P_1 (a) $P(x_1, x_2 \dots x_N|\lambda) = \prod_{i=1}^{N} p(x_i|\lambda) = \prod_{i=1}^{N} (\frac{\lambda^{x_i}}{x_i!} e^{-\lambda}) = \frac{\lambda^{\sum_{i=1}^{N} x_i} e^{-N\lambda}}{\prod_{i=1}^{N} x_i!}$

(b) $\lambda_{ML} = \arg\max_{\lambda} p(x_1, x_2 \dots x_N) = \arg\max_{\lambda} \prod_{i=1}^{N} (\frac{\lambda^{x_i}}{x_i!} e^{-\lambda})$

$\nabla_\lambda \prod_{i=1}^{N} (\frac{\lambda^{x_i}}{x_i!} e^{-\lambda}) = 0$

$\because \ln(\prod_{i}^{N} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}) = \sum_{i}^{N} \ln(\frac{\lambda^{x_i}}{x_i!} e^{-\lambda})$

$\therefore \nabla_\lambda \sum_{i=1}^{N} \ln(\frac{\lambda^{x_i}}{x_i!} e^{-\lambda}) = \sum_{i=1}^{N} \nabla_\lambda \ln(\frac{\lambda^{x_i}}{x_i!} e^{-\lambda}) = \sum_{i=1}^{N} \nabla_\lambda (\ln \frac{\lambda^{x_i}}{x_i!} - \lambda) = \sum_{i=1}^{N} \nabla_\lambda (\ln \lambda^{x_i} - \ln x_i! - \lambda)$

$= \sum_{i=1}^{N} (\frac{x_i \lambda^{x_i - 1}}{\lambda^{x_i}} - 1) = \sum_{i=1}^{N} (x_i \lambda^{-1} - 1) = \sum_{i=1}^{N} (\frac{x_i - \lambda}{\lambda}) = 0$

$\therefore \lambda_{ML} = \frac{1}{N} \sum_{i=1}^{N} x_i$

(c) $\lambda_{MAP} = \arg\max_{\lambda} \ln p(\lambda|x)$

$= \arg\max_{\lambda} \ln \frac{p(x|\lambda) p(\lambda)}{p(x)}$

$= \arg\max_{\lambda} (\ln p(x|\lambda) + \ln p(\lambda) - \ln p(x))$ $\qquad \Gamma(a) = (a-1)!$

$\lambda_{MAP} = \arg\max_{\lambda} (\sum_{i=1}^{N} \ln(\frac{\lambda^{x_i}}{x_i!} e^{-\lambda}) + \ln \frac{b^a \lambda^{a-1} e^{-b\lambda}}{\Gamma(a)})$

$= \arg\max_{\lambda} [\sum_{i=1}^{N} (\ln \lambda^{x_i} - \ln x_i! - \lambda) + \ln \beta^a + \ln \lambda^{a-1} - b\lambda - \ln \Gamma(a)]$

$\nabla_\lambda [\sum_{i=1}^{N} (\ln \lambda^{x_i} - \ln x_i! - \lambda) + \ln \lambda^{a-1} - b\lambda + \text{Const}]$

$= \sum_{i=1}^{N} (\frac{x_i \lambda^{x_i - 1}}{\lambda^{x_i}} - 1) + \frac{(a-1)\lambda^{a-2}}{\lambda^{a-1}} - b$

$= \sum_{i=1}^{N} (x_i \lambda^{-1} - 1) + (a-1)\lambda^{-1} - b = 0$

$\frac{\sum_{i=1}^{N} x_i - \lambda N}{\lambda} + \frac{a-1}{\lambda} - b = 0$

$\sum_{i=1}^{N} x_i - \lambda N + a - 1 - \lambda = 0$

$\lambda_{MAP} = \frac{\sum_{i=1}^{N} x_i + a - 1}{N + b}$

(d) $p(\lambda|x) = \frac{p(x|\lambda) p(\lambda)}{p(x)}$

$\propto p(x|\lambda) p(\lambda)$

$\propto \prod_{i=1}^{N} (\frac{\lambda^{x_i}}{x_i!} e^{-\lambda}) \cdot \frac{\lambda^{a-1} b^a e^{-b\lambda}}{(a-1)!}$

$\propto \prod_{i=1}^{N} (\frac{\lambda^{x_i}}{x_i!}) e^{-(N+b)\lambda} \cdot \frac{\lambda^{a-1} b^a}{(a-1)!}$

$$\propto \frac{\lambda^{\sum_{i=1}^{N} x_i + (a-1)} b^a e^{-(N+b)\lambda}}{\prod_{i=1}^{N} x_i!\,(a-1)!}$$

$$\propto \lambda^{\sum_{i=1}^{N} x_i + a - 1} e^{-(N+b)\lambda}$$

$$\therefore\ p(\lambda|x) \sim \text{Gamma}\left(\sum_{i=1}^{N} x_i + a,\ N+b\right)$$

(e) $$\text{mean} = \frac{\sum_{i=1}^{N} x_i + a}{N+b}$$

$$\text{var} = \frac{\sum_{i=1}^{N} x_i + a}{(N+b)^2}$$

Discuss: λ_{MAP} is the mode of the mean of this posterior distribution, it's the maximum value of this mean. λ_{ML} is similar to the mean except for a, b. When dataset is large λ_{ML} equals λ_{MAP}. λ_{MAP} considers prior distribution compared to λ_{ML}.

P2 $$E[W_{RR}] = E[(\lambda I + X^T X)^{-1} X^T y]$$

$$= (\lambda I + X^T X)^{-1} X^T E[y]$$

$$= (\lambda I + X^T X)^{-1} X^T X w$$

$$\text{Var}[W_{RR}] = E[W_{RR} W_{RR}^T] - E[W_{RR}]E[W_{RR}]^T \qquad \text{Let } A = (\lambda I + X^T X)^{-1}$$

$\because E[yy^T] = \sigma^2 I + X w w^T X^T$

$$= A X^T E[yy^T] X A^T - A X^T X w w^T X^T X A^T$$

$$= A X^T \sigma^2 I X A^T + A X^T \cancel{X w w^T} X^T X A^T - A X^T \cancel{X w w^T} X^T X A^T$$

$$= A X^T \sigma^2 I X A^T$$

$$A = (\lambda I + X^T X)^{-1} = ((X^T X)(\lambda (X^T X)^{-1} + I))^{-1} = (\lambda (X^T X)^{-1} + I)(X^T X)^{-1}$$

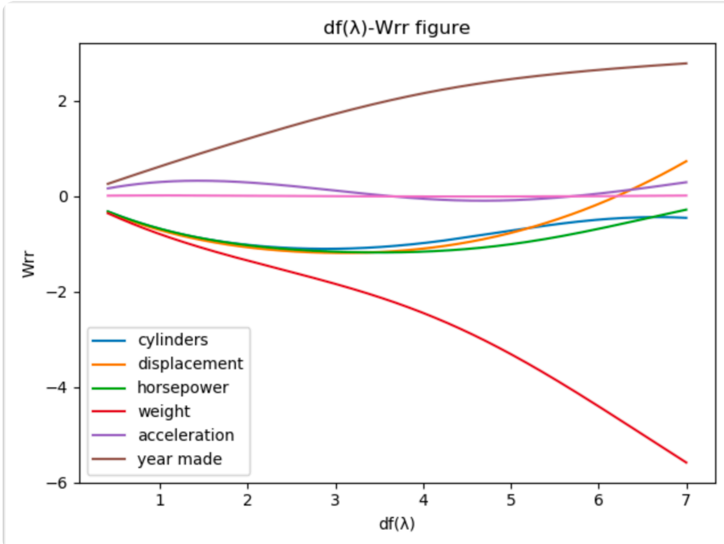
$$\text{Let } Z = (I + \lambda (X^T X)^{-1})^{-1}$$

$$\therefore\ A = Z (X^T X)^{-1} \qquad ((X^T X)^{-1})^T = ((X^T X)^T)^{-1} = (X^T X)^{-1}$$

$$\therefore\ \text{Var}[W_{RR}] = Z (X^T X)^{-1} X^T \sigma^2 X (X^T X)^{-1} Z^T = Z (X^T X) Z^T \sigma^2$$

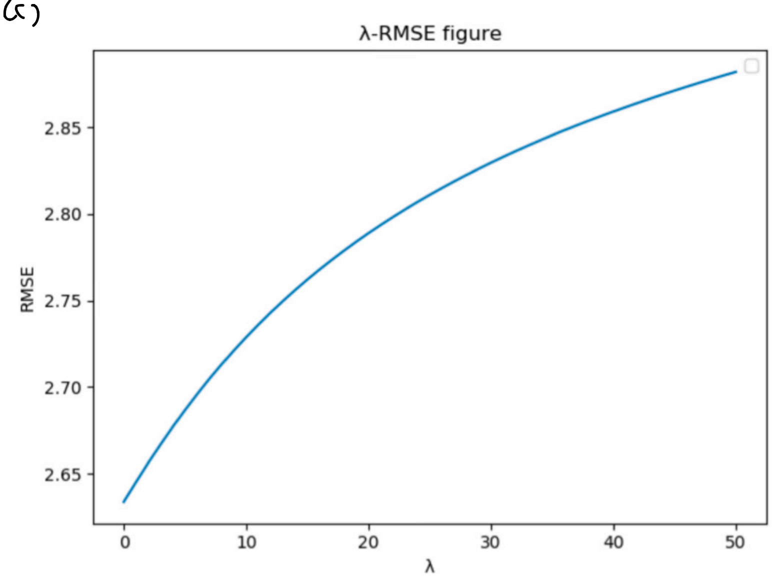
P3. Part 1.

(a)



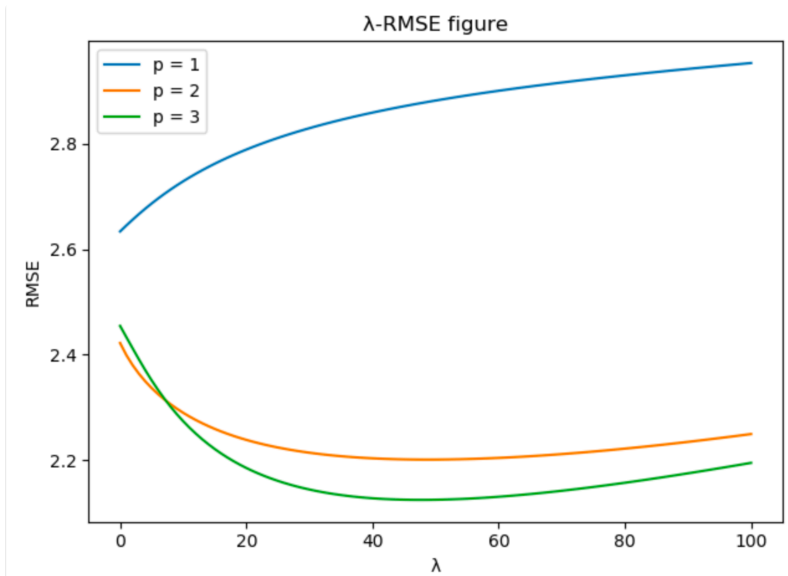
(b) The features "Year made" and "weight" clearly stand out. From the figure, we can see that with the increase of df(λ), "Year made" feature's Wrr increase a lot, which indicating that with the decrease of λ, increase of df(λ), this "Year made" feature Wrr weights more and becomes more important. As for "weight", with the increase of df(λ), decrease of λ, this "weight" feature's Wrr decreases a lot, indicating that this feature weights less and becomes less important.

(c)



From the figure, we can see that with the increase of λ, the "RMSE" becomes larger and larger, which is not idea. For this problem, the figure shows that it's better to choose small λ which can get smaller RMSE. $\lambda = 0$, which is the least square solution, meaning that least square solution is better for this problem.

(d)



Base on the plot, when λ is less than (about) 10, I'll choose $p=2$ since it'll lead to the smallest RMSE. When λ is greater than (about) 10. I'll choose $p=3$ since it'll lead to the smallest RMSE.

When $p=1$. the plot shows that with the increase of λ. RMSE increases. So the idea λ is 0 for $p=1$.

When $p=2$ and 3, the plot shows that RMSE first decreases as λ increasing. When λ is about 40, RMSE starts increasing. So the idea λ for $p=2$ and 3 is 40