CM124/224, Fall 2023
Problem Set 2: Linear and Ridge Regression
Due November 10, 2023 at 11:59 pm PST

1 Course Challenge: Ancestry Inference

The first challenge of the course is now live! Following the instructions detailed in PS1, read the documentation in the challenge website (https://compmed.codalab.click), and explore the dataset provided in the link. The challenge will be open until November 17 – you may experiment with various models to achieve the most accurate predictions you can.

2 Relaxing the assumptions of linear regression [15 pts]

In class, we discussed the assumptions of linear regression (OLS) underlying GWAS. Assume we are analyzing a SNP with three genotypes: CC, CT, TT which we code as 0, 1, and 2 respectively. In a standard GWAS, we test this SNP for association with a quantitative phenotype using linear regression: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where y_i is the phenotype in individual i ($y_i \in \mathbb{R}$), x_i is the genotype in individual i at the SNP ($x_i \in \{0, 1, 2\}$) and $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

We collect a training dataset of $2n$ individuals: $\{(x_i, y_i), i = 1, \ldots, 2n\}$, $x_i \in \{0, 1, 2\}$, $y \in \mathbb{R}$. Due to how the study was designed, the first n individuals were measured in center A while the last n individuals were measured in center B. We are also told that the variance of the phenotype measurements in center B is twice is large as the variance in center A (this type of information is not always known however). We will explore how we can model this data and perform inference (estimate the parameters of the model).

(a) We would now like to write a probabilistic model describing our data. The probabilistic model for the data from center A is given by: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, \ldots, n$, where $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Write the probabilistic model for the data from center B (*i.e.*, the equation relating y_i to x_i for an individual from center B analogous to the one we have for an individual from center A).

Solution:

$$\epsilon_i \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right) \rightarrow y_i \sim \mathcal{N}\left(\beta_0 + \beta_1 x_i, \sigma^2\right), i = 1, \ldots, n$$

$$\epsilon_i \overset{iid}{\sim} \mathcal{N}\left(0, 2\sigma^2\right) \rightarrow y_i \sim \mathcal{N}\left(\beta_0 + \beta_1 x_i, 2\sigma^2\right), i = n+1, \ldots, 2n$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \epsilon_i \overset{iid}{\sim} \mathcal{N}\left(0, 2\sigma^2\right), i = n+1, \ldots, 2n$$

(b) Write the log probability for a single training sample from center A: $\log p(y_i|x_i;\beta_0,\beta_1,\sigma^2)$.
Solution:

$$A \rightarrow p(y_i|x_i;\beta_0,\beta_1,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(y_i-(\beta_0+\beta_1 x_i))^2}{2\sigma^2}$$

$$\log p(y_i|x_i;\beta_0,\beta_1,\sigma^2) = -\frac{(y_i-(\beta_0+\beta_1 x_i))^2}{2\sigma^2} + \log(\sqrt{2\pi\sigma^2})$$

(c) Write the log probability for a single training sample from center B: $\log p(y_i|x_i;\beta_0,\beta_1,\sigma^2)$.
Solution:

$$B \rightarrow p(y_i|x_i;\beta_0,\beta_1,\sigma^2) = \frac{1}{\sqrt{4\pi\sigma^2}}\exp(-\frac{(y_i-(\beta_0+\beta_1 x_i))^2}{4\sigma^2}$$

$$\log p(y_i|x_i;\beta_0,\beta_1,\sigma^2) = -\frac{(y_i-(\beta_0+\beta_1 x_i))^2}{4\sigma^2} + \log(\sqrt{4\pi\sigma^2})$$

(d) Write the log-likelihood function for the full training dataset $\mathcal{LL}(\beta_0,\beta_1,\sigma^2)$. Express the log-likelihood in terms of β_0, β_1, and σ^2, and $\{x_i,y_i)\}_{i=1}^{2n}$. You may use C to represent constant terms that do not depend on β_0, β_1, and σ^2. **Solution:**

$$\mathcal{LL}(\beta_0,\beta_1,\sigma^2) = \sum_{i=1}^{n}\log p_A(y_i|x_i;\beta_0,\beta_1,\sigma^2) + \sum_{i=n+1}^{2n}\log p_B(y_i|x_i;\beta_0,\beta_1,\sigma^2)$$

$$= \sum_{i=1}^{n} -\frac{(y_i-(\beta_0+\beta_1 x_i))^2}{2\sigma^2} + \log(\sqrt{2\pi\sigma^2}) + \sum_{i=n+1}^{2n} -\frac{(y_i-(\beta_0+\beta_1 x_i))^2}{4\sigma^2} + \log(\sqrt{4\pi\sigma^2})$$

$$= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}[(y_i-(\beta_0+\beta_1 x_i))^2] - \frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi) - \frac{1}{4\sigma^2}\sum_{i=n+1}^{2n}[(y_i-(\beta_0+\beta_1 x_i))^2] - \frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(4\pi)$$

$$= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}[(y_i-(\beta_0+\beta_1 x_i))^2] - \frac{1}{4\sigma^2}\sum_{i=n+1}^{2n}[(y_i-(\beta_0+\beta_1 x_i))^2] - n\log(\sigma^2) + C$$

(e) Show that finding the maximum likelihood estimate (MLE) of (β_0, β_1) is the same as finding the value of (β_0, β_1) that minimizes the cost function:

$$J(\beta_0, \beta_1) \quad = \quad A \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 + B \sum_{i=n+1}^{2n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Write down the values of A and B in this cost function.

Solution:

$$\max_{\beta_0,\beta_1} \mathcal{LL}(\beta_0, \beta_1) = \min_{\beta_0,\beta_1} -2\mathcal{LL}(\beta_0, \beta_1)$$

$$= \min_{\beta_0,\beta_1} \frac{1}{\sigma^2} \sum_{i=1}^{n} [(y_i - (\beta_0 + \beta_1 x_i))^2] + \frac{1}{2\sigma^2} \sum_{i=n+1}^{2n} [(y_i - (\beta_0 + \beta_1 x_i))^2] = \min_{\beta_0,\beta_1} J(\beta_0, \beta_1)$$

$$\text{where } A = \frac{1}{\sigma^2}, B = \frac{1}{2\sigma^2}$$

$$= \min_{\beta_0,\beta_1} \sum_{i=1}^{2n} w_i(y_i - (\beta_0 + \beta_1 x_i)^2, w_i = \begin{cases} \frac{1}{\sigma^2} & \text{if } i = 1, \ldots, n \\ \frac{1}{2\sigma^2} & \text{if } i = n+1, \ldots, 2n \end{cases}$$

(f) We denote the vector of parameters $\boldsymbol{\beta} = (\beta_0, \beta_1)$. Let $\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_{2n} \end{pmatrix}$ denote the design matrix and $\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{2n} \end{pmatrix}$ denote the target vector. What are the dimensions of $\boldsymbol{X}$ and $\boldsymbol{y}$?

Solution:

$$X \in \{0, 1, 2\}^{2n \times 2}$$

$$y \in \mathbb{R}^{2n}$$

(g) Show that $J(\boldsymbol{\beta})$ can also be written as:

$$J(\boldsymbol{\beta}) \quad = \quad (\boldsymbol{y} - \boldsymbol{X\beta})^{\mathrm{T}} \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{X\beta})$$

Here $\boldsymbol{W}$ is a $2n \times 2n$ diagonal matrix. Write the expression for $\boldsymbol{W}$ in terms of A and B from the previous question.

Solution:

$$= \min_{\boldsymbol{\beta}} \sum_{i=1}^{2n} w_i (y_i - \boldsymbol{\beta}^T \boldsymbol{x_i})^2, w_i = \begin{cases} \frac{1}{\sigma^2} & \text{if } i = 1, \ldots, n \\ \frac{1}{2\sigma^2} & \text{if } i = n+1, \ldots, 2n \end{cases} = \min_{\boldsymbol{\beta}} J(\boldsymbol{\beta})$$

$$= \min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X\beta})^{\mathrm{T}} \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{X\beta})$$

$$W_{ij} = \begin{cases} \frac{1}{\sigma^2} & \text{if } i = 1, \ldots, n \wedge i = j \\ \frac{1}{2\sigma^2} & \text{if } i = n+1, \ldots, 2n \wedge i = j \\ 0 & o/w \end{cases}$$

$$W = \begin{bmatrix} \frac{1}{\sigma^2} & & & & & \\ & \ddots & & & & \\ & & \frac{1}{\sigma^2} & & & \\ & & & \frac{1}{2\sigma^2} & & \\ & & & & \ddots & \\ & & & & & \frac{1}{2\sigma^2} \end{bmatrix}$$

(h) Show that the optimal value for $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{y}$.

Solution:

$$\begin{aligned}
&\nabla J(\boldsymbol{\beta}) \\
&= \nabla \left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{W}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right] \\
&= \nabla \left[\left(\mathbf{y}^\top - \boldsymbol{\beta}^\top \mathbf{X}^\top\right) \mathbf{W}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right] \\
&= \nabla \left[\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{W}\mathbf{X}\boldsymbol{\beta} - \mathbf{y}^\top \mathbf{W}\mathbf{X}\boldsymbol{\beta} + \mathbf{y}^\top \mathbf{W}\mathbf{y} - \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{W}\mathbf{y}\right] \\
&= \nabla \left(\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{W}\mathbf{X}\boldsymbol{\beta} - \mathbf{y}^\top \mathbf{W}\mathbf{X}\boldsymbol{\beta} + \mathbf{y}^\top \mathbf{W}\mathbf{y} - \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{W}\mathbf{y}\right) \\
&= \nabla \left(\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{W}\mathbf{X}\boldsymbol{\beta}\right) - \nabla \left(\mathbf{y}^\top \mathbf{W}\mathbf{X}\boldsymbol{\beta}\right) + \nabla \left(\mathbf{y}^\top \mathbf{W}\mathbf{y}\right) - \nabla \left(\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{W}\mathbf{y}\right) \\
&= \nabla \left(\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{W}\mathbf{X}\boldsymbol{\beta}\right) - 2\nabla \left(\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{W}\mathbf{y}\right) \\
&= 2\mathbf{X}^\top \mathbf{W}\mathbf{X}\boldsymbol{\beta} - 2\mathbf{X}^\top \mathbf{W}\mathbf{y} = 0.
\end{aligned}$$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{y}$$

I have used the fact that $W = W^T$

3 Data analysis [10 pts]

Files gwas.geno contains the genotypes of 500 individuals at 382 SNPs. File gwas.pheno contains 4 continuous phenotypes for each individual. For each phenotype, perform an association analysis of each SNP for the phenotype. To do this, run linear regression for each phenotype against the genotype at each SNP (coded as 0,1 or 2) with an intercept term.
In R, you can use the lm function:
(https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/lm)
and in Python, you can use the statsmodel.api.OLS function:
(https://www.statsmodels.org/dev/examples/notebooks/generated/ols.html)
to compute linear regression estimates. We would like to find associations while controlling for the FWER. We will use the Bonferroni procedure to control FWER.

For this question, include any plots along with your answers (do not include code).

(a) What is the significance level at which we reject each single SNP test to control the overall FWER at 0.05 ?

Solution: From Bonferroni threshold, reject p values such that

$$p \leq \frac{\alpha}{m}$$

If we control FWER over the test of all SNPs and phenotypes, we would be performing 382*4 tests. Then $\frac{\alpha}{m} = 3.27 \times 10^{-5}$.However, if we control FWER over the test of only all SNPS, we would be performing 382 tests. Then,

$$p \leq \frac{0.05}{382} = 1.3089 \times 10^{-4}$$

(b) For which SNPs and phenotypes, can we reject the null hypothesis of no association at the chosen signficance level ?

Solution:
SNP 43, phenotype 4
SNP 119, phenotype 3
SNP 258, phenotype 2

(c) For each association (*i.e.*, rejected null hypothesis), we need to rule out the possibility that the rejection could be a result of the linear regression model being incorrect. If the p-values are not uniformly distributed, this could indicate that the model assumptions are violated. For each phenotype, do the p-values across the SNPs look uniformly distributed ?

Solution:

I have plotted the histogram of p values and also applied Kolmogorov-Smirnov test to see whether p values follow uniform distribution. Test rejected the only fourth phenotype which is observed from the Figure 1. Therefore, for phenotypes 1,2 and 3, we can conclude that p-values across the SNPs look uniformly distributed.
I have also checked out the QQ plot of p values to find out whether they follow an uniform distribution. The result at Figure 2 also shows that p values that belong to phenotype 4

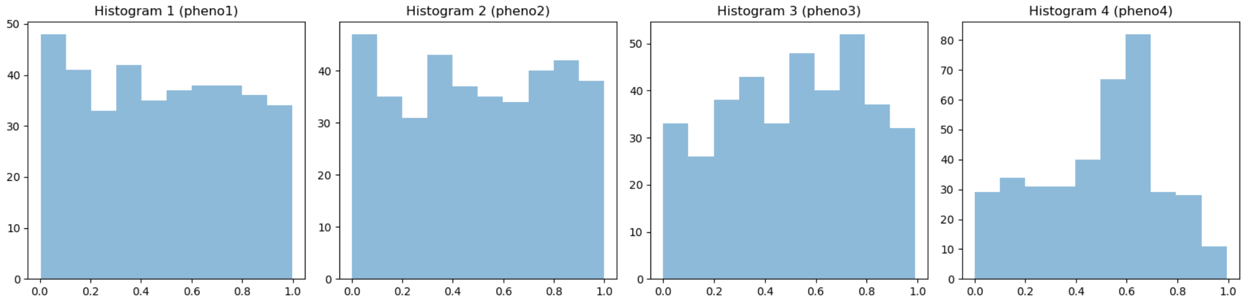

Figure 1: Distribution of p values

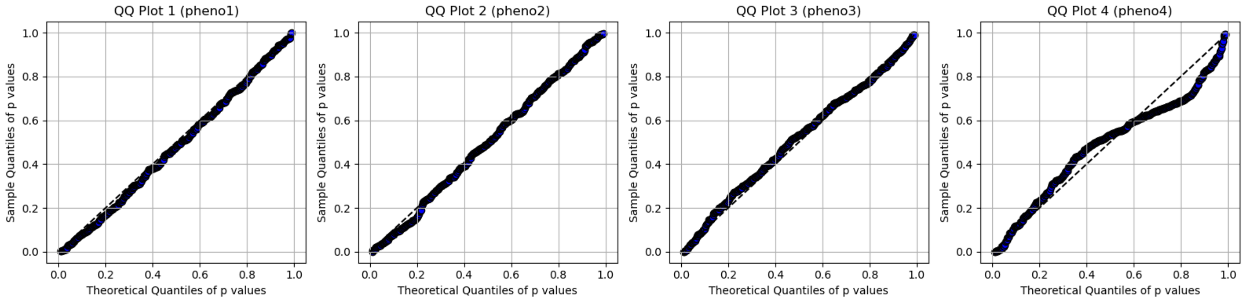

Figure 2: QQ plots of p values

does not follow a linear line, indicating p values are not uniformly distributed for phenotype 4. Also, there are some slight deviation at the phenotype 3, therefore it is also possible that phenotype 3 might not be following uniform distribution.

(d) An additional check is to plot the relationship between phenotype (on y-axis) vs genotype (on x-axis) for the SNPs that are discovered to be associated. Since the genotype takes values in $\{0, 1, 2\}$, we can use either use a boxplot (boxplot in R or Python) or add random noise to the genotype to aid visualization. Plot this relationship for each SNP-phenotype association. Which of the associations might be a result of model violation? In each of these cases, what assumption do you think is violated ? **Solution:**

From Figure 3 (LEFT), we can see that phenotype for genotype 2 is not normally distributed. Therefore, error is not normally distributed. Linear assumption is violated.

From Figure 3(MIDDLE), we can also see that phenotype for genotype 0 is not normally distributed. One half of the phenotype given genotype 0 is between 0 to 5, whereas other half consists of values between 5 to 100. Therefore, error is not normally distributed. Linear assumption is violated. Also, the dependence of the mean phenotype on genotype is not additive.

From Figure 3(RIGHT), we can see that phenotype for all genotypes are normally distributed. Therefore, error is normally distributed. Linear assumption is not violated. There is a clear relationship between SNP 258 and phenotype 2. Also, the dependence of the mean phenotype on genotype is not additive.

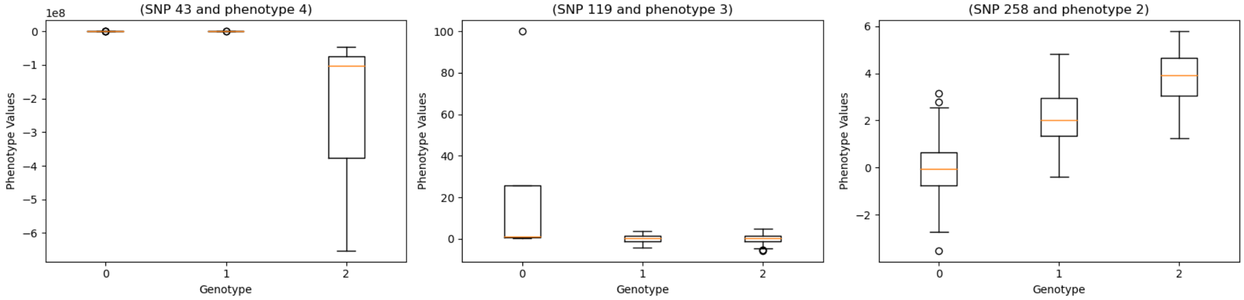

Figure 3: Box Plots

4 Ridge regression [10 pts]

Consider ridge regression where we have n pairs of inputs and outputs, $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^{n}$ where $\boldsymbol{x}_i \in \mathbb{R}^m$. The outputs are centered so we don't need a bias term in our regression. We have shown in class that the ridge regression estimator is given by $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_m\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$ where $\boldsymbol{X} = [\boldsymbol{x}_1 \ldots \boldsymbol{x}_n]^{\mathrm{T}}$ is a $n \times m$ matrix.

(a) Given genotype matrix $\boldsymbol{X}$ and vector of phenotypes $\boldsymbol{y}$, we would like to obtain a linear predictor from $\boldsymbol{X}$ to $\boldsymbol{y}$. In a typical GWAS study, why would ridge regression be preferable over linear regression to accomplish this?

Solution:

Multicollinearity: GWAS typically involves a large number of genetic SNPs as predictor variables. These markers can be highly correlated with each other due to linkage disequilibrium, which can lead to multicollinearity in the data. Ridge regression can handle multicollinearity effectively by introducing a regularization term that shrinks the coefficients of correlated predictors, preventing overfitting.

Overfitting: In GWAS, the number of predictors can be larger than the number of observations (samples or individuals). This high-dimensional data can lead to overfitting in a standard linear regression model, where it may capture noise in the data and perform poorly on new samples. Ridge regression reduces the risk of overfitting by adding a penalty term to the linear regression objective function, which discourages large coefficient values.

Model stability: Note that, X^TX might not be invertible, therefore ridge regression tends to produce more stable and interpretable models compared to linear regression in the context of GWAS. It reduces the sensitivity of the model to small changes in the data, leading to more reliable results.

Improved generalization: Ridge regression can improve the generalization performance of the model. This means that the model is more likely to perform well on new, unseen data because it reduces the risk of fitting the training data too closely.

Variable selection: While ridge regression doesn't perform variable selection in the same way as Lasso regression, it can still help in prioritizing important genetic markers by shrinking less relevant markers' coefficients toward zero. This can be valuable in GWAS to identify the most influential genetic variants associated with a trait or disease.

(b) You are given the genotypes $\boldsymbol{X}$ (ridge.training/test.geno) and phenotypes $\boldsymbol{y}$ (ridge.training/test.pheno) of 1000 individuals. Implement ridge regression. For each of training and test data, plot the mean squared error (MSE) against different parameter settings of $\lambda = 0.001, 2, 5, 8$. How does ridge regression behave as the λ value increases? **Solution:**

I have used the given range and another range of lambda values to clearly explain the behaviour of ridge regression.

From Figure 4, it is easily observable that train MSE increases when we increase the value of lambda since model starts to underfit to the data. Regularization term dominates the actual loss function.

From Figure 5, test MSE starts to decrease since our model starts not to overfit to the model, but this decrease stops when lambda = 100 which is shown at Figure 8.

From Figure 6, we can see that for low values of lambda model overfits to the data which shows that training MSE is too low and for high values of lambda model starts to underfit to the data.

From Figure 7, we can see that for too low and too high values of lambda the MSE is high. There is an optimal lambda which is 100. Therefore, lambda value should be chosen such that model should neither overfit nor underfit.

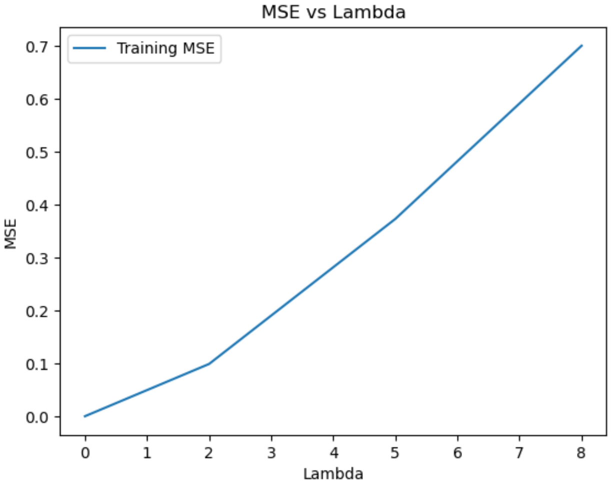

Figure 4: Train MSE vs lambdas(0.001,2,5,8)

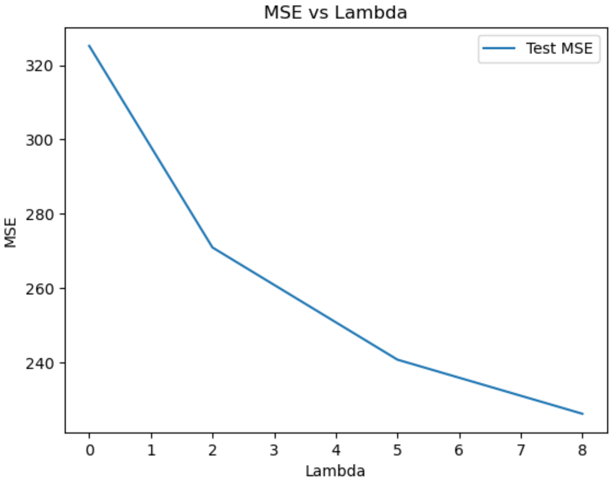

Figure 5: Test MSE vs lambdas(0.001,2,5,8)

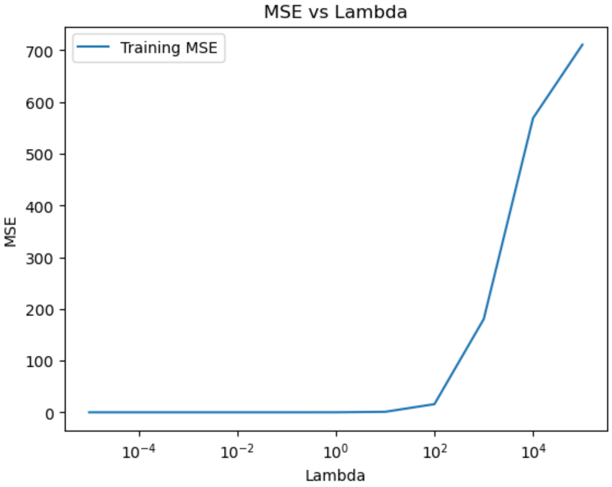

Figure 6: Train MSE vs lambdas(1e-5,..1e5)

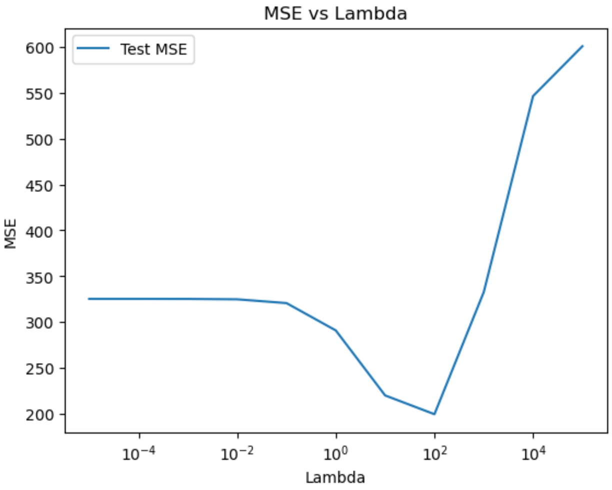

Figure 7: Test MSE vs lambdas(1e-5,..1e5)