6.8 CONCEPTUAL EXERCISES

1. We perform best subset, forward stepwise, and backward stepwise selection on a single data set. For each approach, we obtain p + 1 models, containing 0, 1, 2,...,p predictors. Explain your answers:

(a) Which of the three models with k predictors has the smallest training RSS?

If all the three models have k predictors each, then the model selected by best subset will have the smallest training RSS. Best subset approach searches the whole combinations of predictors that can be formed and than chooses the correct model, in Forward stepwise, the model with k features, will derive k-1 features from the previous model, and hence it doesn’t search over the whole combinations. Similarly backward stepwise selection also looks for a subset of combinations.

(b) Which of the three models with k predictors has the smallest test RSS?

We can be certain about the which approach will yield the model which has the smallest training error, but this can’t be transferred to the test data. The best select approach will give the model with best training error, but it can overfit the data and may not perform that good on test data.

(c) True or False:

i. The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the (k+1)-variable model identified by forward stepwise selection.

TRUE

ii. The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the (k + 1)- variable model identified by backward stepwise selection.

TRUE

iii. The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the (k + 1)- variable model identified by forward stepwise selection.

FALSE. The predictors chosen by forward stepwise selection and backward stepwise selection are independent of each other,

iv. The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the (k+1)-variable model identified by backward stepwise selection.

FALSE.

v. The predictors in the k-variable model identified by best subset are a subset of the predictors in the (k + 1)-variable model identified by best subset selection.

FALSE. Best subset approach looks for the all possible combinations. There's no certainity that predictors in k variable model will be a subset of predictors in the k+1 variable model.

2. For parts (a) through (c), indicate which of i. through iv. is correct. Justify your answer.

(a) The lasso, relative to least squares, is:

iii. Less flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.
Since, lasso results in some feature to have zero coefficients, therefore it will always be less flexible than least squares. Also, lasso decreses with variance at a cost of slight increase in bias.

(b) Repeat (a) for ridge regression relative to least squares

iii. Less flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.
The reason is the same. Ridge also decreases the variance at a cost of slight increase in the bias.

(c) Repeat (a) for non-linear methods relative to least squares.

ii. More flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.
Non linear methods are more flexible than least squares which is a linear method. Also, as flexibility increases, variance also increases, hence this option is the correct one.

3. Suppose we estimate the regression coefficients in a linear regression model by minimizing
(eq 6.8 in text)
for a particular value of s. For parts (a) through (e), indicate which of i. through v. is correct. Justify your answer.

We can visualize s as inversely propotional to lambda, in figure 6.9 in the text, it may help with this answer. So, as s increases, see the changes in the curve from right to left.

(a) As we increase s from 0, the training RSS will:

iv. Steadily decrease
As we increase s, the effect of regularization will decrease. So, we are lifting the restrictions

that we imposed on s. The model will behave more as least square, and training error will keep on decreasing as we increase s.

(b) Repeat (a) for test RSS.

ii. Decrease initially, and then eventually start increasing in a U shape.
As we can look from the graph (fig. 6.9 in chapter), as we go from right to left, the test error first decreases, reaches a minimum and then increases.

(c) Repeat (a) for variance.

iii. Steadily increase.
As we increase s, the model is fitting the data better and as training error decreases, variance increases. We can also conform this from graph, as we go from right to left, variance increases.

(d) Repeat (a) for (squared) bias.

iv. Steadily decrease.
With increase in s, the bias decreases steadily. This can be confirmed through the graph, as we go from right to left, the black line describing the square of bias is decreasing steadily.

(e) Repeat (a) for the irreducible error

v. Remain constant.
Irreducable error, as the name suggests cannot be reduced, and remains constant independent of the method used for fitting.

4. Suppose we estimate the regression coefficients in a linear regression model by minimizing

(eq 6.7)

for a particular value of λ. For parts (a) through (e), indicate which of i. through v. is correct. Justify your answer.

Be careful, this time its lambda, in previous question it was s. They are inversely related to each other,

(a) As we increase λ from 0, the training RSS will:

iii. Steadily increase.
As we increase λ, the regularization effect in the model will increase. The flexibility of the model decreases and RSS increases.

(b) Repeat (a) for test RSS.

ii. Decrease initially, and then eventually start increasing in a U shape.
We can see it from the graph (fig. 6.9), as the flexibility increases, test error (purple line) first decreases, reaches a min, and then increases.

(c) Repeat (a) for variance.

iv. Steadily decrease.
Increasing value of lambda causes the model to become less and less flexible. As flexibility decreases, variance decreases.

(d) Repeat (a) for (squared) bias.

iii. Steadily increases
(squared) bias increases with lambda. From the graph we can conclude that increasing lambda will result in decreasing variance and increasing bias.

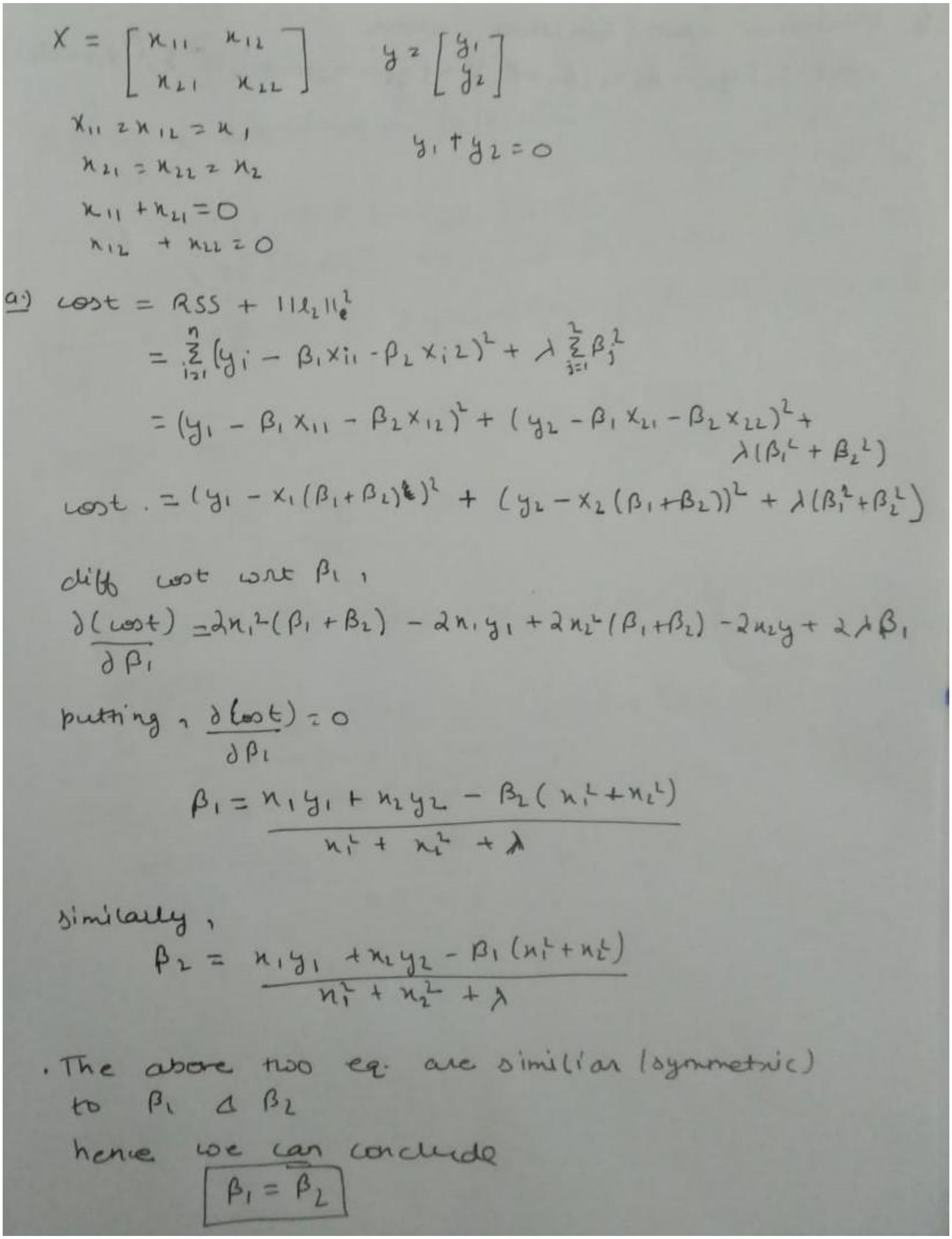
(e) Repeat (a) for the irreducible error.

v. Remains constant
Irreducible error is independent of the model, and remains constant.

5. It is well-known that ridge regression tends to give similar coefficient values to correlated variables, whereas the lasso may give quite different coefficient values to correlated variables.
We will now explore this property in a very simple setting. Suppose that n = 2, p = 2, x11 = x12, x21 = x22. Furthermore, suppose that y1 +y2 = 0 and x11 +x21 = 0 and x12 +x22 = 0, so that the estimate for the intercept in a least squares, ridge regression, or lasso model is zero: βˆ0 = 0.

(a) Write out the ridge regression optimization problem in this setting.
← answer in the image below→

(b) Argue that in this setting, the ridge coefficient estimates satisfy β^1 = β^2.



$$X = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} \qquad y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

$x_{11} = x_{12} = x_1$ $\qquad y_1 + y_2 = 0$

$x_{21} = x_{22} = x_2$

$x_{11} + x_{21} = 0$

$x_{12} + x_{22} = 0$

a.) $\text{cost} = RSS + \|\ell_2\|_e^2$

$$= \sum_{i=1}^{n} (y_i - \beta_1 x_{i1} - \beta_2 x_{i2})^2 + \lambda \sum_{j=1}^{2} \beta_j^2$$

$$= (y_1 - \beta_1 x_{11} - \beta_2 x_{12})^2 + (y_2 - \beta_1 x_{21} - \beta_2 x_{22})^2 + \lambda(\beta_1^2 + \beta_2^2)$$

$$\text{cost} = (y_1 - x_1(\beta_1 + \beta_2))^2 + (y_2 - x_2(\beta_1 + \beta_2))^2 + \lambda(\beta_1^2 + \beta_2^2)$$

diff cost wrt β_1,

$$\frac{\partial(\text{cost})}{\partial \beta_1} = 2x_1^2(\beta_1 + \beta_2) - 2x_1 y_1 + 2x_2^2(\beta_1 + \beta_2) - 2x_2 y + 2\lambda\beta_1$$

putting, $\frac{\partial(\text{cost})}{\partial \beta_1} = 0$

$$\beta_1 = \frac{x_1 y_1 + x_2 y_2 - \beta_2(x_1^2 + x_2^2)}{x_1^2 + x_2^2 + \lambda}$$

similarly,

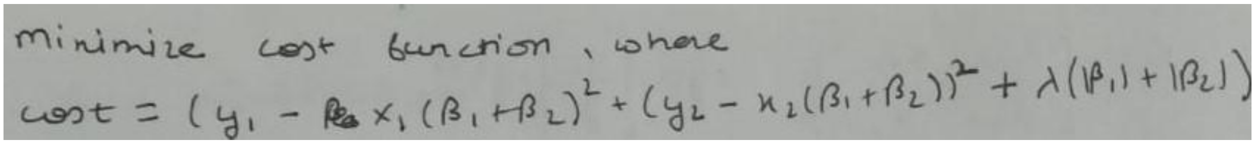
$$\beta_2 = \frac{x_1 y_1 + x_2 y_2 - \beta_1(x_1^2 + x_2^2)}{x_1^2 + x_2^2 + \lambda}$$

- The above two eq. are similiar (symmetric) to β_1 & β_2

hence we can conclude

$$\boxed{\beta_1 = \beta_2}$$

(c) Write out the lasso optimization problem in this setting.



minimize cost function , where

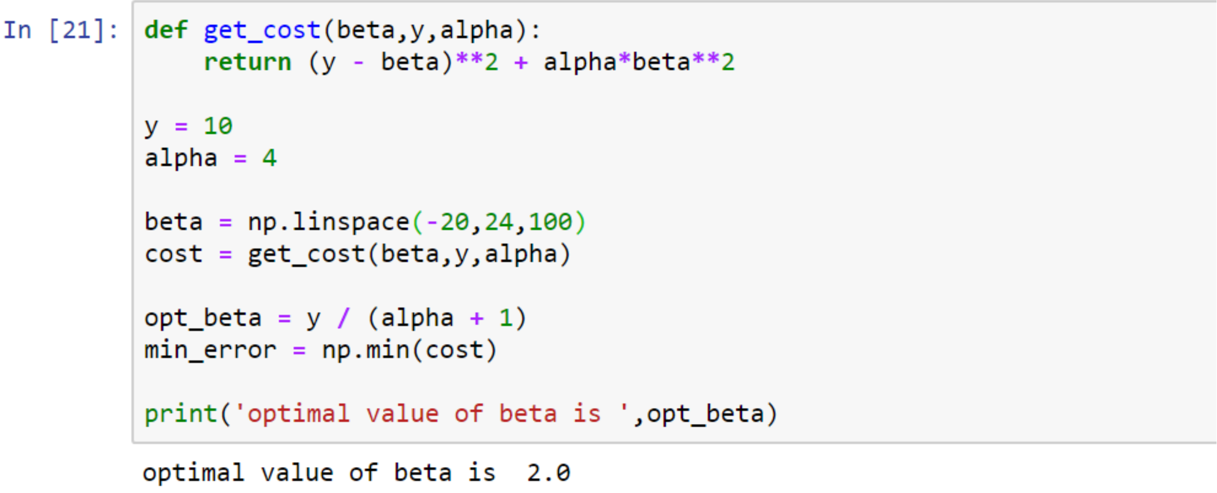
$$cost = (y_1 - x_1(\beta_1+\beta_2)^2 + (y_2 - x_2(\beta_1+\beta_2))^2 + \lambda(|\beta_1| + |\beta_2|)$$

(d) Argue that in this setting, the lasso coefficients βˆ1 and βˆ2 are not unique—in other words, there are many possible solutions to the optimization problem in (c). Describe these solutions.

Will add this in the future ☺

6. We will now explore (6.12) and (6.13) further.

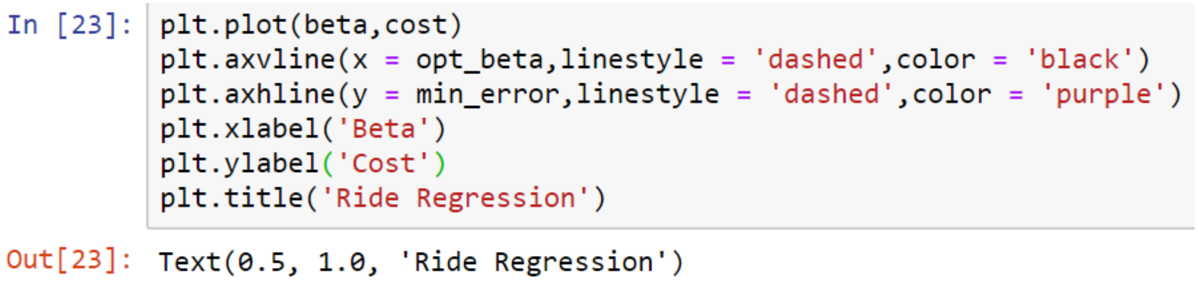
(a) Consider (6.12) with p = 1. For some choice of y1 and λ > 0, plot (6.12) as a function of β1. Your plot should confirm that (6.12) is solved by (6.14).



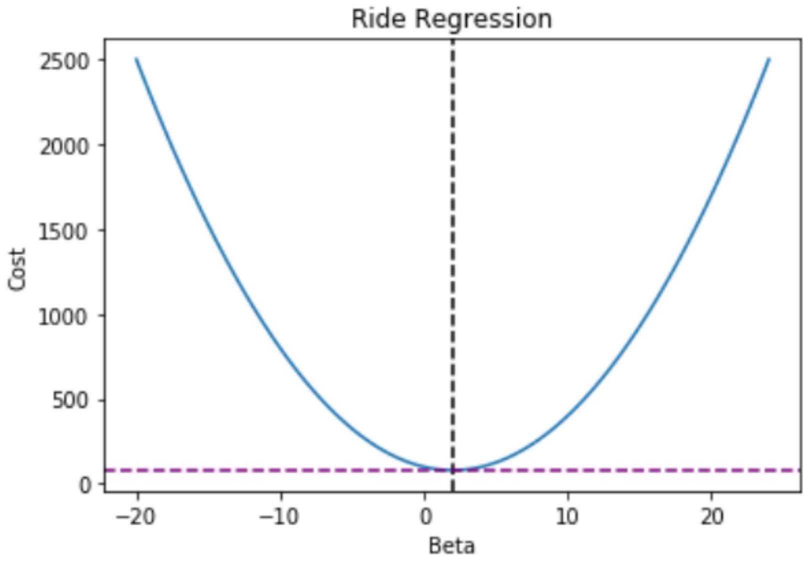
In [21]:

```
def get_cost(beta,y,alpha):
    return (y - beta)**2 + alpha*beta**2

y = 10
alpha = 4

beta = np.linspace(-20,24,100)
cost = get_cost(beta,y,alpha)

opt_beta = y / (alpha + 1)
min_error = np.min(cost)

print('optimal value of beta is ',opt_beta)
```

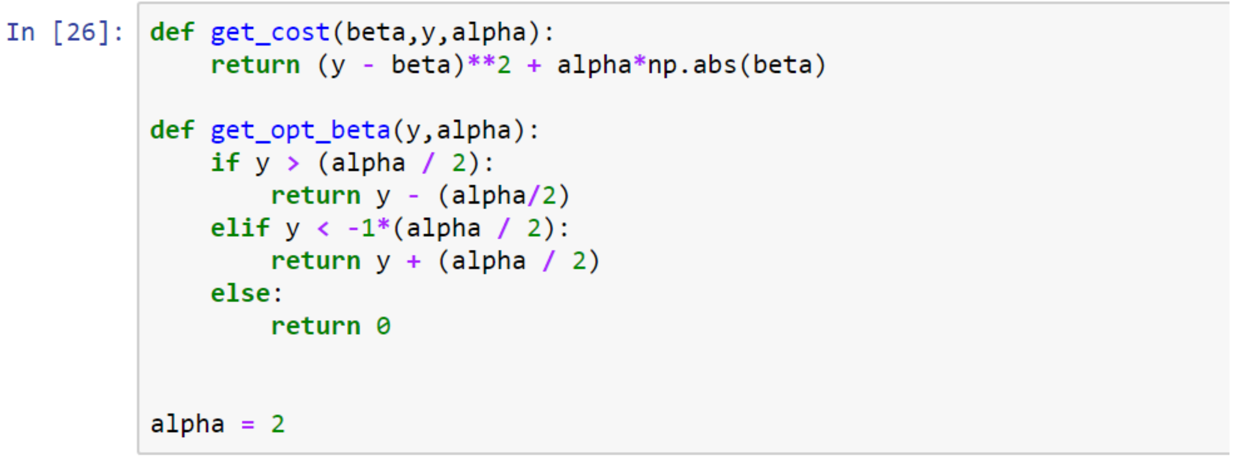
```
optimal value of beta is  2.0
```



```
In [23]: plt.plot(beta,cost)
         plt.axvline(x = opt_beta,linestyle = 'dashed',color = 'black')
         plt.axhline(y = min_error,linestyle = 'dashed',color = 'purple')
         plt.xlabel('Beta')
         plt.ylabel('Cost')
         plt.title('Ride Regression')

Out[23]: Text(0.5, 1.0, 'Ride Regression')
```



(b) Consider (6.13) with p = 1. For some choice of y1 and λ > 0, plot (6.13) as a function of β1. Your plot should confirm that (6.13) is solved by (6.15).



```
In [26]: def get_cost(beta,y,alpha):
             return (y - beta)**2 + alpha*np.abs(beta)

         def get_opt_beta(y,alpha):
             if y > (alpha / 2):
                 return y - (alpha/2)
             elif y < -1*(alpha / 2):
                 return y + (alpha / 2)
             else:
                 return 0


         alpha = 2
```

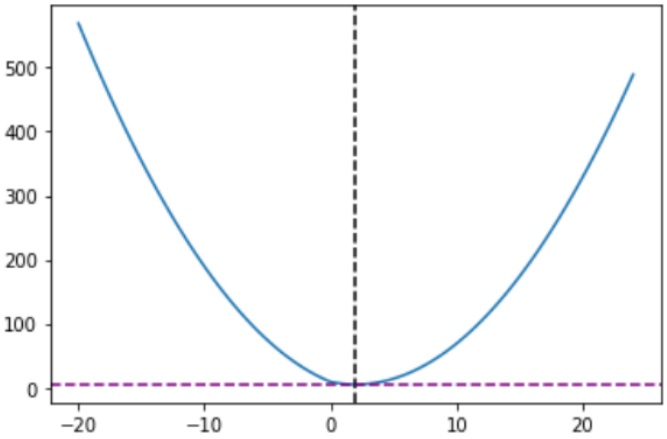
In [29]:

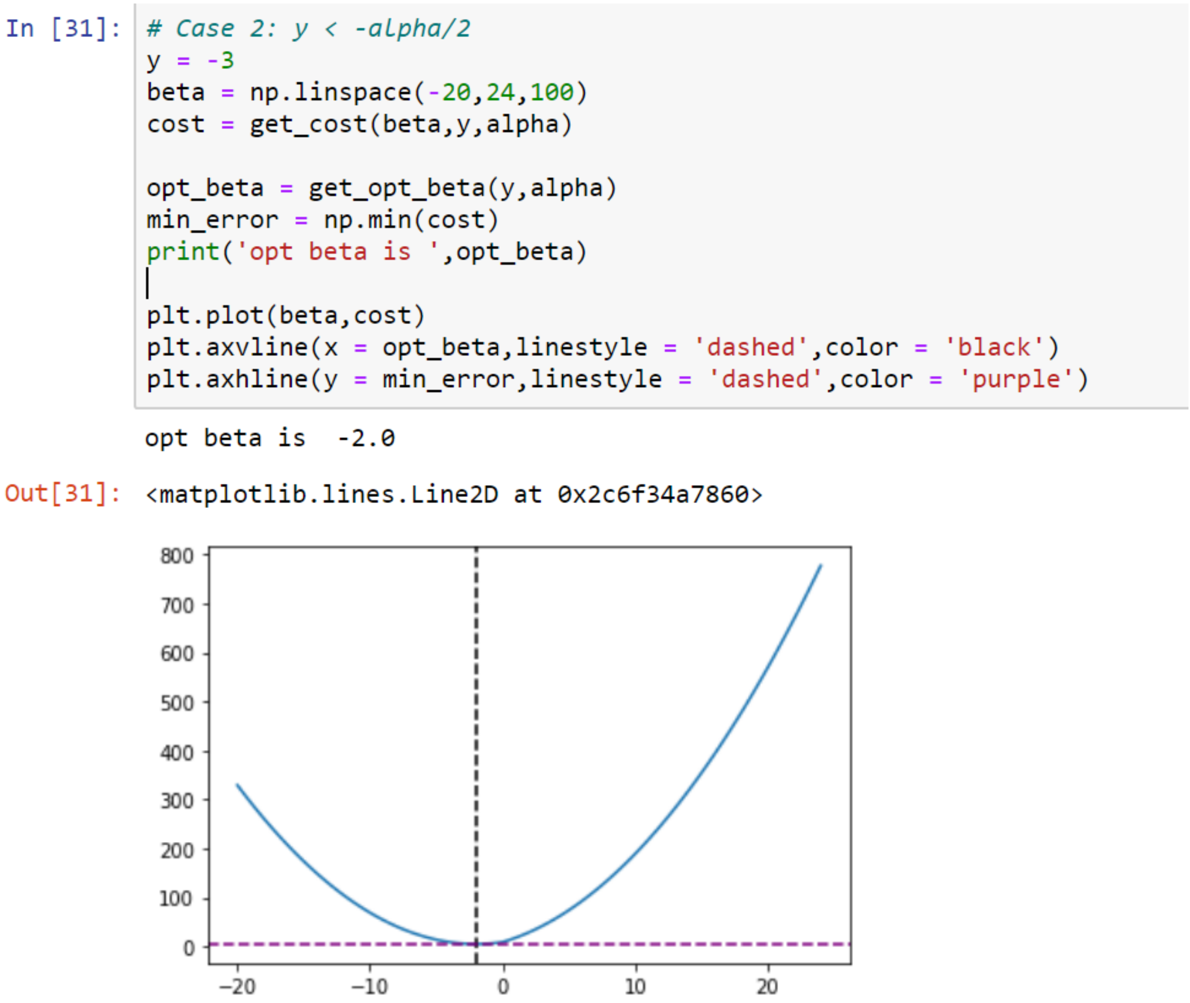
```
# Case 1: y > alpha/2
y = 3
beta = np.linspace(-20,24,100)
cost = get_cost(beta,y,alpha)

opt_beta = get_opt_beta(y,alpha)
min_error = np.min(cost)
print('opt beta is ',opt_beta)

plt.plot(beta,cost)
plt.axvline(x = opt_beta,linestyle = 'dashed',color = 'black')
plt.axhline(y = min_error,linestyle = 'dashed',color = 'purple')
```

```
opt beta is  2.0
```

Out[29]:

```
<matplotlib.lines.Line2D at 0x2c6f342d828>
```





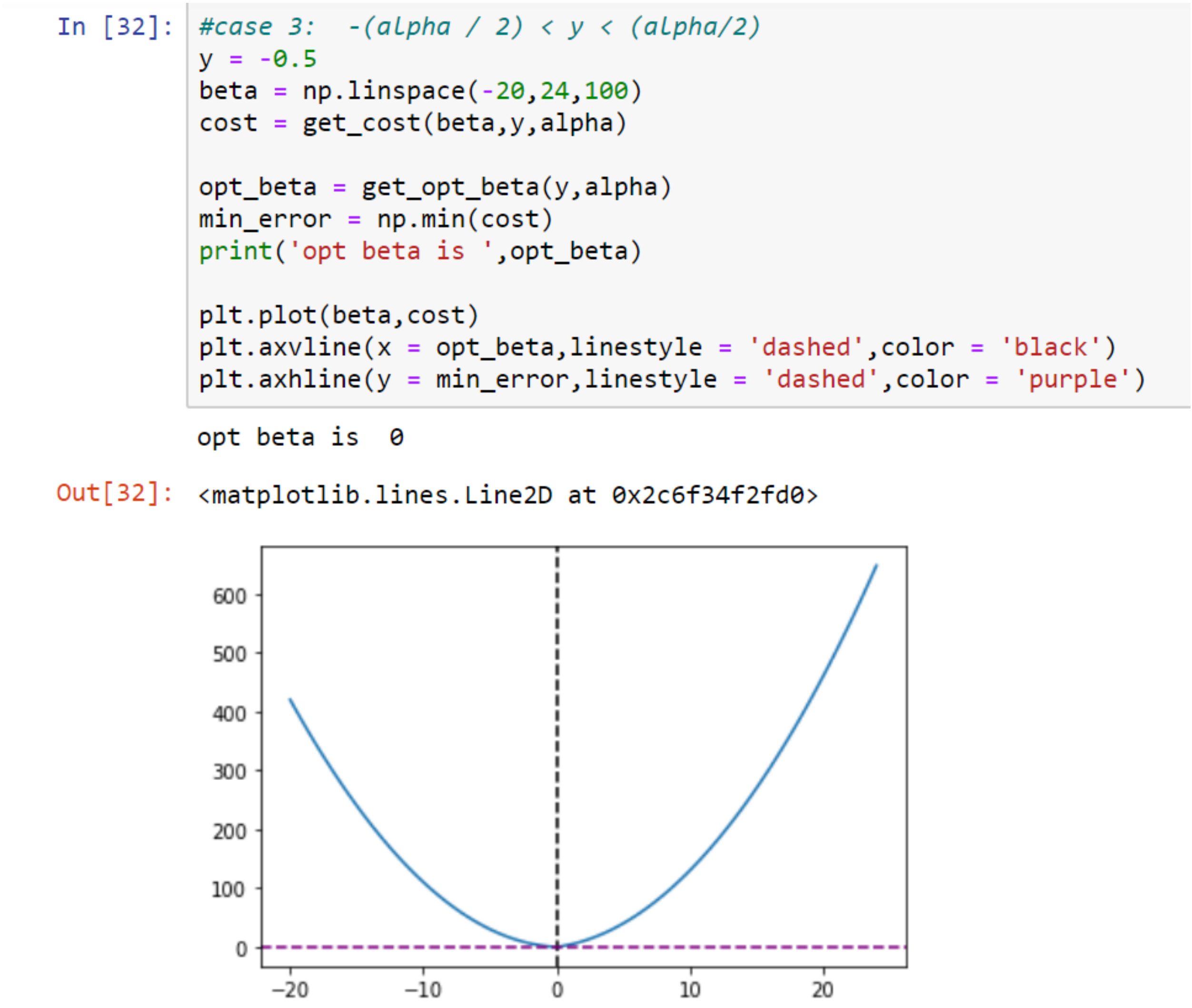
In [31]:

```
# Case 2: y < -alpha/2
y = -3
beta = np.linspace(-20,24,100)
cost = get_cost(beta,y,alpha)

opt_beta = get_opt_beta(y,alpha)
min_error = np.min(cost)
print('opt beta is ',opt_beta)

plt.plot(beta,cost)
plt.axvline(x = opt_beta,linestyle = 'dashed',color = 'black')
plt.axhline(y = min_error,linestyle = 'dashed',color = 'purple')
```

```
opt beta is  -2.0
```

Out[31]:

```
<matplotlib.lines.Line2D at 0x2c6f34a7860>
```



```
In [32]: #case 3:  -(alpha / 2) < y < (alpha/2)
         y = -0.5
         beta = np.linspace(-20,24,100)
         cost = get_cost(beta,y,alpha)

         opt_beta = get_opt_beta(y,alpha)
         min_error = np.min(cost)
         print('opt beta is ',opt_beta)

         plt.plot(beta,cost)
         plt.axvline(x = opt_beta,linestyle = 'dashed',color = 'black')
         plt.axhline(y = min_error,linestyle = 'dashed',color = 'purple')

         opt beta is  0

Out[32]: <matplotlib.lines.Line2D at 0x2c6f34f2fd0>
```

7. We will now derive the Bayesian connection to the lasso and ridge regression discussed in Section 6.2.2.

(a) Suppose that yi = β0+ p j=1 xijβj+i where 1,...,n are independent and identically distributed from a N(0, σ2) distribution. Write out the likelihood for the data.

(b) Assume the following prior for β: β1,...,βp are independent and identically distributed according to a double-exponential distribution with mean 0 and common scale parameter b: i.e. p(β) = 1 2b exp(−|β|/b). Write out the posterior for β in this setting.

(c) Argue that the lasso estimate is the mode for β under this posterior distribution.

(d) Now assume the following prior for β: β1,...,βp are independent and identically distributed according to a normal distribution with mean zero and variance c. Write out the posterior for β in this setting.

(e) Argue that the ridge regression estimate is both the mode and the mean for β under this posterior distribution.

Will update in future. ☺