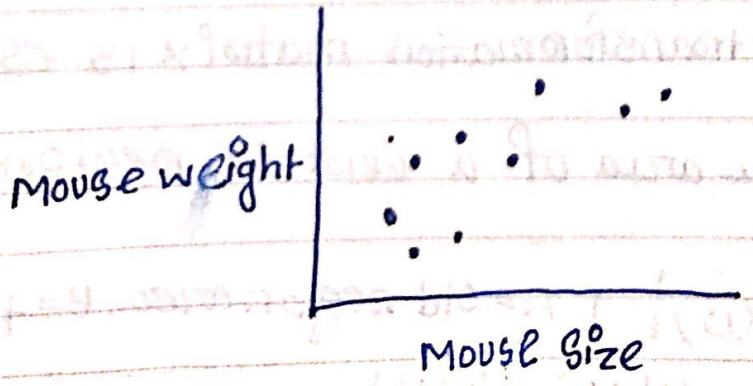
Homework-2

Problem 1

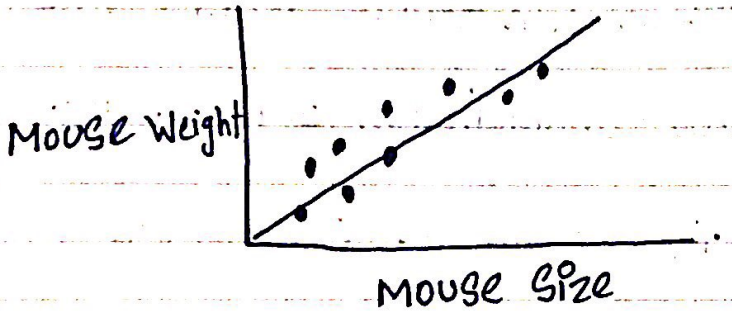
(a) Ridge regression is used when there is not enough data in training set or there is multicollinearity amongst regression predictors variables in a model.

But if there is enough data in the training set and we entirely know about the nature of independent variables in the dataset. Standard linear regression can appropriately predict from the test set.

Let us suppose that we have a decent amount of dataset where we predict mice weight from mice size.

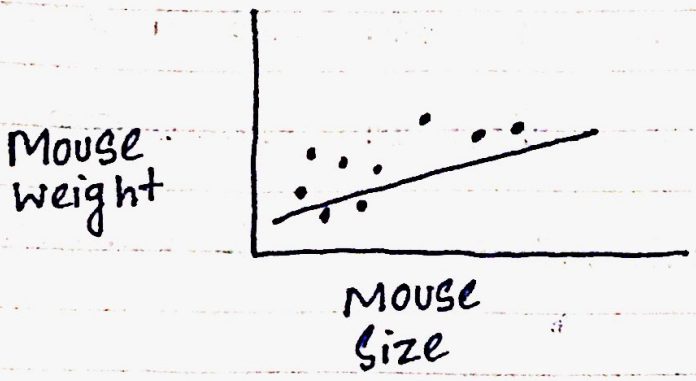


In Standard linear regression, we use AKA least square to model the relationship between weight and size. Least square line accurately reflects the relationship between size and weight.



On the other hand if we use ridge regression in this case it introduces a small amount of bias into how the new line is fit to the data. It returns a small amount of bias and there is a significant drop in variance.

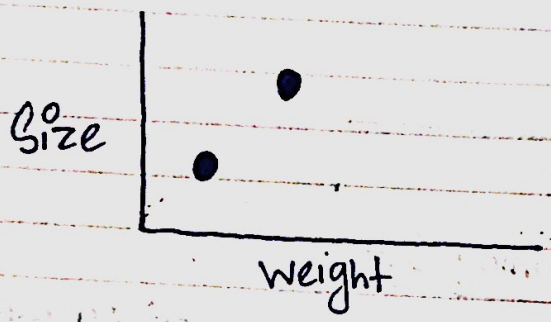
from equation (2), we can see that $\|w\|^2$ add a penalty to the traditional least square method. $(\frac{1}{2})=\lambda$ determines how severe the penalty is. As a result the model finally looks like the following:



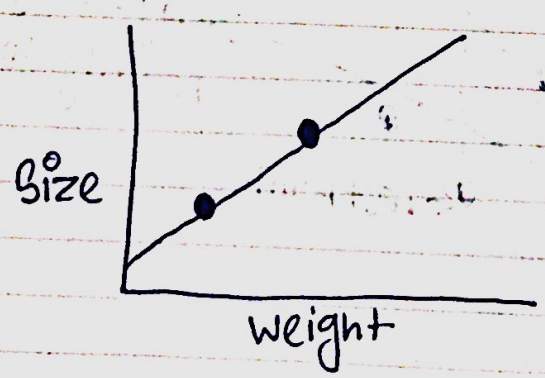
So from the above description we can finally realize that the ridge regression clearly underfits our model, where linear regression fit the model accurately. Ridge regression makes the slope smaller making mouse size less sensitive to mouse weight.

(b)

Let us get back to ovr previous model. This time we just have the two measurements in training set.



When we have a lot of measurements, we can be fairly confident that the least squares line accurately reflects the relationship between size and weight. But as we have only two measurments now we fit a new line with least square.



Here the minimum sum of squared residuals = 0
So, the new line is over fit to the training data.

Now, with ridge regression we introduce a small amount of bias into how the new line is fit to the data

and we get a significant droop in variance.

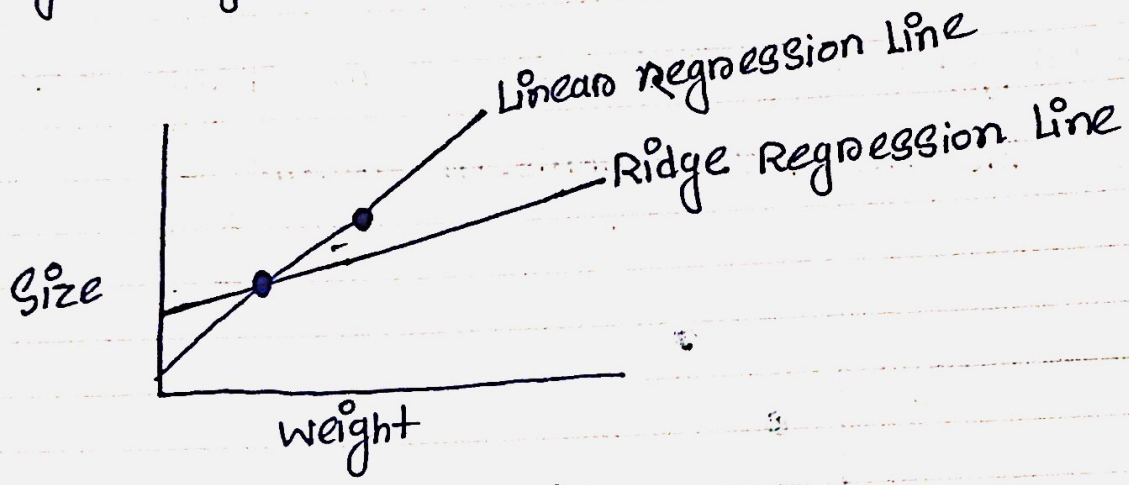


Fig: Training Set

For ridge regression line here, the sum of squared residuals plus the ridge regression penalty is smaller than the sum of the squared residuals of linear regression line.

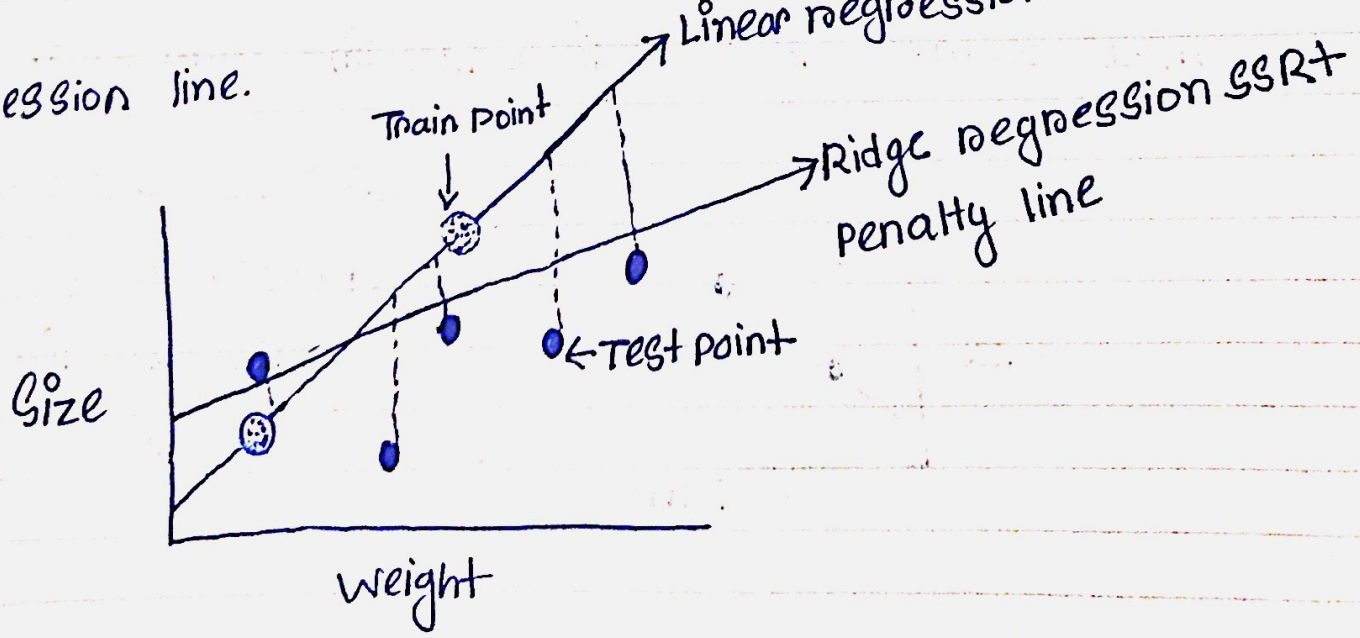
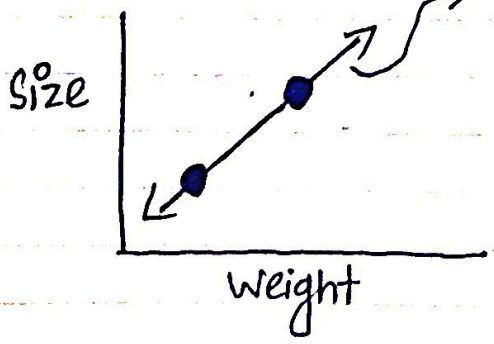


Fig: Test Set

Ridge regression finds new line that doesn't fit the training data by introducing bias with penalizer but in return provides better long term predictions.

(@) The regularization parameter ($\frac{n}{2} = \lambda$) controls the size of coefficient and amount of regularization. Let's get back to our previous model.

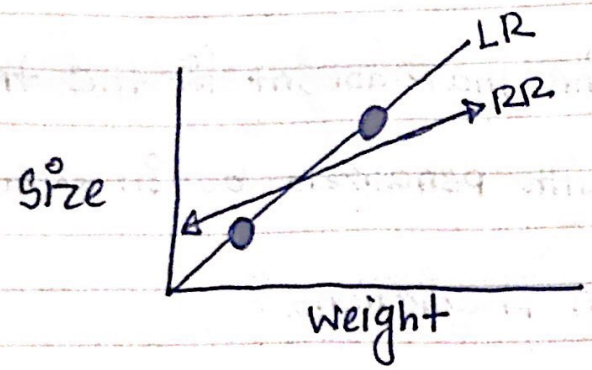
Suppose, $\lambda = 0$, then the ridge regression penalty is also zero. That means the ridge regression line will only minimize the sum of the squared residuals.



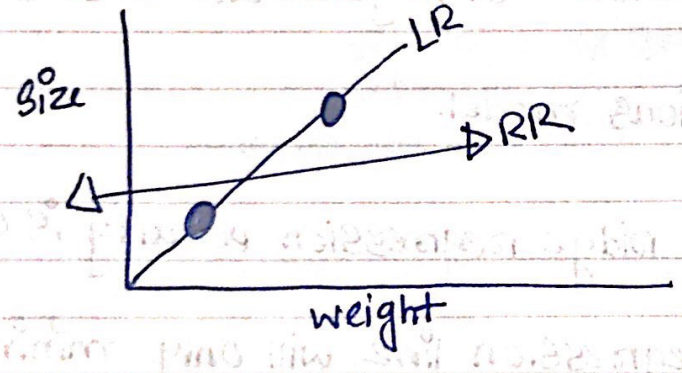
Least square line = ridge regression line

In this case, the bias and variance will be the same as standard linear regression

When $\lambda = 1$, the ridge regression line ended up with a smaller slope than the least square line.



when $\lambda=2$, the slope gets smaller.



The larger we make λ the slope gets asymptotically close to zero. with the increase of λ, our prediction of size become less and less sensitive to weight.

So, from above discussion we find out that, as the model's n, the regularizer gets larger, the model becomes more biased, drops variance and

vice-versa. Summary:

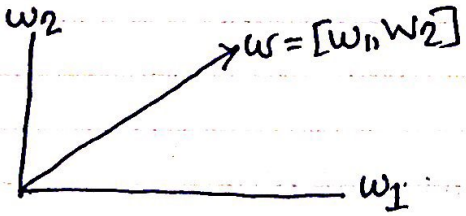
Large λ:
high bias, low variance
Small λ:
low bias, high variance

(c) Given equation:

$$\min_w \|Xw - y\|_2^2 + \frac{\eta}{2}\|w\|_2^2$$

So if we have a vector w in a two dimensional space

the norm of the vector:



Euclidian norm $= \sqrt{w_1^2 + w_2^2}$
$= l_2$ norm

lp norms:

$(w_1^p + w_2^p + \ldots\ldots\ldots + w_d^p)^{\frac{1}{p}}$ [w has d elements]

$= \|w\|_p$

Suppose, $F(w) = \|Xw - Y\|_2^2$

Regularizing linear models:

$F(w) + \lambda \|w\|_2^2 \rightarrow$ l2 regularization.

Ridge regression is l2 norm regularization of linear regression.

$$F(w) = \sum_{i=1}^{N} (y_i - w^T x_i)^2 + \lambda \|w\|_2^2$$

We want to find a w that minimizes $F(w)$, objective function.

$\|w\|_2^2$ can be written as $w^T w$

$$\frac{\partial F(w)}{\partial w} = 0$$

$F(w)$ can also be written as:

$$F(w) = (Y - Xw)^T (Y - Xw) + \lambda w^T w$$

$$\Rightarrow \frac{\partial F(w)}{\partial w} = -2(Y - Xw)X^T + 2\lambda w \qquad \left[\frac{\partial}{\partial x}(x^T x) = 2x\right]$$

Setting $\frac{\partial F(w)}{\partial w} = 0$

$\Rightarrow -2X^T(Y - Xw) + 2\lambda w = 0$

$\Rightarrow w = (X^TX + \lambda I)^{-1} X^TY$

(e) (i) The cost function of linear regression:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

We need to choose θ to minimize $J(\theta)$

Lets suppose we have 4 features. For each observation we will have: x, x^2, x^3, x^4, y. For each x, x^2, x^3, x^4 we will have parameter θ_i

$$h(x) = \theta_0 + \theta_1 * x + \theta_2 * x^2 + \theta_3 * x^3 + \theta_4 x^4$$

And, we have two data points: (2, 3, 0, 4, 1) and (1, 2, 0, 1, 1)

Now for $J(\theta)$ to be zero, there are numerous solution.

After entering two data points:

$$J(\theta) = \frac{1}{2}(\theta_0 + 2\theta_1 + 3\theta_2 + 4\theta_4 - 1)^2 + \frac{1}{2}(\theta_0 + \theta_1 + 2\theta_2 + \theta_4 - 1)^2$$

From above equations we can say, there is no single solutions of those equations.. There will be infinite

number of solutions. So, deriving any generalized solution form is not possible.

When there is multicollinearity in featureset being cost function of linear model is somewhat convex, nothing stops steepest descent algorithm from converging.

For a given dataset, the mean square error of a linear model = $\sigma^2 Tr((X^TX)^{-1})$

Now with multicollinearity X^TX will have almost linearly dependent columns, leading some eigenvalues to be very small.

$$Tr((X^TX)^{-1}) = \frac{1}{\lambda_1(X^TX)} + \dots \frac{1}{\lambda_p(X^TX)} \geq \frac{1}{\lambda_{min}(X^T}$$

So, any closed form solution isn't possible

(ii) Ridge regression can handle efficiently the cases when there are more features than samples. Also, it can handle multicollinear

The closed form of ridge regression:

$$(X^TX + \lambda I)^{-1} X^TY$$

Ridge regression can find a solution with cross validation and the ridge regression penalty that favors smaller parameter values. As it performs L2 regularization, it adds penalty equivalent to square the magnitude of coefficient which minimizes the sum of square of coefficients to reduce the impact of correlated predictors.

Problem 2:

(a) From the 2D scatter plot for each feature, where each point associates with a sample, and the two coordinates are the feature value and the house price of the sample. The top three features that are most correlated:

1. RM (average number of rooms per dwelling)
- has strong positive correlation with price.

2. LSTAT (lowered status of the population)
- has strong negative correlation with price.

3. B (1000 (Bk - 0.63)^2 where Bk is the proportions of blacks by town - has positive correlation with price.

(b) By visualizing the matrix using heatmap we can find top three features:

1. RM - Stroongly & positively corrorelated with price
2. LSTAS - Strongly & negatively corrorelated with price
3. PTRATIO has next stroong negative corrorelation with price

So it is evident that the first two features (RM, LSTAS) also showed stroong corrorelation in 2(a). But we can notice from the heatmap that PTRATIO is negatively corrorelated with house price (corrorelation is bettero than othero features) that was not obseroved in 2(a)