Problem 1

Problem 1

$p(x_i; \eta \mid \pi, r) = \binom{x_i+r-1}{x_i} \pi^{x_i} (1-\pi)^r$

(a) Joint Likelihood

$L = \prod_{i=1}^{n} p(x_i) = \binom{x_1+r-1}{x_1} \pi^{x_1} (1-\pi)^r \cdot$

$= \binom{x_2+r-1}{x_2} \pi^{x_2} (1-\pi)^r \cdot \binom{x_3+r-1}{x_3} \pi^{x_3} (1-\pi)^r \cdots \binom{x_n+r-1}{x_n} \pi^{x_n} (1-\pi)^r$

$= \boxed{\binom{x_1+r-1}{x_1} \binom{x_2+r-1}{x_2} \cdots \binom{x_n+r-1}{x_n} \pi^{\sum x_i} (1-\pi)^{nr}}$

(b) $\hat{\pi}_{ML} = \underset{\pi}{\operatorname{argmax}}\, L$

$\underset{\pi}{\operatorname{argmax}}\, C \pi^{\sum x_i} (1-\pi)^{nr} = \underset{\pi}{\operatorname{argmax}} \ln \left[C \pi^{\sum x_i} (1-\pi)^{nr} \right]$

$l = \ln C + \sum x_i \ln \pi + nr \ln(1-\pi)$

$\frac{\partial l}{\partial \pi} = \frac{\sum x_i}{\pi} - \frac{nr}{1-\pi} = 0$

$\Rightarrow \sum x_i (1-\pi) - nr\pi = 0$

$\Rightarrow \sum x_i - \pi \sum x_i = \pi n r$

$\Rightarrow \sum x_i = \pi (\sum x_i + nr)$

$\Rightarrow \boxed{\hat{\pi}_{ML} = \frac{\sum x_i}{\sum x_i + nr} = \frac{\bar{X}}{\bar{X}+r}}$ $\quad \bar{X} = \frac{\sum x_i}{n}$

c) Prior $p(\pi) = \text{beta}(a, b)$ Pg 2

derive MAP estimate $\hat{\pi}_{MAP}$ for π.

Find $p(\pi \mid x_1, x_2 \ldots x_n, r)$ Use

$$\boxed{P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)} = \frac{P(A|B) \cdot P(B)}{\sum_B P(A|B) \cdot P(B)}} \to \text{Baye's rule}$$

$$p(\pi \mid x_1, x_2 \ldots x_n, r) = \frac{P(x_1, x_2 \ldots x_n \mid \pi, r) \cdot p(\pi)}{\int_\pi p(x_1, x_2 \ldots x_n \mid \pi, r) \cdot p(\pi) \, d\pi}$$

$$\propto p(x_1, x_2 \ldots x_n \mid \pi, r) \cdot p(\pi)$$

$$\propto \prod_{i=1}^{n} \left[\binom{x_i + r - 1}{x_i} \pi^{x_i} (1-\pi)^{r} \right] \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \pi^{a-1} (1-\pi)^{b-1}$$

ignore the constant term.

$$\Rightarrow \propto \prod_{i=1}^{n} \left\{ \binom{x_i + r - 1}{x_i} \right\} \underbrace{\pi^{\sum x_i + a - 1} (1-\pi)^{nr + b - 1}}_{\sim \text{beta}(\sum^n x_i + a,\ nr + b)}$$

Now, we want $\hat{\pi}_{MAP}$

$$= \arg\max_\pi \binom{x_1 + r - 1}{x_1} \binom{x_2 + r - 1}{x_2} \cdots \binom{x_n + r - 1}{x_n} \pi^{\sum x_i + a - 1} (1-\pi)^{nr + b - 1}$$

$$= \arg\max_\pi \sum_{i=1}^{n} \ln \binom{x_i + r - 1}{x_i} + \left(\sum x_i + a - 1\right) \ln \pi + (nr + b - 1) \ln(1-\pi)$$

$$\frac{\partial}{\partial \pi} = \frac{\sum x_i + a - 1}{\pi} - \frac{(nr + b - 1)}{1 - \pi} = 0$$

$$\Rightarrow \left(\sum x_i + a - 1\right)(1 - \pi) = (nr + b - 1)\pi$$

Pg 3

$$\Rightarrow \left(\sum x_i + a - 1\right) = \pi\left[(nr+b-1) + \left(\sum x_i + a - 1\right)\right]$$

$$\Rightarrow \sum x_i + a - 1 = \pi\left[\sum x_i + nr + a + b - 2\right]$$

$$\Rightarrow \boxed{\hat{\pi}_{MAP} = \frac{\sum x_i + a - 1}{\sum x_i + nr + a + b - 2}}$$

d) Using baye's rule in previous step, we have already identified the kernel of the posterior distribution.

Posterior distribution is beta $(\underbrace{\sum x_i + a}_{\to a}, \underbrace{nr + b}_{\to b})$

e) $\mathbb{E}(\pi) = \frac{a}{a+b} \Rightarrow \frac{\sum x_i + a}{\sum x_i + a + nr + b}$

$Var(\pi) = \frac{ab}{(a+b)^2(a+b+1)} = \frac{\left(\sum x_i + a\right)(nr+b)}{\left(\sum x_i + a + nr + b\right)^2\left(\sum x_i + a + nr + b + 1\right)}$

We observe the following:

- prior: $\pi \sim beta(a, b)$; posterior: $\pi \sim beta\left(\sum x_i + a, nr + b\right)$

  So there is Conjugacy, the posterior is in the same family. We augment the parameters a and b with $\sum x_i$ and nr respectively.

- $\hat{\pi}_{MAP} = \frac{\sum x_i + a - 1}{\sum x_i + nr + a + b - 2} \neq \mathbb{E}(\pi_{posterior}) = \frac{\sum x_i + a}{\sum x_i + a + nr + b}$

- As $n \to \infty$, $\sum x_i \to \infty$ (since $x_i \geq 0$), $\Rightarrow Var(\pi) \to 0$

Pg 4

- $\hat{\pi}_{MAP} = \dfrac{\sum x_i + a - 1}{\sum x_i + nr + a + b - 2}$; $\hat{\pi}_{ML} = \dfrac{\sum x_i}{\sum x_i + nr}$

$\hat{\pi}_{MAP} = \hat{\pi}_{ML}$ when $a = b = 1$

and $\pi \sim bet(a,b) \Rightarrow \pi \sim Uniform\ [0,1]$ when $a=b=1$

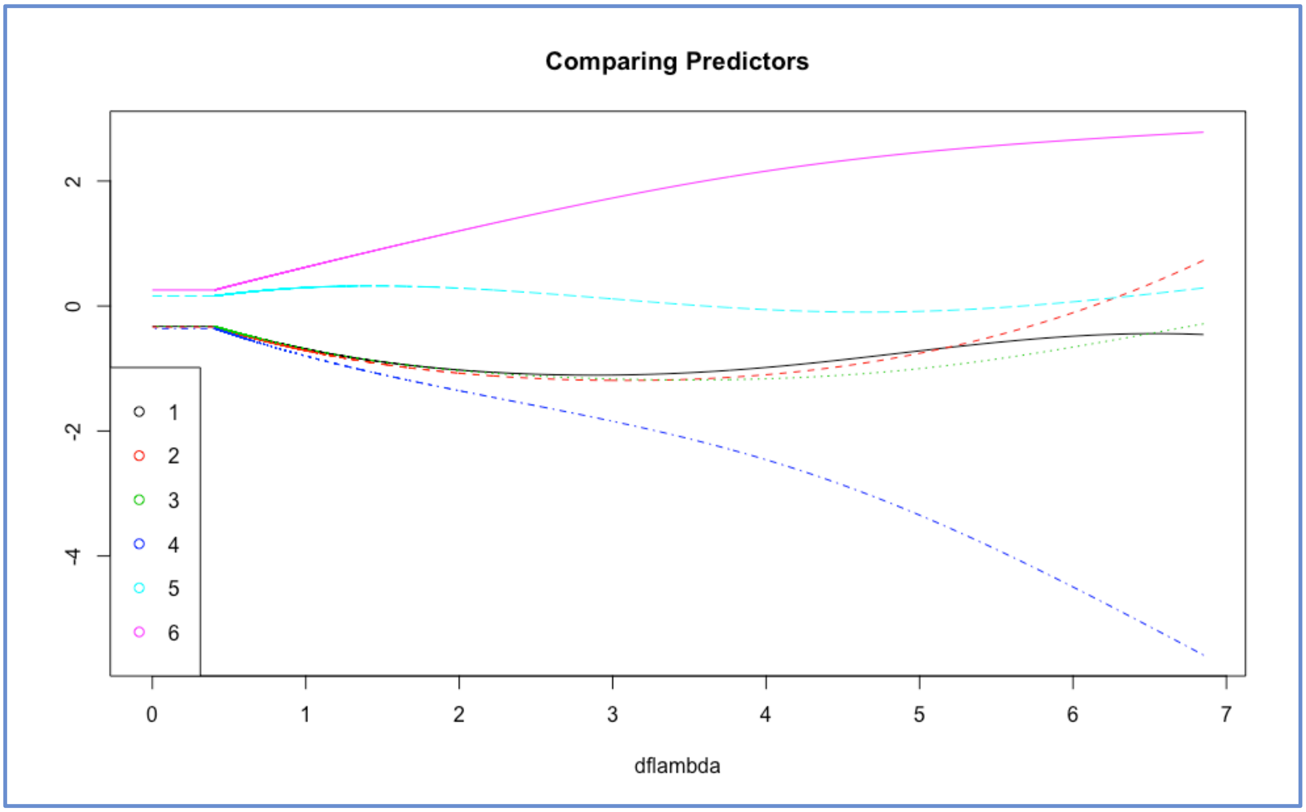
when we believe that our prior distribution of π is $U[0,1]$, our estimates for π using ML and MAP are the same

Problem 2

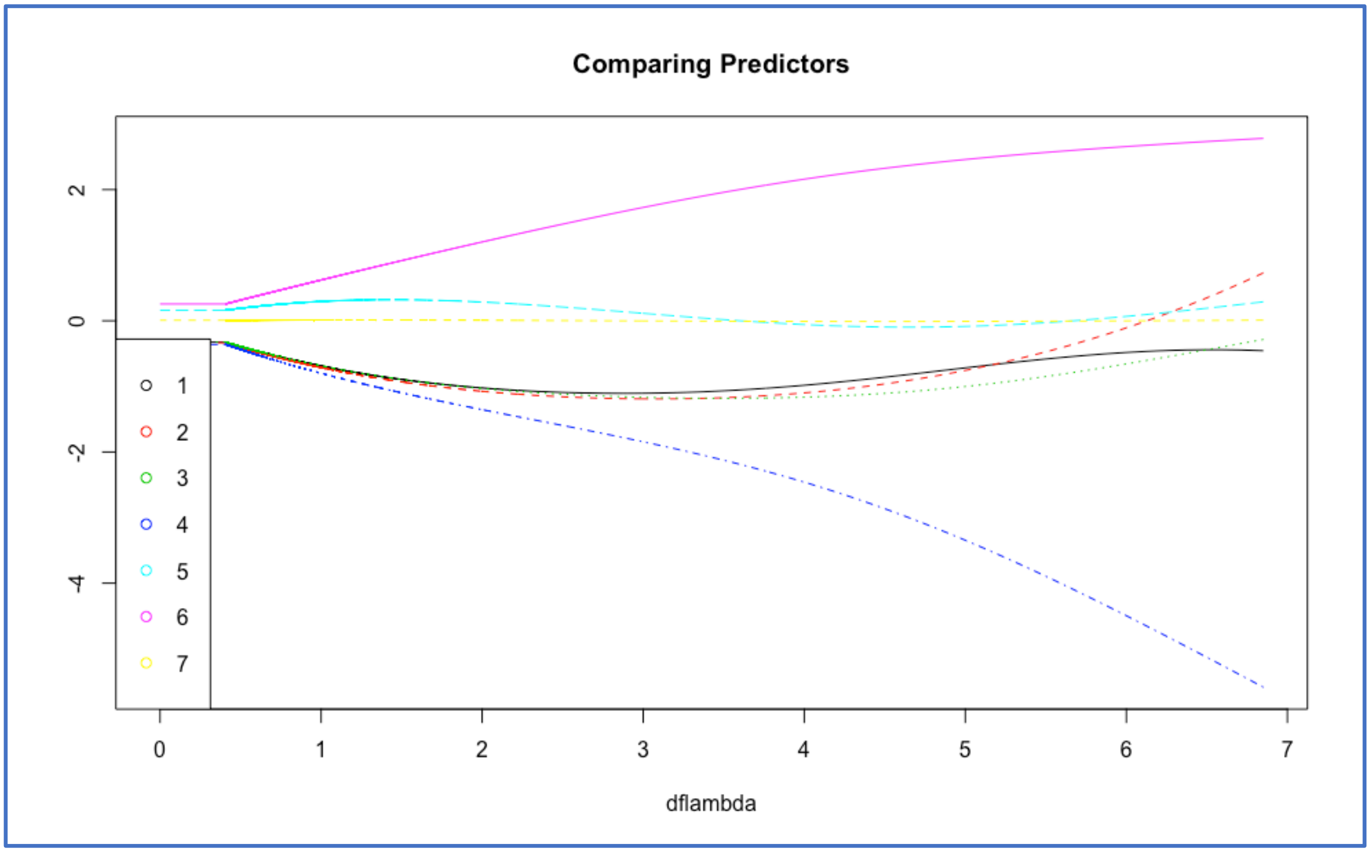
Part 1

a)

The following plot does NOT include the weight for the intercept



The next plot includes 7 predictors, including one for the intercept:



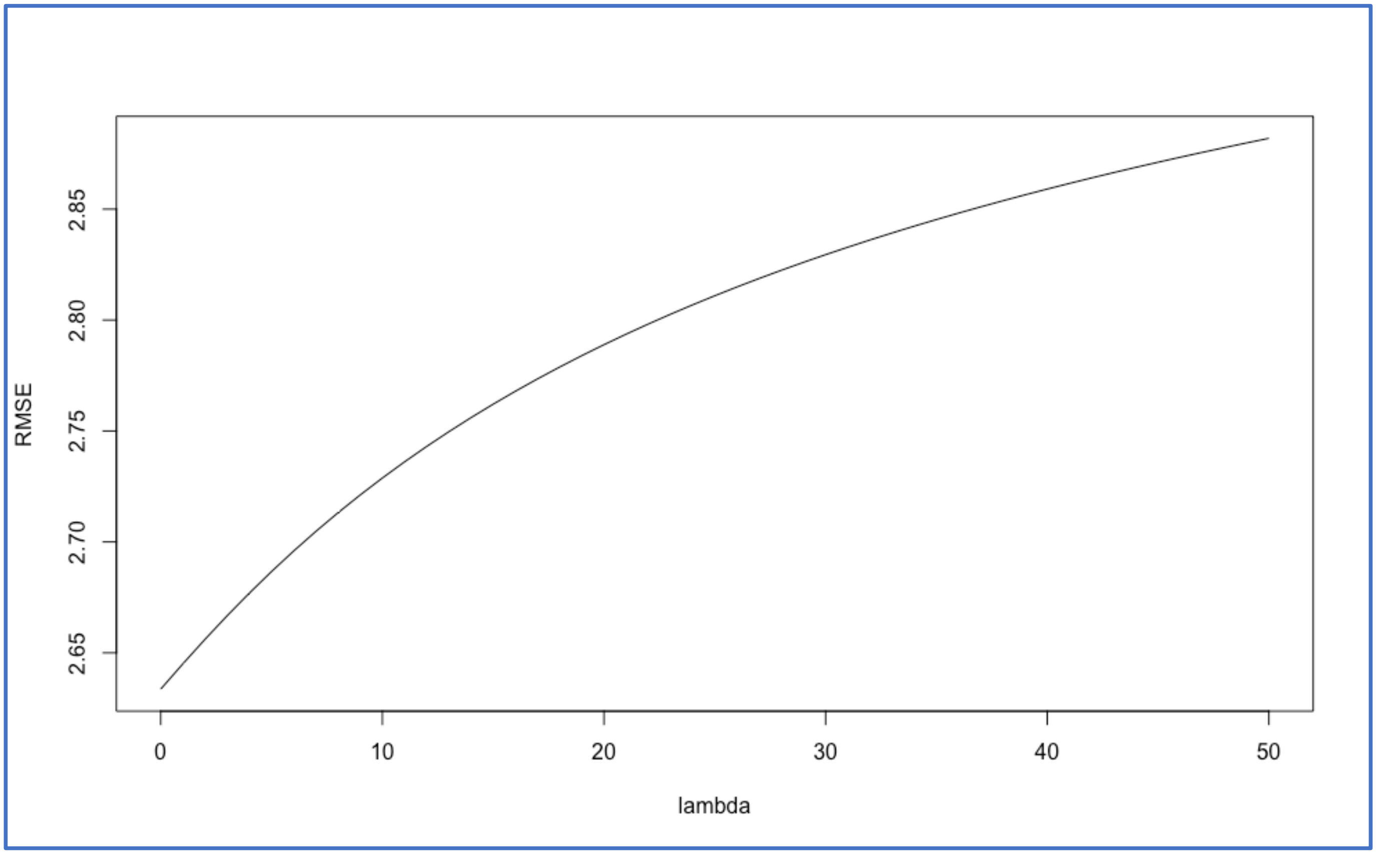
Part 1
b)
As lambda converges to 5000, all the coefficients converge to 0. However, we see that even for large values of lambda, the absolute value of the 4^{th} and 6^{th} predictors is farthest from 0. This means that these two predictors help predict the dependent variable, miles per gallon, the best.

The 4^{th} predictor is weight
The 6^{th} predictor is model year

According to this plot, weight has a negative relationship with mpg. The higher the weight of a car, the lower its mileage.

Model year has a positive relationship with mpg. The newer the car, the higher its mpg.

Part 1
c)



From this graph, we can infer the following:

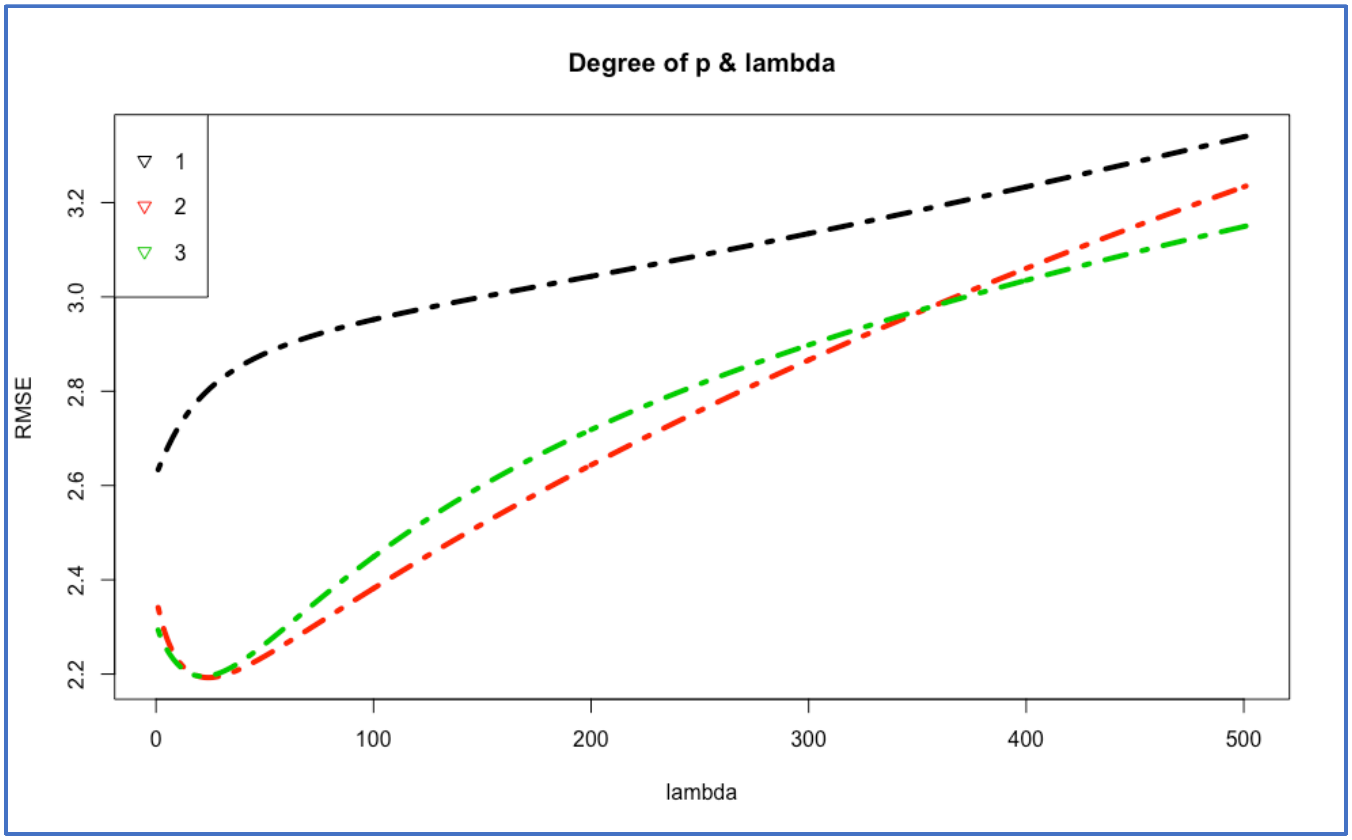
1. The optimal value of lambda is lambda = 0

2. lambda=0 corresponds to the least squares solution. Thus, least squares is a better fit for this data set, as compared to ridge regression.

3. When lambda>0, our estimates for the true parameters are no longer unbiased. However, although bias increases, ridge regression reduces the variance in the model. From the above plot, we infer that the reduction in the variance is not greater than the increase in the bias. This is why RMSE is increasing as lambda increases.

4. This makes sense intuitively. There are only 7 predictors (including intercept) in this data set. And n= 350, which is very large. So n>>p. Hence, ridge does not achieve a big decrease in the variance which is low to start off with. This is why OLS is a better estimator.

Part 2
d)



The numbers 1, 2 and 3 in the legend denote the degree of the polynomial used in the fitted model.

a) We see that p=1 yields much higher RMSE for any lambda

b) We see that p=2 and p=3 yield very similar RMSE values

It seems like we could choose either p=2 or 3, since they both yield similar results.

We should go ahead **and choose p=2** because:

- we see that it outperforms p=3 for low values of lambda.

- Also, increasing predictors in the regression will increase the variance. So unless there is a significant reduction in the bias, we should stick to the simpler model. Hence, we should choose p=2

The ideal value of lambda is **lambda = 23**. We found this using the following code:

```
> which.min(testErrorMat[,2])
[1] 23
```