ISLR Questions:

3) ISLR 6.8.3

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right) \text{ subject to } \sum_{j=1}^{p}|\beta_j| \leq s$$

a) iv) is TRUE as 's' is increased, there is a less constraint on the model and it should always have a better training error (if 's' is increased, then the best model using a budget of 's' would be included when using a budget of 's').

if we consider $p=2$ for predictors, so the subject to constraint is $|\beta_1| + |\beta_2| \leq s$

when $s=0$ then $|\beta_1| + |\beta_2| \leq 0$ i.e. β_1 & β_2 are 0 & the model's training RSS will be at its maximum. As 's' increases from 0, the co-efficients (β_1 & β_2) will start to approach their least squares estimates values causing the training RSS to steadily decrease to the ordinary least square RSS. Hence here iv is TRUE.

b) ii) is TRUE - test error will improve (decrease) to a point and then will worsen (increase) as constraints loosen and model overfits. If we see the same constraint then the coefficients are 0 when $s=0$, the model has a high test RSS. As we increase s from 0, the coefficients will assume non-zero values causing the

model to start to fit well to test data decreasing test RSS. However, there will reach a point when the coefficients approach ordinary least squares values and the model will start overfitting to the training data, which will increase test RSS. Therefore, as we increase s from 0, the test RSS will decrease initially and then eventually start increasing in a U- shape.

c) iii) is TRUE as variance always increases with fewer constraints. when s=0, the model basically predicts a constant leading to virtually no variance. As 's' increases from 0, the model starts including more coefficients and their values become highly dependent on training data. Therefore, as 's' increases from 0, the variance steadily increases.

d) iv) is TRUE - bias always decreases with more model flexibility. when s=0, the model is basically predicting a constant which is far from the actual value, thus the squared bias is high. As s increases, the coefficients become non-zero and fit the training data better which causes bias to decrease. Therefore, as 's' increases from 0 the squared bias steadily decreases.

c) v) is TRUE as the irreducible error is a constant value, not related to model selection. Since the irreducible error is model independent, the choice of 's' does not affect the irreducible error associated with a model. Therefore, as 's' increases from 0, the irreducible error remains constant.

4) ISLR , 6.8.5

Ridge Regression : Minimize $\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij}\right)^2 + \lambda\sum_{i=1}^{p}\hat{\beta}_i^2$

a) By substituting $\hat{\beta}_0 = 0$ and $n = p = 2$ specified in the problem into the general form of Ridge Regression shown above, we get

$$(y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12})^2 + (y_2 - \hat{\beta}_1 x_{21} - \hat{\beta}_2 x_{22})^2 + \lambda(\hat{\beta}_1^2 + \hat{\beta}_2^2)$$

b) Expanding the equation from part a.

$$(y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12})^2 + (y_2 - \hat{\beta}_1 x_{21} - \hat{\beta}_2 x_{22})^2 + \lambda(\hat{\beta}_1^2 + \hat{\beta}_2^2)$$
$$= (y_1^2 + \hat{\beta}_1^2 x_{11}^2 + \hat{\beta}_2^2 x_{12}^2 - 2y_1\hat{\beta}_1 x_{11} + 2\hat{\beta}_1\hat{\beta}_2 x_{11}x_{12} - 2y_1\hat{\beta}_2 x_{12}$$
$$+ (y_2^2 + \hat{\beta}_1^2 x_{21}^2 + \hat{\beta}_2^2 x_{22} - 2y_2\hat{\beta}_1 x_{21} + -2\hat{\beta}_2 x_{22} y_2 + 2\hat{\beta}_1\hat{\beta}_2 x_{21}x_{22}$$
$$+ \lambda\hat{\beta}_1^2 + \lambda\hat{\beta}_2^2$$

we are given that : $x_{11} = x_{12}$, $x_{21} = x_{22}$, $y_1 + y_2 = 0$, $x_{11} + x_{21} = 0$ and $x_{12} + x_{22} = 0$

let $x_{11} = x_{12} = x_1$ & $x_{21} = x_{22} = x_2$

substituting values in equation:

$$= Y_1^2 - 2Y_1X_1(\hat{\beta}_1 + \hat{\beta}_2) + X_1^2(\hat{\beta}_1 + \hat{\beta}_2)^2 + Y_2^2 - 2Y_2X_2(\hat{\beta}_1 + \hat{\beta}_2) + X_2^2(\hat{\beta}_1 + \hat{\beta}_2)^2 + \lambda(\hat{\beta}_1^2 + \hat{\beta}_2^2) \longrightarrow \text{equation } (1)$$

taking the partial derivative to $\hat{\beta}_1$ and setting equation to 0 to minimize:

$$-2Y_1X_1 + X_1^2(2\hat{\beta}_1 + 2\hat{\beta}_2) - 2Y_2X_2 + X_2^2(2\hat{\beta}_1 + 2\hat{\beta}_2) + \lambda 2\hat{\beta}_1 = 0$$

$$-2Y_1X_1 + 2X_1^2(\hat{\beta}_1 + \hat{\beta}_2) - 2Y_2X_2 + 2X_2^2(\hat{\beta}_1 + \hat{\beta}_2) + 2\lambda\hat{\beta}_1 = 0$$

dividing equation by 2

$$-Y_1X_1 + X_1^2(\hat{\beta}_1 + \hat{\beta}_2) - Y_2X_2 + X_2^2(\hat{\beta}_1 + \hat{\beta}_2) + \lambda\hat{\beta}_1 = 0$$

$$X_1Y_1 - X_1^2\hat{\beta}_2 + X_2Y_2 + X_2^2\hat{\beta}_2 = X_1^2\hat{\beta}_1 + X_2^2\hat{\beta}_1 + \lambda\hat{\beta}_1$$

$$\hat{\beta}_1(X_1^2 + X_2^2 + \lambda) = X_1Y_1 + X_2Y_2 - \hat{\beta}_2(X_1^2 + X_2^2)$$

$$\beta_1 = \frac{X_1Y_1 + X_2Y_2 - \hat{\beta}_2(X_1^2 + X_2^2)}{X_1^2 + X_2^2 + \lambda} \longrightarrow (2)$$

taking derivative of equation (1) with respect to $\hat{\beta}_2$ and setting it to 0.

$$-2Y_1X_1 + 2X_1^2(\hat{\beta}_1 + \hat{\beta}_2) - 2Y_2X_2 + 2X_2^2(\hat{\beta}_1 + \hat{\beta}_2) + 2\lambda\hat{\beta}_2 = 0$$

$$X_1 Y_1 - X_1^2 \beta_1 + X_2 Y_2 - X_2^2 \hat{\beta}_1 = \lambda \hat{\beta}_2 + X_1^2 \hat{\beta}_2 + X_2^2 \hat{\beta}_2$$

$$\hat{\beta}_2 (\lambda + X_1^2 + X_2^2) = X_1 Y_1 + X_2 Y_2 - \hat{\beta}_1 (X_1^2 + X_2^2)$$

$$\hat{\beta}_2 = \frac{X_1 Y_1 + X_2 Y_2 - \hat{\beta}_1 (X_1^2 + X_2^2)}{\lambda + X_1^2 + X_2^2} \longrightarrow (3)$$

Since equation (2) = equation (3) we can say that $\hat{\beta}_1 = \hat{\beta}_2$

c) Lasso : ~~Minimize~~ ~~$(Y_1 - \hat{\beta}_1 X_{11})^2 + (Y_2 - \hat{\beta}_1 X_{21})$~~

Similar to Ridge Regression, we can write the optimization problem for this case as:

Minimize : $(Y_1 - \hat{\beta}_1 X_{11} - \hat{\beta}_2 X_{12})^2 + (Y_2 - \hat{\beta}_1 X_{11} - \hat{\beta}_2 X_{12})^2 + \lambda(|\hat{\beta}_1| + |\hat{\beta}_2|$

d) Consider the lasso constraints as:

i) $|\hat{\beta}_1| + |\hat{\beta}_2| \leq s \longrightarrow (1)$

ii) $(Y_1 - \hat{\beta}_1 X_{11} - \hat{\beta}_2 X_{12})^2 + (Y_2 - \hat{\beta}_1 X_{21} - \hat{\beta}_2 X_{12})^2 \longrightarrow (2)$

equation (2) can be simplified by using the assumptions

$X_{11} = X_{12}$, $X_{21} = X_{22}$, $X_{11} + X_{21} = 0$, $X_{12} + X_{22} = 0$ &

$Y_1 + Y_2 = 0$

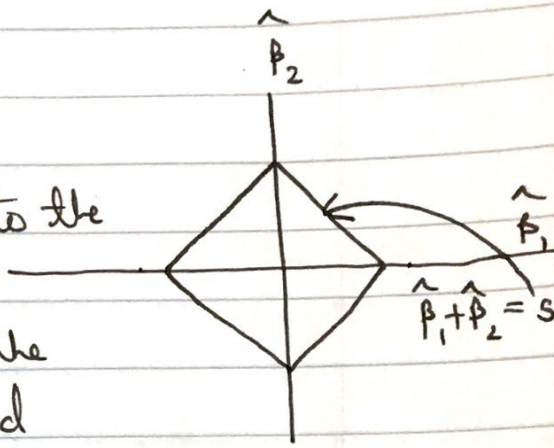
which minimizes to 自

$$Y_1 - (\hat{\beta}_1 + \hat{\beta}_2) X_{11})^2 \longrightarrow (3)$$

from the 1st constraint

$|\hat{\beta}_1| + |\hat{\beta}_2| \leq s$ → this corresponds to the lasso diamond

$\hat{\beta}_1 + \hat{\beta}_2 = s$ → this corresponds to the edge on the lasso diamond



solution to the optimization problem:

$$\hat{\beta}_1 + \hat{\beta}_2 = \frac{Y_1}{X_{11}}$$

As values for $\hat{\beta}_1$ & $\hat{\beta}_2$ vary along the line $\hat{\beta}_1 + \hat{\beta}_2 = \frac{Y_1}{X_{11}}$, the contours related to equation ③ will touch the lasso diamond edge at different points.

for this reason, the entire edge $\hat{\beta}_1 + \hat{\beta}_2 = s$ can be a solution to the lasso optimization problem as well as edge at $\hat{\beta}_1 + \hat{\beta}_2 = -s$.

Here, the lasso optimization problem does not have a unique solution.

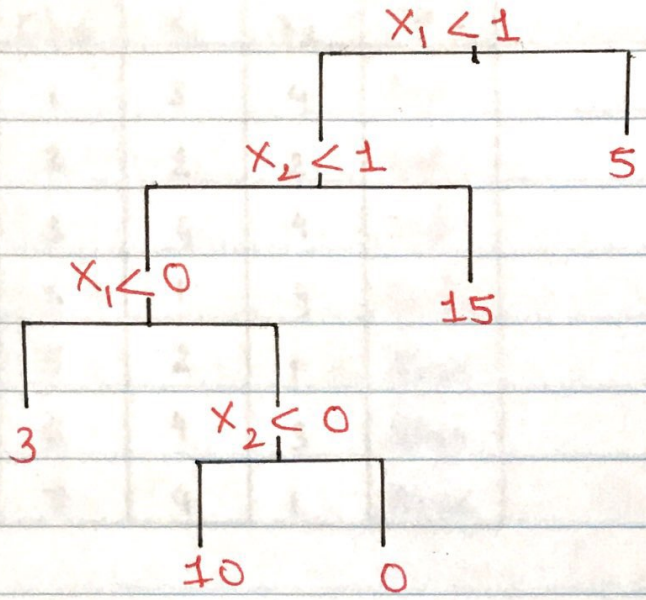
The general forms of solutions are the two line segments as follows:

① $\hat{\beta}_1 + \hat{\beta}_2 = s$ such that $\hat{\beta}_1 \geq 0$ and $\hat{\beta}_2 \geq 0$

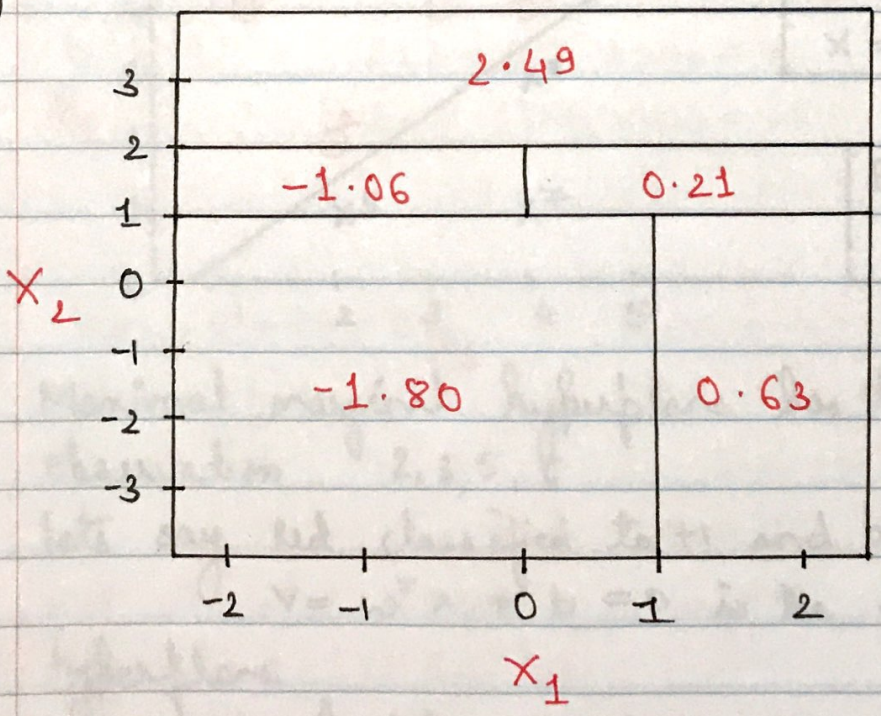
② $\hat{\beta}_1 + \hat{\beta}_2 = -s$ such that $\hat{\beta}_1 \leq 0$ & $\hat{\beta}_2 \leq 0$

5) ISLR 8.4.5

a)


$X_1 < 1$
$X_2 < 1$
5
$X_1 < 0$
15
3
$X_2 < 0$
10
0

b)

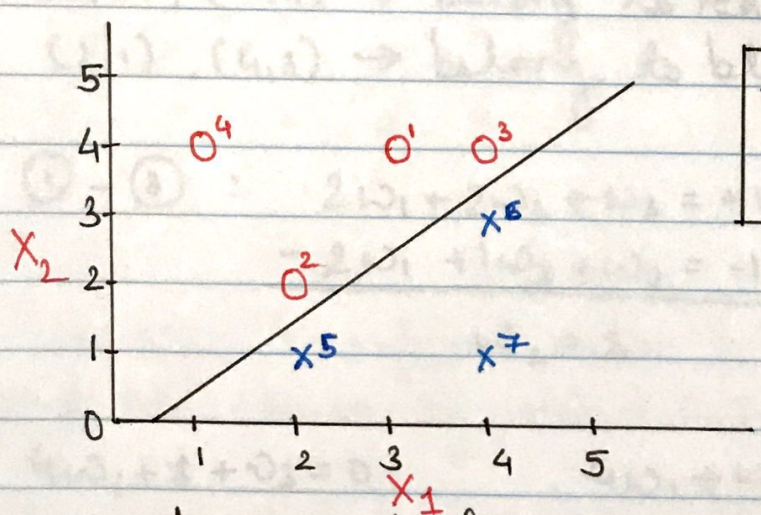

2.49
-1.06
0.21
-1.80
0.63
3
2
1
0
-1
-2
-3
X_2
-2
-1
0
1
2
X_1

6) ISLR 9.7.3

a)

Obs	X_1	X_2	Y
1	3	4	Red
2	2	2	Red
3	4	4	Red
4	1	4	Red
5	2	1	Blue
6	4	3	Blue
7	4	1	Blue

b)



Legends
O = Red
X = Blue

Equation of hyperplane
$1 - 2X_1 + 2X_2 = 0$

Maximal marginal hyperplane has to be in between observation 2, 3, 5, 6

lets say Red classified to +1 and blue classified to -1

$Y = w^T x_i + b = 0$ is the equation for a hyperplane

for this problem it takes like

$0 = w_1 x_1 + w_2 x_2 + w_3$

positive points $[(2,2),(4,4)]$

negative points [(2,1), (4,3)]

Hyperplane linear equations:

① Obs 2 : $2w_1 + 2w_2 + w_3 = 1$
② Obs 3 : $4w_1 + 4w_2 + w_3 = 1$
③ Obs 5 : $2w_1 + 1w_2 + w_3 = -1$
④ Obs 6 : $4w_1 + 3w_2 + w_3 = -1$

Co-ordinates : (2,2), (4,4), (2,1), (4,3)
(2,2), (4,4) → belong to red
(2,1), (4,3) → belong to blue

① − ③ :
$$\begin{array}{r} 2w_1 + 2w_2 + w_3 = +1 \\ -2w_1 + 1w_2 + w_3 = -1 \\ \hline w_2 = 2 \end{array}$$

$4w_1 + 7 + w_3 = 0$
$2w_1 + 2(2) + w_3 = 1$

$$\begin{array}{r} 4w_1 + w_3 + 7 = 0 \\ -2w_1 + w_3 + 3 = 0 \\ \hline 2w_1 + 4 = 0 \end{array}$$

$\therefore w_1 = -2$

put $w_1 = -2$ & $w_2 = 2$ in eq^ ①

$2(-2) + 2(2) + w_3 = 1$

$\therefore w_3 = 1$

hence equation of hyperplane : $\boxed{1 - 2x_1 + 2x_2 = 0}$

c) From the equation of the hyperplane

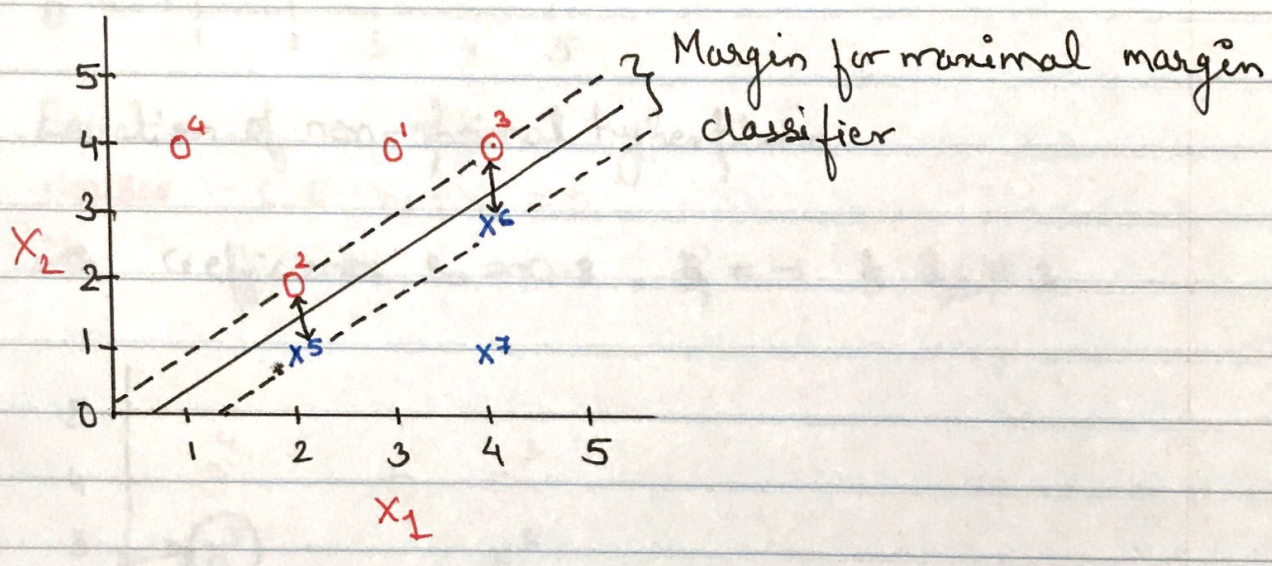
$1 - 2X_1 + 2X_2 = 0$ Red $= +1$ & Blue $= -1$

Hence, classification rule is:

$$0.5 - X_1 + X_2 > 0$$

with $\beta_0 = 0.5$, $B_1 = -1$ & $B_2 = 1$

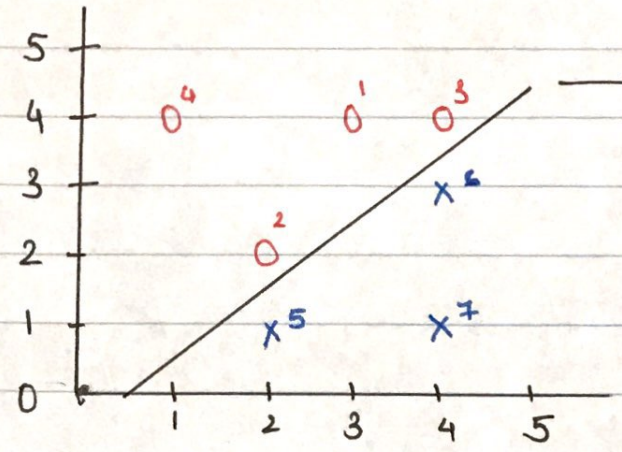
d)



e) There are 4 support vectors located at observations 2, 3, 5 & 6 and are shown by arrows in the graph shown in d)

f) A slight movement in the 7th observation would not affect the maximal marginal hyperplane because its position after a slight movement would be outside the margin

g)



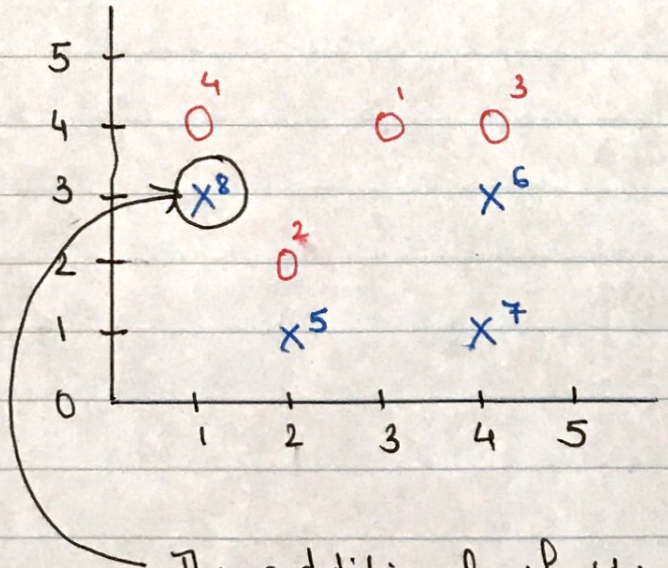
→ Non-optimal seperating hyperplane because it is closer to the blue data points than the red.

Equation of non-optimal hyperplane

$-0.8 - X_1 + X_2 = 0$

with coefficients $B_0 = -0.8$, $B_1 = -1$ & $B_2 = 1$

h)



The additional of this observation makes it so the two classes are no longer seperable by a hyperplane.