Homework 1. Zihao Zhang zz2763.

Problem 1.

(a) since x_i is i.i.d. $(i = 1, 2, \cdots, \infty)$

$$p(x_1, \cdots, x_N) = \prod_{i=1}^{N} p(x_i|\lambda)$$

$$= \frac{\lambda^{\sum_{i=1}^{N} x_i}}{\prod_{i=1}^{N} x_i!} e^{-\lambda N}$$

is the joint likelihood distribution of data $(x_1, \cdots, x_N)$

(b) when $p(x_1, \cdots, x_N)$ reaches maximimum

$$\frac{\partial}{\partial \lambda} \ln p(x_1, \cdots, x_N) = 0$$

then $\frac{\partial}{\partial \lambda} \left(\sum_{i=1}^{N} x_i \ln \lambda - \sum_{i=1}^{N} \ln x_i! - \lambda N \right) = 0$ $\quad \left(\frac{\partial^2}{\partial \lambda^2} \ln p(X) < 0 \right)$

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

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

(c) given $p(\lambda) = \text{gamma}(a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}$

$$p(\lambda | x_1, \cdots, x_N) = \prod_{i=1}^{N} p(x_i|\lambda) \cdot p(\lambda) \Big/ \int p(X|\lambda) p(\lambda) d\lambda$$

therefore $\lambda_{MAP} = \underset{\lambda}{\text{argmax}}\, p(\lambda | x_1, \cdots, x_N)$

$$= \underset{\lambda}{\text{argmax}} \prod_{i=1}^{N} p(x_i|\lambda) \cdot p(\lambda) \quad (\text{since } p(X) \text{ does not depend on } \lambda)$$

$$= \underset{\lambda}{\text{argmax}} \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} \cdot \frac{\lambda^{\sum_{i=1}^{N} x_i}}{\prod_{i=1}^{N} x_i!} e^{-\lambda N}$$

$$\frac{\partial}{\partial \lambda} \ln \left[\frac{b^a}{\Gamma(a)} \cdot \frac{1}{\prod_{i=1}^{N} x_i!} \cdot e^{-(b+N)\lambda} \cdot \lambda^{a-1+\sum_{i=1}^{N} x_i} \right] = 0$$

$$\Rightarrow \quad -(b+N) + \left(a - 1 + \sum_{i=1}^{N} x_i\right) \cdot \frac{1}{\lambda} = 0$$

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

(d) according to Bayes rule,

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

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

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

Hence $p(\lambda|X)$ is gamma distribution and $p(\lambda|X) \sim gamma(a+\sum_{i=1}^{N} x_i, b+N)$

(e) given a gamma function $p(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}$

$$E[\lambda] = \frac{a}{b}, \quad Var[\lambda] = \frac{a}{b^2}$$

hence $E[\lambda_{MAP}] = \frac{a+\sum_{i=1}^{N} x_i}{b+N}$

$$Var[\lambda_{MAP}] = \frac{a+\sum_{i=1}^{N} x_i}{(b+N)^2}$$

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

Problem 2

(a) in ridge regression, loss function $\mathcal{L} = \|y - Xw\|^2 + \lambda\|w\|^2$

we can solve $w_{RR} = (\lambda I + X^TX)^{-1}X^Ty$

given $w_{ML} = (X^TX)^{-1}X^Ty$

hence $w_{RR} = (\lambda I + X^TX)^{-1} \cdot (X^TX) \cdot w_{ML}$

$= (\lambda(X^TX)^{-1} + I)^{-1} w_{ML}$

$E[w_{RR}] = (\lambda(X^TX)^{-1} + I)^{-1} E[w_{ML}]$

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

$= (\lambda I + X^TX)^{-1} \cdot X^TXw$

and $Var[w_{RR}] = Var[(\lambda(X^TX)^{-1} + I)^{-1} \cdot w_{ML}]$

$= (\lambda(X^TX)^{-1} + I)^{-1} \cdot Var[w_{ML}] \cdot ((\lambda(X^TX)^{-1} + I)^{-1})^T$

$= Z\sigma^2(X^TX)^{-1}Z^T$, given $Z = (\lambda(X^TX)^{-1} + I)^{-1}$.

(b) SVD of X can be $X = USV^T$

hence $(X^TX)^{-1} = (VSU^T \cdot USV^T)^{-1}$

$= VS^{-2}V^T$

in (a), we have got $w_{RR} = (\lambda(X^TX)^{-1} + I)^{-1} w_{LS}$

$= (\lambda VS^{-2}V^T + I)^{-1} w_{LS}$

$= V(\lambda S^{-2} + I)^{-1} V^T w_{LS}$

M is used to denote $(\lambda S^{-2} + I)^{-1}$, as the sigular values.

and $S = diag(S_{ii})$, $S^{-1} = diag(S_{ii}^{-1})$

so $M = (\lambda S^{-2} + I)^{-1}$

$= [\lambda \cdot diag(S_{ii}^{-2}) + diag(1)]^{-1}$

$= [diag(\frac{\lambda}{S_{ii}^2} + 1)]^{-1}$

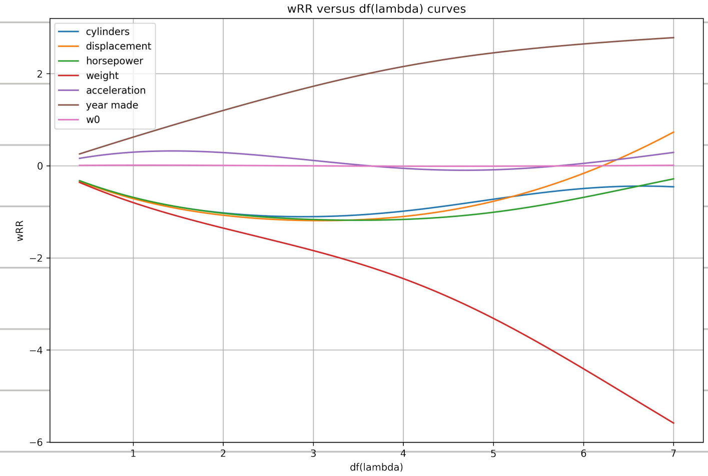
$= diag(\frac{S_{ii}^2}{\lambda + S_{ii}^2})$

therefore we derives $w_{RR} = VMV^T w_{LS}$, as a function of w_{LS},

the sigular values, and V of matrix X.

Problem 3

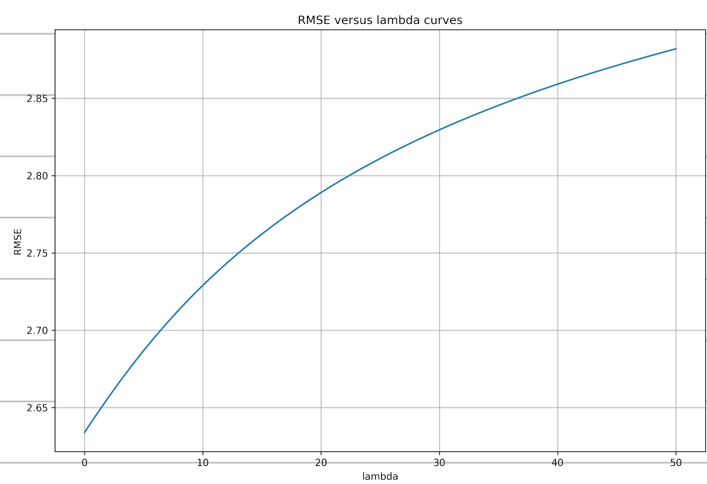
(a) $w_{RR} - df(\lambda)$ curves



(b) The dimensions are 'year made' and 'weight'

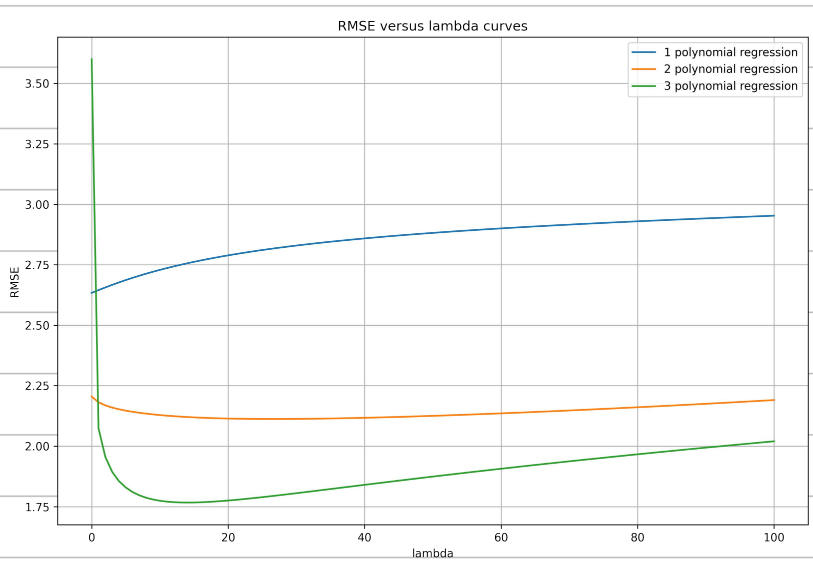
We can get that these two dimensions have significant influence on the prediction of y

(c)



This figure indicates that a smaller λ results in smaller RMSE. I prefer to choose least squares for this problem, and $\lambda = 0$

(d)



I prefer to choose $p=3$, because it results in smaller RMSE. In this model, obviously, ridge regression performs better than least squares. The ideal λ might be 14 instead of 0. I think λ would change with the model we choose, an underfitting model like linear regression may not advantage ridge regression and regularization.