Name: Xun Xue UNI: xx2241

(a) joint likelihood of the data $(x_1 \dots x_N)$

$$P(x_1 \dots x_N | \pi, r) = \prod_{i=1}^{N} \binom{x_i + r - 1}{x_i} \pi^{x_i} (1-\pi)^r$$

(b) $$\ln P(x_1 \dots x_N | \pi, r) = \sum_{i=1}^{N} \ln \binom{x_i + r - 1}{x_i} \pi^{x_i} (1-\pi)^r = \sum_{i=1}^{N} \left(\ln \binom{x_i + r - 1}{x_i} + x_i \ln \pi + r \ln(1-\pi) \right)$$

$$\nabla_\pi \ln P(x_1 \dots x_N | \pi, r) = \sum_{i=1}^{N} \left(\frac{x_i}{\pi} - \frac{r}{1-\pi} \right) = 0$$

$$\therefore \frac{\sum_{i=1}^{N} x_i (1-\pi) - nr\pi}{\pi(1-\pi)} = 0 \qquad \left(\sum_{i=1}^{N} x_i + nr \right) \pi = \sum_{i=1}^{N} x_i$$

$$\therefore \hat{\pi}_{ML} = \frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} x_i + Nr}$$

(c) from Bayes rule, we know that $P(\pi | x_1 \dots x_N) = \dfrac{P(x_1 \dots x_N | \pi) \cdot P(\pi)}{P(x_1 \dots x_N)}$

$$P(\pi | x_1 \dots x_N) = \frac{\prod_{i=1}^{N} \binom{x_i + r - 1}{x_i} \pi^{x_i} (1-\pi)^r \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \pi^{a-1} (1-\pi)^{b-1}}{P(x_1 \dots x_N)}$$

$$\nabla_\pi \ln P(\pi | x_1 \dots x_N) = \sum_{i=1}^{N} \left(\frac{x_i}{\pi} - \frac{r}{1-\pi} \right) + \frac{a-1}{\pi} + \frac{1-b}{1-\pi} = 0$$

$$\therefore \hat{\pi}_{MAP} = \frac{\sum_{i=1}^{N} x_i + a - 1}{\sum_{i=1}^{N} x_i + a + b + Nr - 2}$$

(d) from Bayes rule, we know that $P(\pi | x_1 \dots x_N) \propto P(x_1 \dots x_N | \pi) \cdot P(\pi)$

$$\therefore P(\pi | x_1 \dots x_N) \propto \prod_{i=1}^{N} \binom{x_i + r - 1}{x_i} \pi^{x_i} (1-\pi)^r \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \pi^{a-1} (1-\pi)^{b-1}$$

$$\propto \prod_{i=1}^{N} \binom{x_i + r - 1}{x_i} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \pi^{\sum_{i=1}^{N} x_i + a - 1} (1-\pi)^{Nr + b - 1}$$

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

$\therefore P(\pi | x_1 \dots x_N)$ can be identify as a Beta distribution, $\text{Beta}\left(\sum_{i=1}^{N} x_i + a, Nr + b \right)$

Name: Xun Xue UNI: xx2241

(e) ① As we known in Beta distribution:

Mean: $E(\pi) = \dfrac{\sum_{i=1}^{N} x_i + a}{\sum_{i=1}^{N} x_i + Nr + a + b}$

Variance: $Var(\pi) = \dfrac{(\sum_{i=1}^{N} x_i + a)(Nr+b)}{(\sum_{i=1}^{N} x_i + Nr + a + b)^2 (\sum_{i=1}^{N} x_i + Nr + a + b + 1)}$

② Discuss the relationship with $\hat{\pi}_{ML}$ and $\hat{\pi}_{MAP}$

As we can see in part (b) and part (c)

$$\hat{\pi}_{ML} = \frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} x_i + Nr} \qquad \hat{\pi}_{MAP} = \frac{\sum_{i=1}^{N} x_i + a - 1}{\sum_{i=1}^{N} x_i + a + b + Nr - 2}$$

when $a=1$ and $b=1$, $\hat{\pi}_{ML} = \hat{\pi}_{MAP}$. since in this case the prior distribution is a uniform distribution, so $P(\pi|x_1 \dots x_N) \propto P(x_1 \dots x_N|\pi)$ and $\nabla_\pi P(\pi|x_1 \dots x_N) = \nabla_\pi P(x_1 \dots x_N|\pi)$

Therefore $\hat{\pi}_{ML}$ can be regard to a special case for $\hat{\pi}_{MAP}$ when $P(\pi) = beta(1,1)$.

The posterior distribution of π is the Bayesian Inference. Comparing with point estimates such as $\hat{\pi}_{ML}$ or $\hat{\pi}_{MAP}$, it takes a step further by characterizing uncertainty about the values in π using Bayes rule.

Comparing $E(\pi)$ and $\hat{\pi}_{MAP}$. we noticed that when $N \to \infty$, $E(\pi) = \hat{\pi}_{MAP}$

$\hat{\pi}_{MAP}$ seeks the most probable value π according to its posterier distribution therefore $\hat{\pi}_{MAP}$ is equal to the mode of the posterior distribution of π, which is also $\dfrac{\sum_{i=1}^{N} x_i + a - 1}{\sum_{i=1}^{N} x_i + a + b + Nr - 2}$. when $N \to \infty$. the Beta distribution will be very sharp so the mode and mean will be the same value, therefore when $N \to \infty$, $E(\pi) = \hat{\pi}_{MAP}$

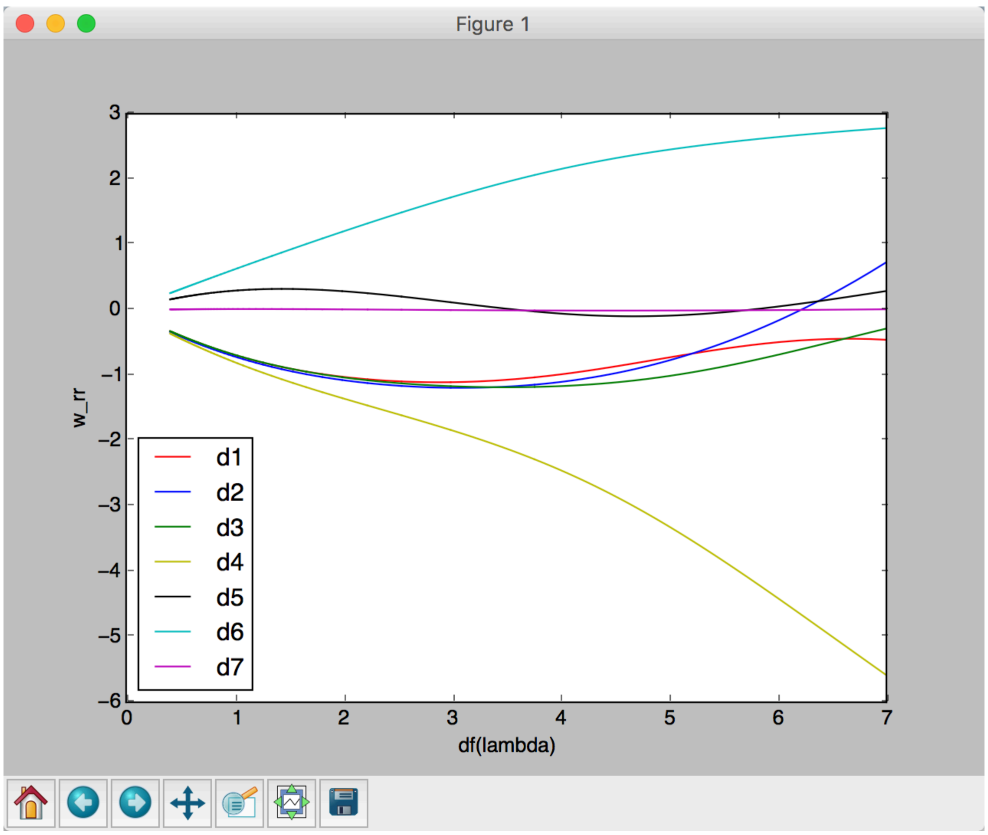
Answers of coding part
Name: Xun Xue
UNI: xx2241

Problem 2

Part 1

(a)

The plot of 7 values in w_{RR} as a function of df(λ)

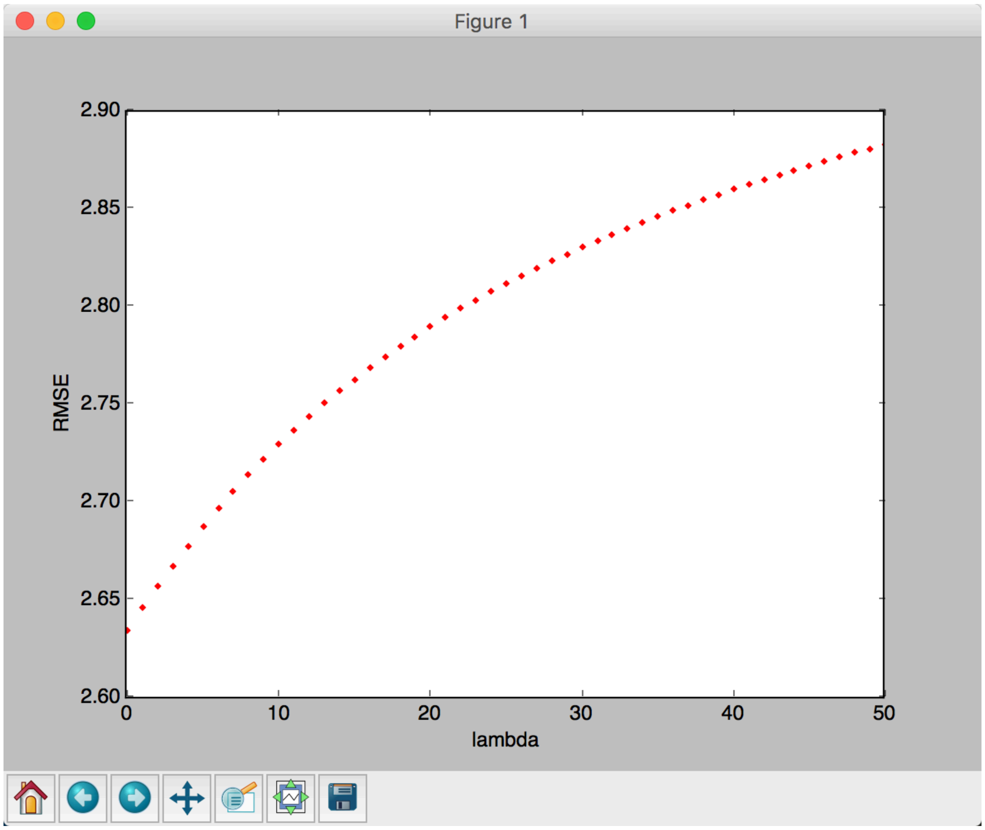


(b)

The 4th dimension (car weight) and the 6th dimension (car year) clearly stand out over the other dimensions. This indicates that these two features have a bigger influence on the result y (miles per gallon for the car). As the degree of freedom decrease, the constraint in these two features is obvious.

(c)

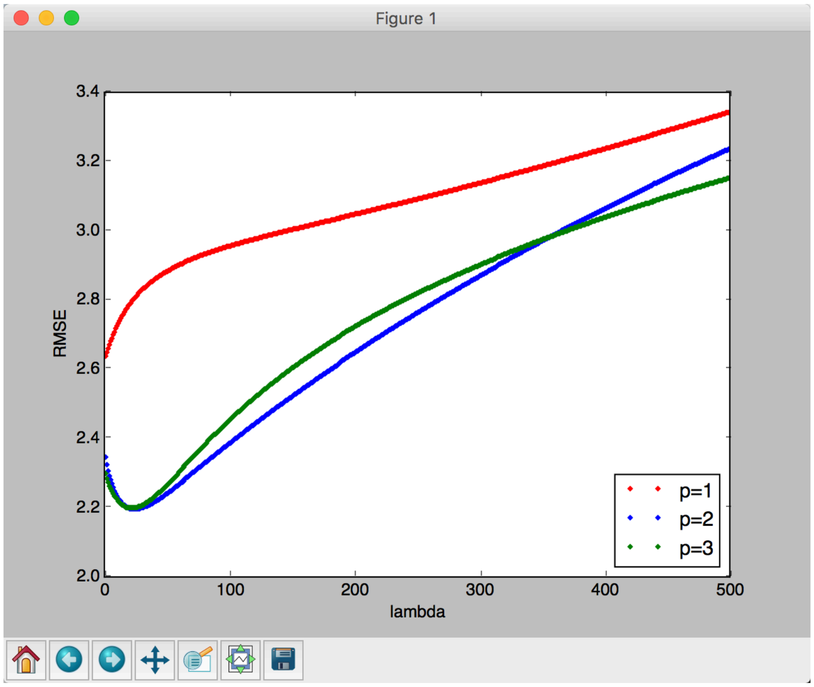
The plot of RMSE as a function of λ.



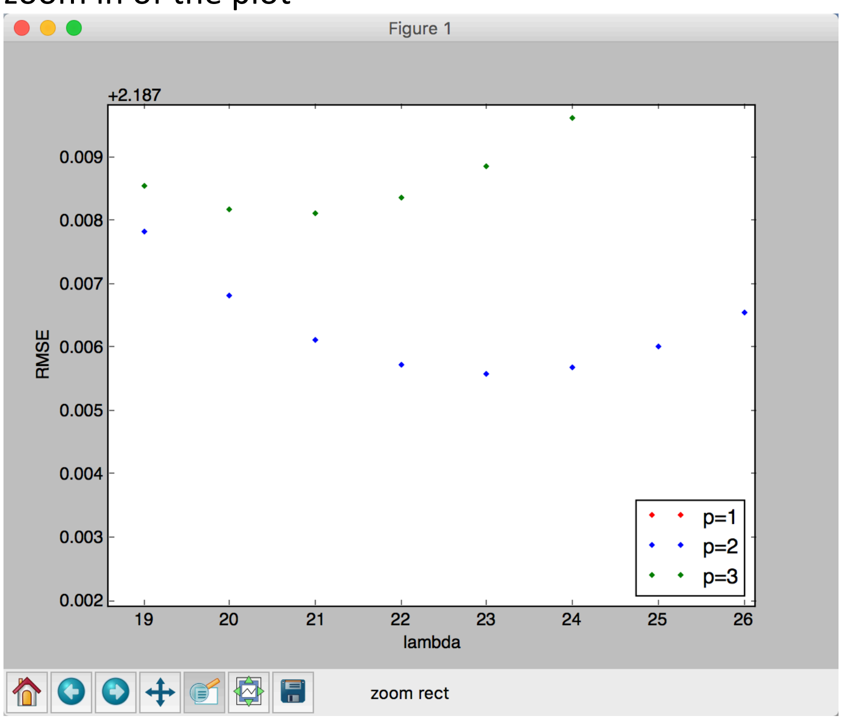
From this figure we can conclude that when λ grows larger the prediction will be worse. When λ=0, which is just the least square approach, has the smallest RMSE. Therefore, for this problem least square is a better choice than ridge regression.

(d)

The plot of RMSE as a function of λ when p=1, p=2 and p=3.



zoom in of the plot



As we can see from the zoom-in of plot, it will reach the minimum of RMSE when p=2 and λ=23. Since from the plot we can see the minimum of p=2 and p=3 is very similar, so choosing p=2 and p=3 all make sense.

For this problem, when p=1, the RMSE value increase with the increase of λ. When p=2 and p=3, the RMSE value decrease at first and then increase with the increase of λ.

The ideal value of λ is 0 when p=1. The ideal value of λ is 23 when p=2. The ideal value of λ is 21 when p=3。