. The variance is same but the mean is different



Problem 3

(a)

First we define the fitted equation to be an equation:

$$\hat{Y} = \beta_0 + \beta_1 X$$

Now, for each observed response Y_i, with a corresponding predictor variable X_i, so we would like to minimize the sum of the squared distances of each observed response to its fitted value.

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$$

Thus, we set the partial derivatives of $SSE(\beta_0, \beta_1)$ with respect β_0 and β_1 equal to zero

$$\begin{aligned}\frac{dSSE}{d\beta_0} &= \sum_{i=1}^{n} 2(-1)(Y_i - \beta_0 - \beta_1 X_i) = 0 \\ &\Rightarrow \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i) = 0\end{aligned}$$

$$\begin{aligned}\frac{dSSE}{d\beta_1} &= \sum_{i=1}^{n} 2(-X_i)(Y_i - \beta_0 - \beta_1 X_i) = 0 \\ &\Rightarrow \sum_{i=1}^{n} X_i(Y_i - \beta_0 - \beta_1 X_i) = 0\end{aligned}$$

The we could get 2 normal equations:

$$\beta_0 n + \beta_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i$$

$$\beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i$$

For the first normal equation, we could get

$$\beta_0 = \frac{\sum_{i=1}^{n} Y_i - \beta_1 \sum_{i=1}^{n} X_i}{n}$$

Substitute into the second normal equation, yields,

$$\begin{aligned}\frac{\sum_{i=1}^{n} Y_i - \beta_1 \sum_{i=1}^{n} X_i}{n} \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2 &= \sum_{i=1}^{n} X_i Y_i \\ \beta_1\Big(\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}\Big) &= \sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n} \\ \beta_1\Big(\sum_{i=1}^{n} X_i^2 - 2\frac{(\sum_{i=1}^{n} X_i)^2}{n} + \frac{(\sum_{i=1}^{n} X_i)^2}{n}\Big) &= \sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n} - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n} + \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n} \\ \beta_1\Big(\sum_{i=1}^{n} X_i^2 - 2\sum_{i=1}^{n} X_i \frac{\sum_{i=1}^{n} X_i}{n} + \sum_{i=1}^{n}\big(\frac{\sum_{i=1}^{n} X_i}{n}\big)^2\Big) &= \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \bar{Y} - \sum_{i=1}^{n} Y_i \bar{X} + \sum_{i=1}^{n} \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n^2} \\ \beta_1 \sum_{i=1}^{n}\Big(X_i^2 - 2X_i \frac{\sum_{i=1}^{n} X_i}{n} + \big(\frac{\sum_{i=1}^{n} X_i}{n}\big)^2\Big) &= \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \bar{Y} - \sum_{i=1}^{n} Y_i \bar{X} + \sum_{i=1}^{n} \bar{X}\bar{Y} \\ \beta_1 \sum_{i=1}^{n}(X_i^2 - \bar{X})^2 &= \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \\ \Rightarrow \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\end{aligned}$$

Thus we could have

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

(b)

First, we rewrite $\hat{\beta_1}$ as

$$\hat{\beta_1} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})Y_i}{S_{xx}} = \sum_{i=1}^n \frac{X_i - \bar{X})Y_i}{S_{xx}} = \sum_{i=1}^n c_i Y_i$$

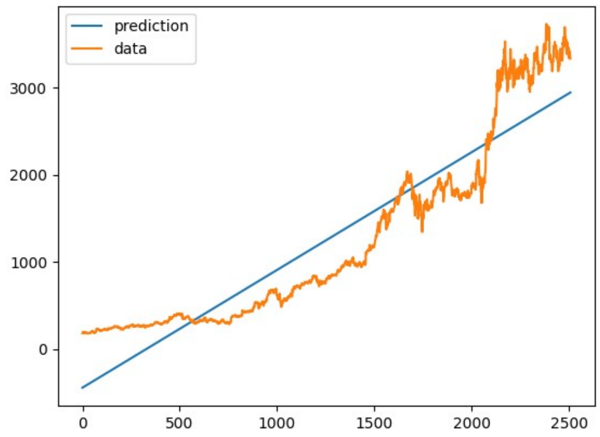
and we could have $\sum_{i=1}^n c_i = \sum_i \frac{X_i - \bar{X}}{S_{xx}} = \frac{n\bar{X} - n\bar{X}}{S_{xx}} = 0$. Also, $E[\epsilon_i] = 0$. Then, we have

$$\begin{aligned}
E[\hat{\beta_1}] &= \sum_{i=1}^n c_i E[Y_i] \\
&= \sum_{i=1}^n c_i E[\beta_0 + \beta_1 X_i + \epsilon_i] \\
&= \beta_0 \sum_{i=1}^n c_i + \beta_1 \sum_{i=1}^n c_i X_i + \sum_{i=1}^n c_i E[\epsilon_i] \\
&= \beta_1 \sum_{i=1}^n \frac{(X_i - \bar{X})X_i}{\sum_{i=1}^n (X_i - \bar{X})^2} \\
&= \beta_1 \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \\
&= \beta_1
\end{aligned}$$

$$\begin{aligned}
E[\hat{\beta_0}] &= E[\bar{Y} - \hat{\beta_1}\bar{X}] \\
&= E[\frac{\sum_{i=1}^n Y_i}{n} - \frac{\sum_{i=1}^n \hat{\beta_1} X_i}{n}] \\
&= \frac{\sum_{i=1}^n E[\beta_0 + \beta_1 X_i - \hat{\beta_1} X_i]}{n} \\
&= \frac{\sum_{i=1}^n (\beta_0 + \beta_1 X_i - \beta_1 X_i)}{n} \\
&= \beta_0
\end{aligned}$$

Problem 4

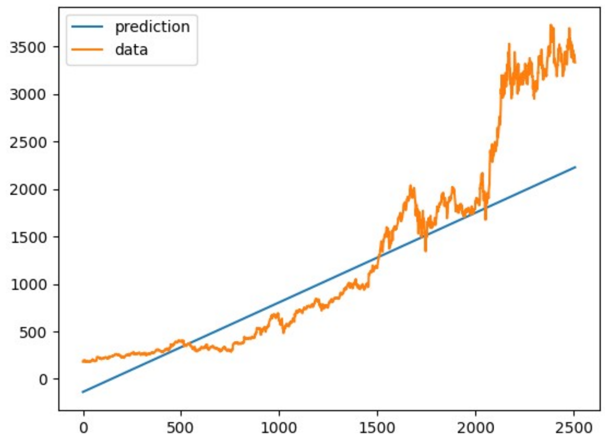
a)



b)

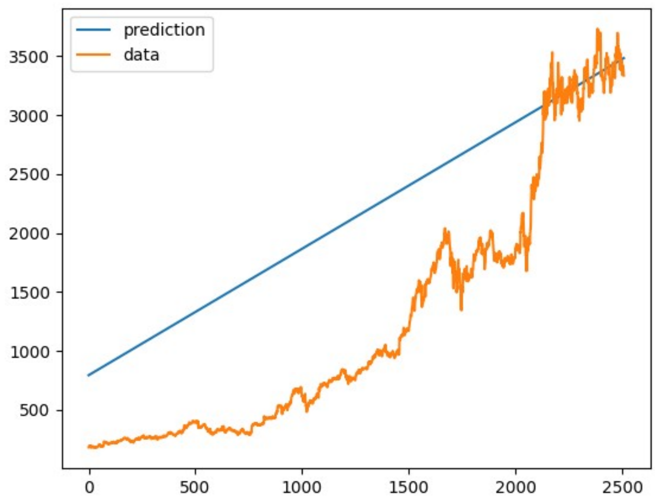
Prediction = 2189s
Real = 3318

%age inc = 51%



c)

Prediction = 5869



d) Prediction = 4368. 25% lower

e)
The company’s stock value increased by 51% due to the pandemic. It’s value would have been 25% lower in 10 year if there was no pandemic assuming the advantage received won’t be negated by some other factors.

Note: The observations made are subjective. Other answers may also receive full marks. As there were confusions regarding the date ranges in the questions, we will allow different possible numerical values.

Problem 5

bd = number of bedrooms
bth = number of bathrooms
sqft = square feet
fl = floor number
am1 – am7 = amenities
age = building age

a)
Rent = -401.32389624*bd + 1376.28742978*bth + 4.91662231*sqft – 819.84864379

b)
Rent = -352.18459684*bd + 1284.64143349*bth + 4.8537158*sqft – 31.96502037*fl -1082.55710925

c)
Rent = -351.72811588*bd + 1288.16221231*bth + 4.84464694*sqft - 32.09372554*fl + 36.23342659*am1 + 148.85190701*am2 -162.24810373*am3 + 120.14434719*am4 -15.30013149*am5 -72.08247686*am6 + 4.82835113*am7 -1094.27990489

d)
Rent = -325.61072006*bd + 1162.4380147*bth + 4.96254301*sqrt + 22.91936795*fl + 39.55090181*am1 + 151.29489546*am2 -135.38844076*am3 + 94.9498263*am4 -24.46241797*am5 -91.20938663*am6 -18.87343528*am7 -6.74874057*age -602.52281811
I chose age of building because it has good correlation with rent.

e)
There is an improvement in SSE from a) to b) but not from b) to c). c) to d) again has improvement.

Problem 6

a)

$H \equiv RV$ for the soil type.

The two hypotheses are: $\boldsymbol{H_0}: H = 0$ and $\boldsymbol{H_1}: H = 1$ with $P(H = 0) = p$ and $P(H = 1) = (1 - p)$

Observations of water concentration metric $\boldsymbol{w} = \{w_1, \dots w_n\}$

$f_W(w|H = 0) = N(w; -\mu, \sigma^2)$ and $f_W(w|H = 1) = N(w; \mu, \sigma^2)$

Also $w_i s$ are conditionally independent of each other given the hypothesis/soil type.

$$P(H = 0|\boldsymbol{w}) = \frac{P(\boldsymbol{w}|H = 0)P(H = 0)}{P(\boldsymbol{w})} \qquad \textit{By Bayes theorem}$$

$$\Rightarrow \; P(H = 0|\boldsymbol{w}) = \frac{P(H=0)}{P(\boldsymbol{w})} \prod_{i=1}^{n} f_W(w_i|H = 0) \quad \because (w_i|H = h) \perp (w_i|H = h)$$

$$\Rightarrow \; P(H = 0|\boldsymbol{w}) = c.p.\exp\left(-\frac{\sum_i (w_i + \mu)^2}{2\sigma^2}\right)$$

We choose $H_0 (C = 0)$ if $P(H = 0|\boldsymbol{w}) \geq P(H = 1|\boldsymbol{w})$, i.e.

$$c.p.\exp\left(-\frac{\sum_i (w_i + \mu)^2}{2\sigma^2}\right) \geq c.(1-p).\exp\left(-\frac{\sum_i (w_i - \mu)^2}{2\sigma^2}\right)$$

$$\Rightarrow \exp\left(-\frac{\sum_i (w_i + \mu)^2 - \sum_i (w_i - \mu)^2}{2\sigma^2}\right) \geq \frac{(1-p)}{p}$$

$$\Rightarrow \exp\left(-\frac{2\mu \sum_i w_i}{\sigma^2}\right) \geq \frac{(1-p)}{p}$$

$$\left(\sum_i w_i\right) \leq \frac{\sigma^2}{2\mu} \ln\left(\frac{p}{1-p}\right)$$

b)

For $P(H_0) = 0.1$, the hypothesis selected are: 0 1 0 0 1 0 1 1 0 1

For $P(H_0) = 0.3$, the hypothesis selected are: 0 1 0 0 1 0 1 1 0 1

For $P(H_0) = 0.5$, the hypothesis selected are: 0 1 0 0 1 0 1 1 0 1

For $P(H_0) = 0.8$, the hypothesis selected are: 0 1 0 0 1 0 1 1 0 1 c)

We choose H_0 i.e. $C = 0$ iff

$$\left(\sum_i w_i\right) \leq \frac{\sigma^2}{2\mu}\ln\left(\frac{p}{1-p}\right)$$

We choose H_1 iff

$$\left(\sum_i w_i\right) > \frac{\sigma^2}{2\mu}\ln\left(\frac{p}{1-p}\right)$$

$$P(C=0|H=1) = P\left(\left(\sum_i w_i\right) \leq \frac{\sigma^2}{2\mu}\ln\left(\frac{p}{1-p}\right) |\ (H=1)\right)$$

$\because w_i|(H=0) \sim N(-\mu,\sigma^2)$

$$\Rightarrow \left(\sum_i w_i\right)|(H=0) \sim N(-n\mu, n\sigma^2)$$

$$\Rightarrow \left(\sum_i w_i\right)|(H=1) \sim N(n\mu, n\sigma^2)$$

$$\Rightarrow P(C=0|H=1) = \Phi\left(\frac{\frac{\sigma^2}{2\mu}\ln\left(\frac{p}{1-p}\right) - n\mu}{\sqrt{n\sigma^2}}\right) \quad \because if\ X \sim N(\mu,\sigma^2) \Rightarrow \frac{X-\mu}{\sigma} \sim N(0,1)$$

Similarly,

$$P(C=1|H=0) = P\left(\left(\sum_i w_i\right) > \frac{\sigma^2}{2\mu}\ln\left(\frac{p}{1-p}\right) |\ (H=0)\right)$$

$$\Rightarrow P(C=1|H=0) = 1 - \Phi\left(\frac{\frac{\sigma^2}{2\mu}\ln\left(\frac{p}{1-p}\right) + n\mu}{\sqrt{n\sigma^2}}\right)$$

$$\therefore AEP = (1-p).\Phi\left(\frac{\frac{\sigma^2}{2\mu}\ln\left(\frac{p}{1-p}\right) - n\mu}{\sqrt{n\sigma^2}}\right) + p.\left(1 - \Phi\left(\frac{\frac{\sigma^2}{2\mu}\ln\left(\frac{p}{1-p}\right) + n\mu}{\sqrt{n\sigma^2}}\right)\right)$$