1. Ridge regression Problem 3.41 $\hat{\beta_0} = \frac{1}{N} y_i = \overline{y}$.

$$\hat{\beta} = \arg\min \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}.$$

$$= \arg\min \{ \vec{Y}^T - \vec{\beta}^T \vec{X}^T)(\vec{Y} - \vec{X}\vec{\beta})$$

where $\vec{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ $\vec{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$ $\vec{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$

$$f(\vec{\beta}) = \vec{Y}^T\vec{Y} - \vec{Y}^T\vec{X}\vec{\beta} - \vec{\beta}^T\vec{X}^T\vec{Y} + \vec{\beta}^T\vec{X}^T\vec{X}\vec{\beta} + \lambda\vec{\beta}^T\vec{\beta}$$

$$\frac{\partial f}{\partial \beta} = -\vec{X}^T\vec{Y} - \vec{X}^T\vec{Y} + 2\vec{X}^T\vec{X}\vec{\beta} + 2\lambda\vec{\beta} = 0.$$

$$-\vec{X}\vec{Y} + \vec{X}^T\vec{X}\beta + \lambda\vec{\beta} = 0 \Rightarrow \vec{\beta} = (\vec{X}^T\vec{X} + \lambda I)^{-1}\vec{X}^T\vec{Y}$$

$$\hat{\beta^c} = \arg\min \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} (x_{ij} - \overline{x}_j)\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

$$\hat{\beta_0^c} = \frac{1}{N} y_i = \overline{y} \Rightarrow \hat{\beta^c} = \arg\min \{ (\vec{Y}^T - \vec{\beta}^{cT}\vec{X}^{cT})(\vec{Y} - \vec{X}^c\vec{\beta}^c) + \lambda \sum_{j=1}^{p} \beta_j^{c2} \}$$

where $\vec{X}^c = \begin{bmatrix} x_{11} - \overline{x}_1 & x_{12} - \overline{x}_2 & \cdots & x_{1p} - \overline{x}_p \\ x_{21} - \overline{x}_1 & x_{22} - \overline{x}_2 & & x_{2p} - \overline{x}_p \\ x_{31} - \overline{x}_1 & \vdots & & \vdots \\ \vdots & & & \\ x_{n1} - \overline{x}_1 & x_{n2} - \overline{x}_2 & & x_{np} - \overline{x}_p \end{bmatrix}.$ $\vec{\beta}^c = \begin{bmatrix} \beta_1^c \\ \beta_2^c \\ \vdots \\ \beta_p^c \end{bmatrix}$

$$\Rightarrow \vec{\beta}^c = (\vec{X}^{cT}\vec{X}^c + \lambda I)^{-1}\vec{X}^{cT}\vec{Y}.$$

2. $\beta_i \sim N(0, J)$.

$$P(\beta_j \mid Y) = \frac{P(Y \mid \beta_j)\, P(\beta_j)}{P(Y)} \qquad P(Y \mid \beta) = P(Y).$$

$$P(\beta \mid Y) = \frac{P(Y \mid \beta)\, P(\beta)}{P(Y)}$$

Since y_i are independent

$$\sim \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(y_1-\beta_0-X_1^T\beta)^2}{2\sigma^2}} \cdots \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(y_N-\beta_0-X_N^T\beta)}{2\sigma^2}} \frac{1}{\sqrt{2\pi J^2}}\, e^{-\frac{\beta^2}{2J^2}} \cdots e^{-\frac{\beta_p^2}{2J^2}}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N e^{-\frac{(y_1-\beta_0-X_1^T\beta)^2}{2\sigma^2} - \frac{(y_2-\beta_0-X_2^T\beta)^2}{2\sigma^2} - \cdots \frac{(y_N-\beta_0-X_N^T\beta)^2}{2\sigma^2}} \left(\frac{1}{\sqrt{2\pi J^2}}\right)^p e^{-\frac{\beta_1^2+\beta_2^2+\beta_p^2}{2J^2}}$$

$$\log P(\beta \mid Y) = \underline{-N\log\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{N}(y_i-\beta_0-X_i^T\beta)^2\right] \underline{-p\log\sqrt{\pi J^2}} - \frac{1}{2J^2}\left[\beta_1^2+\beta_2^2+\cdots\beta_p^2\right]$$

$$= C - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\beta_0-X_i^T\beta)^2 - \frac{1}{2J^2}[\beta_1^2+\cdots\beta_p^2]$$

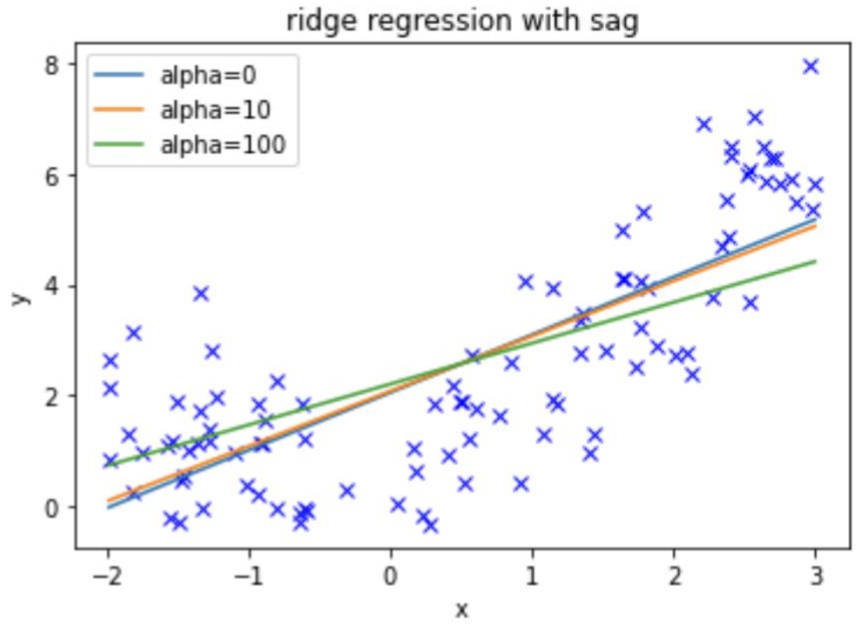
$$= C_1 - \sum_{i=1}^{N}\Big(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\Big)^2 - \frac{\sigma^2}{J^2}\sum_{j=1}^{p}\beta_j^2$$

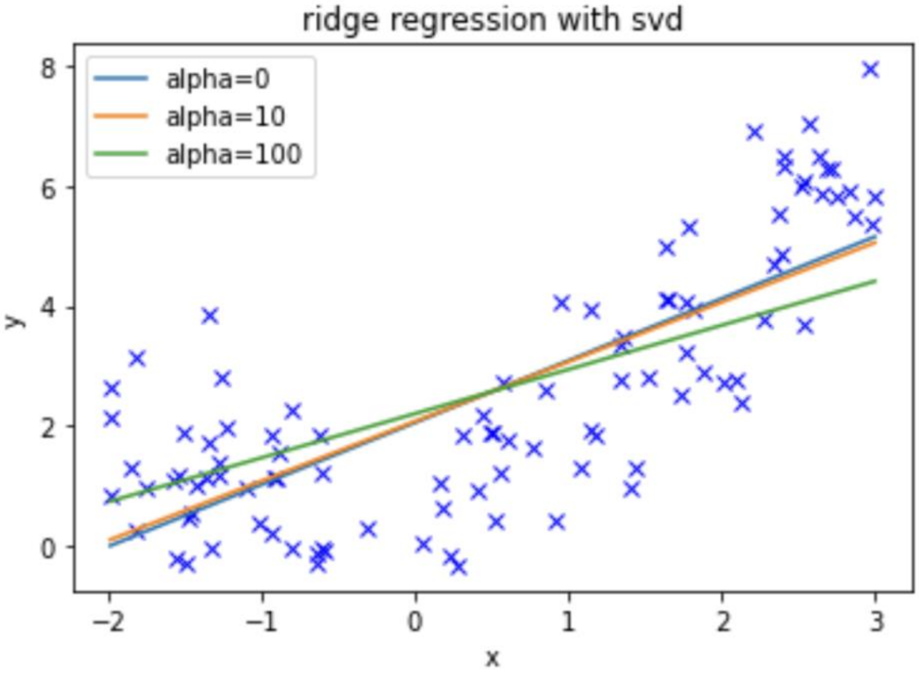
$$-\log P(\beta \mid Y) \sim C_2 + \sum_{i=1}^{N}\Big(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\Big)^2 - \underset{\uparrow}{\frac{\sigma^2}{J^2}}\sum_{j=1}^{p}\beta_j^2.$$

$$\lambda = \frac{\sigma^2}{J^2}.$$

3. a)

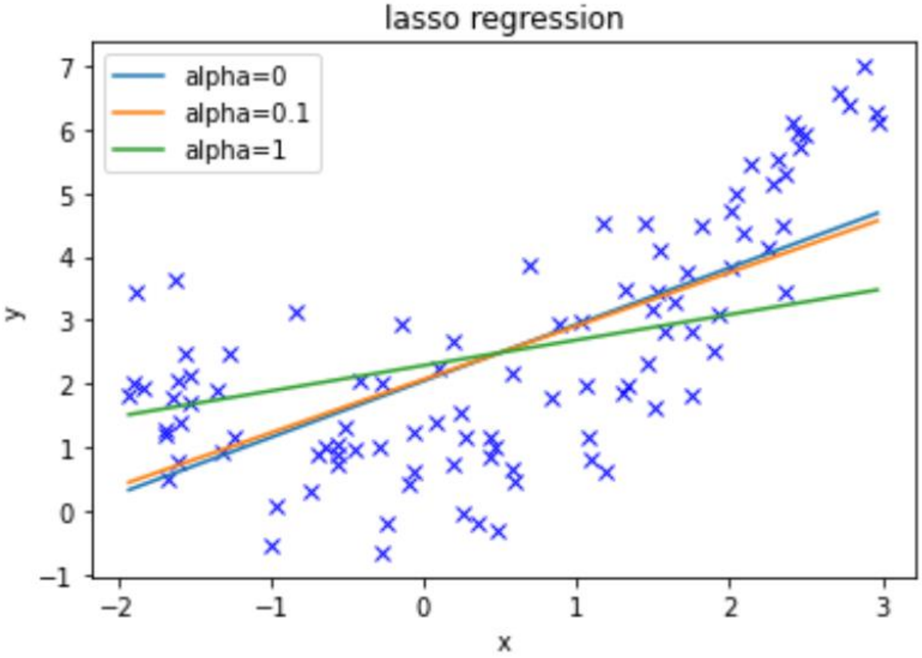
```
plt.show()
```





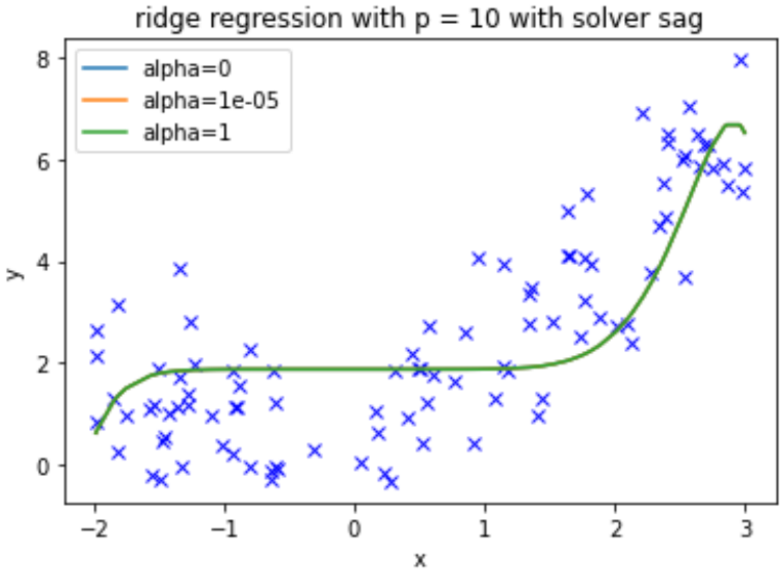
Sag and svd give almost the same result. When alpha is higher, the regression is flatter.

b)

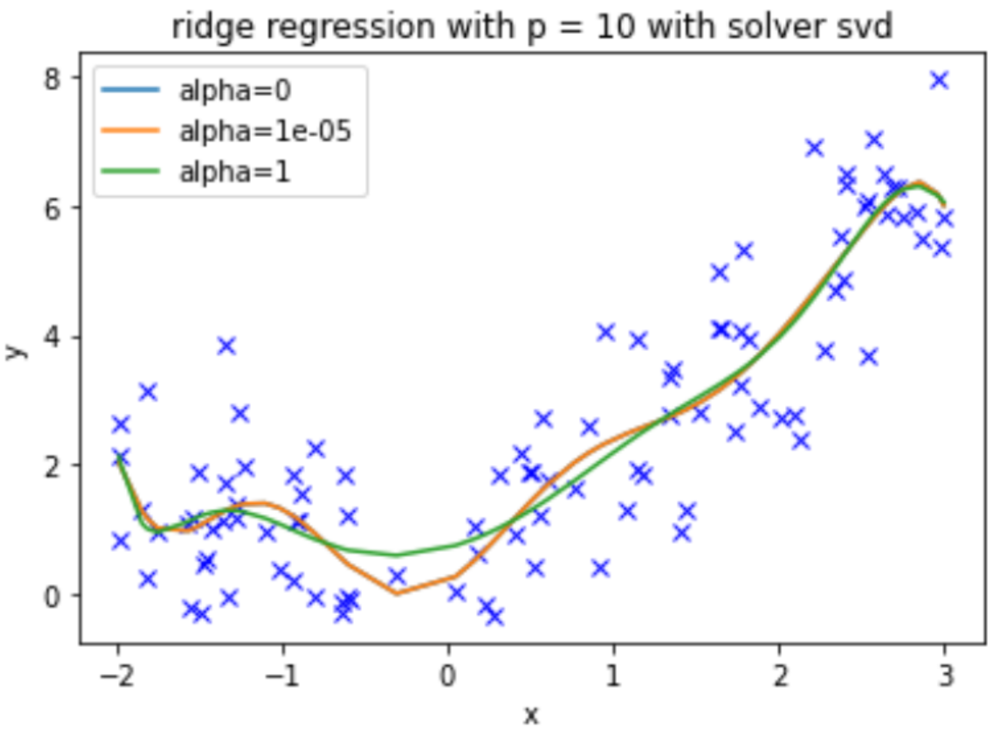


Similar. When alpha is higher, the regression is flatter.
When we have a less constrain (less alpha) in Lasso regression, the regression is more flatter.

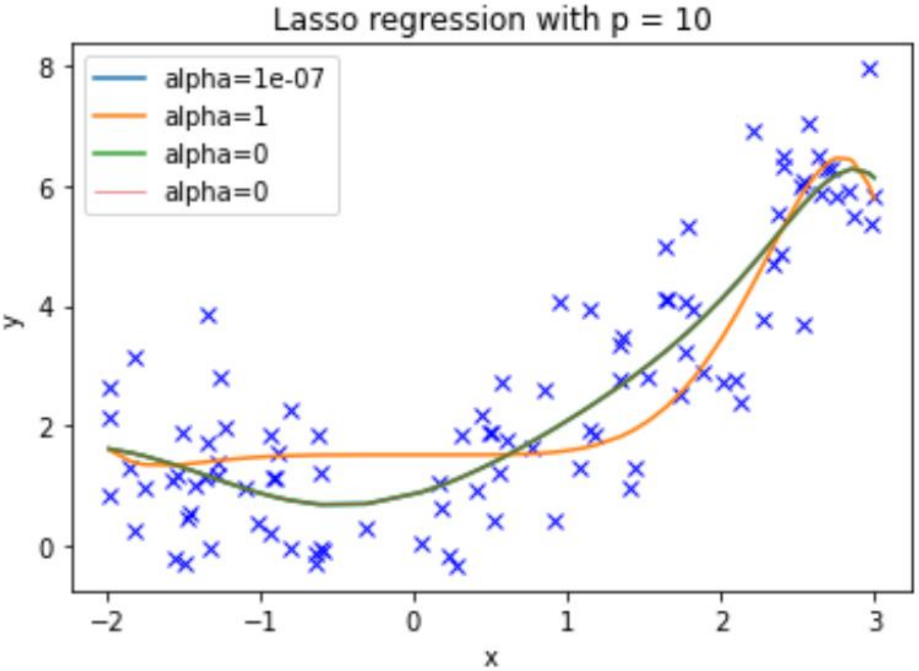
c)



The constrains are too small and can’t vary the result that much. So we have almost similar three lines.

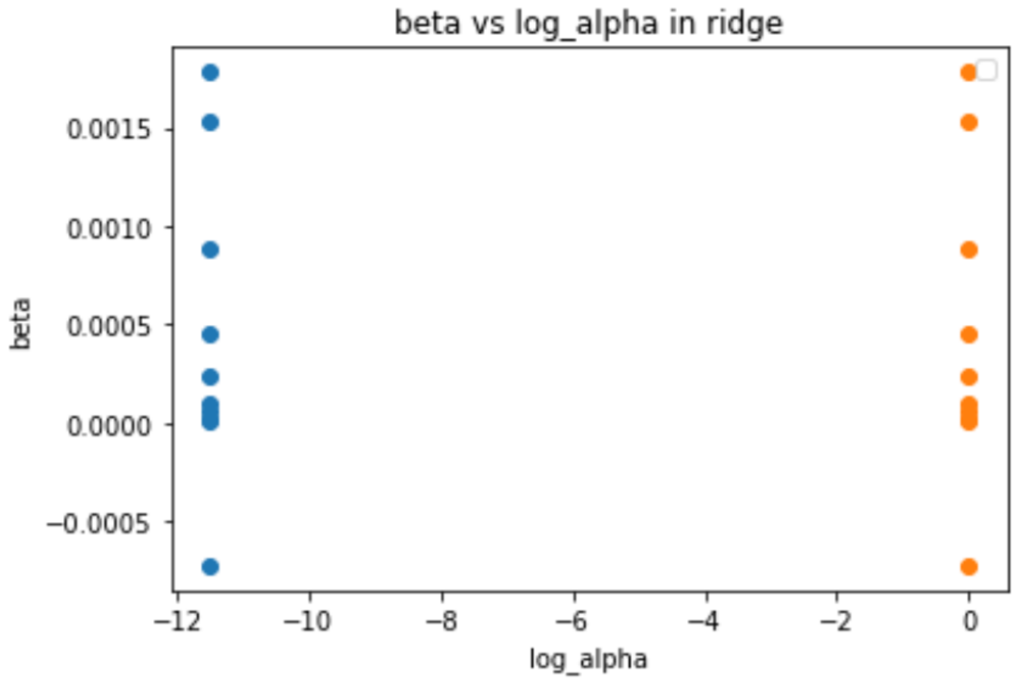

ridge regression with p = 10 with solver svd
alpha=0
alpha=1e-05
alpha=1
y
8
6
4
2
0
−2
−1
0
1
2
3
x

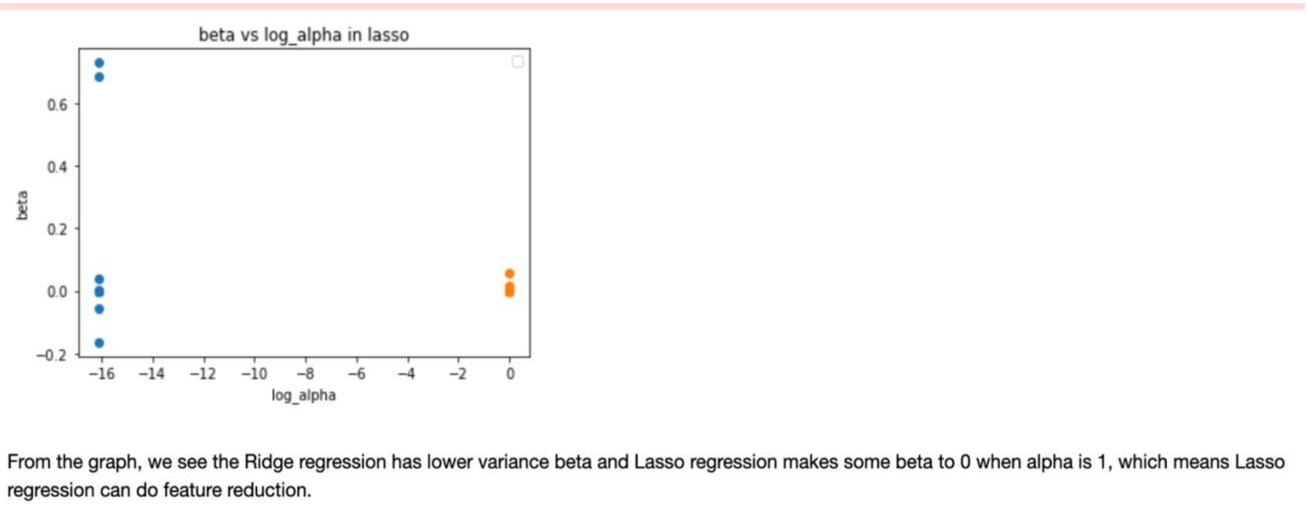
d)


Lasso regression with p = 10
alpha=1e-07
alpha=1
alpha=0
alpha=0
y
8
6
4
2
0
−2
−1
0
1
2
3
x

Compared to ridge regression, the lasso regression has more variance. And when we vary alpha, the regression result changes more. When alpha is one, lasso regression is flatter than ridge regression at the beginning.

e)





From the graph, we see the Ridge regression has lower variance beta and Lasso regression makes some beta to 0 when alpha is 1, which means Lasso regression can do feature reduction.

And ridge regression does shrink the coefficients.