Question 3

The logistic model for q classes can be defined as:

$$P(Y_i = k \mid X_i) = \frac{e^{x_i B_k}}{\sum_{k'=1}^{q} e^{x_i B_k}} = P_{ik}$$

where x_i = i^{th} row of the feature matrix, X

B_k = k^{th} column of the parameter matrix, B

As calculated in class, we want to maximize $L(B)$ using log-likelihood

$$L(B) = \prod_{i=1}^{n} \prod_{k=1}^{q} P(y[i] = k \mid x[i,:])^{1\!\!1[i,k]}$$

lets take log of $L(B)$

$$\log L(B) = \log\left(\prod_{i=1}^{n} \prod_{k=1}^{q} P(y[i] = k) \mid X[i,:]\right)^{1\!\!1[i,k]}$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{q} 1\!\!1[i,k] \cdot \log P(y[i] = k \mid X[i,:])$$

Substituting the P_{ik} function

$$= \sum_{i=1}^{n} \sum_{k=1}^{q} \mathbb{1}[i,k] \cdot \log \left(\frac{e^{X[i,:] B[:,k]}}{\sum_{k'=1}^{q} e^{X[i,:] B[:,k']}} \right)$$

simplifying wrt log

$$= \sum_{i=1}^{n} \sum_{k=1}^{q} \mathbb{1}[i,k] \cdot \left(X[i,:] B[:,k] - \log \left(\sum_{k'=1}^{q} e^{X[i,:] B[:,k']} \right) \right)$$

Let's calculate the gradient for this $\log L(B)$ function.

We will do it partially for each term

$$\frac{d}{dB} \left(X[i,:] \; B[:,k] \right) = X[i,j] \quad \text{—} \; 1$$

$$\frac{d}{dB} \left[\log \left(\sum_{k'=1}^{q} e^{X[i,:] B[:,k']} \right) \right] = \frac{X[i,j] \cdot e^{X[i,:] B[:,k']}}{\sum_{k'=1}^{q} e^{X[i,:] B[:,k']}}$$

$$= X[i,j] \cdot P_{ik} \quad \text{—} \; 2$$

Lets substitute these for the entire derivative

Furthermore $\mathbb{1}[i,k]$ is treated as a constant.

$$\frac{d}{dB} \log L(B) = \sum_{i=1}^{n} \left(\mathbb{1}[i,k] \cdot X[i,j] - X_{i,j} P_{ik} \right)$$

taking $X[i,j]$ as common

$$= \sum_{i=1}^{n} X[i,j]\,(\mathbb{1}[i,k] - P_{ik})$$

Converting to matrix form,
we get

$$\boxed{\frac{d \log L(\beta)}{d\beta} = X^T(I - P)}$$

Question 1A

Given

$$X = [x_1, x_2, \ldots, x_n]^T$$

$$y = [y_1, y_2, \ldots, y_n]^T$$

the model can be defined as

$y = XB$ where B is the parameter matrix to be estimated.

Then Mean Square Error $L(B)$ is given by

$$L(B) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - Bx_i)^2$$

Since, the goal is to minimize MSE wrt B

We can iteratively update B using

$$B^{(k+1)} = B^{(k)} - \alpha \frac{dL(B)}{dB}$$

Assume that the iterative update rule for the parameter, B behaves like a fixed point iteration

$\boxed{B_{t+1} = B_t - \alpha\, l(B)}$ where $l(B)$ is the derived loss function for MSE.

Convergence for this iteration:

$$|B_{t+1} - B^*| < |B_t - B^*|$$

where B^* is the explicit value that minimizes MSE.

This means for every iteration B_{t+1} should reach closer to the B^*.

This means

$$|B_{t+1} - B^*| = |B_t - \alpha \frac{d}{dB} l(B) - B^*)$$

Since we derived $\frac{d\, l(B)}{dB}$. Let's use that

for eigen value analysis.

$$\left| 1 - \frac{2\alpha}{n} \lambda_{max}(X^T X) \right| < 1$$

For linear regression case where X is a vector

$X^T X$ is scalar

$$X^T X = \sum_{i=1}^{n} x_i^2$$

Then the condition for convergence implies

$$-1 < 1 - \frac{2\alpha}{n} \sum_{i=1}^{n} x_i^2 < 1$$

simplifying

$$0 < \frac{2\alpha}{n} \sum_{i=1}^{n} x_i^2 < 2$$

range for α

$$0 < \alpha < \frac{n}{\sum_{i=1}^{n} x_i^2}$$

Question 1 B

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y - xB)^2$$

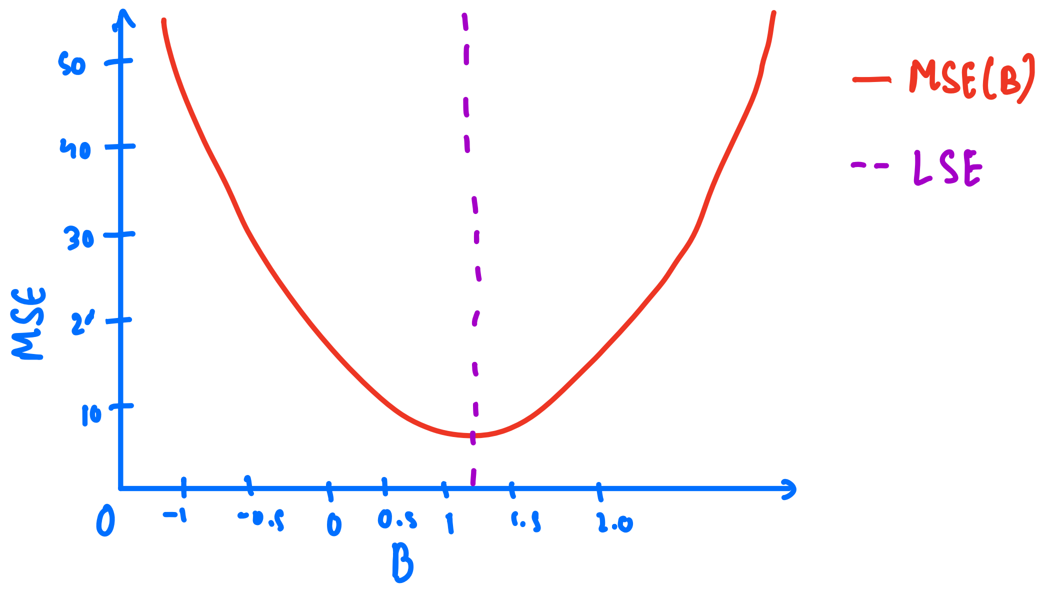
This is a quadratic function, therefore the plot will be parabolic.

Since the LSE of B minimizes the MSE, we set the derivative of MSE wrt B to 0

$$-\frac{2}{n} \sum_{i=1}^{n} x (y - xB) = 0$$

solving for B

$$B_{LSE} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$



MSE
50
40
30
2'
10
0
-1
-0.5
0
0.5
1
1.5
2.0
B
MSE(B)
LSE