Cross-Validation of OLS and CLS Regression

Matthew Sun

Abstract

The purpose of this assignment was to implement ordinary least squares (OLS) and constrained least squares (CLS)regression on a data set of ten points with two outliers. The CLS constraint was chosen to be 8. 5-fold cross-validation was performed on both regression lines. The training and testing mean errors and standard deviations were found for both models. It was found that the OLS model generally had lower training and testing mean errors, while the CLS model consistently had lower standard deviations. It was also found that the training errors of the CLS model were more representative of the testing errors than the OLS model. Although the OLS model had lower mean errors, it was observed that the CLS model better represented the linear trend present in the non-outlier data points since the OLS model was more biased toward outliers.

1 Introduction

The objective of this assignment was to perform ordinary least squares (OLS) and constrained least squares (CLS) regression on a custom data set. The independent data is a design matrix X where the first column contains integers from 0 to 9 and the second column is a column of 1s. The dependent data was a vector with entries $y_i = ex_{i1} + \pi$ where y_1 was subtracted by 5 and y_{10} had 3 added to it to create two outliers.

1.1 The Lagrange Equation

Given some continuous and differentiable functions $f : \mathbb{R}^n \to \mathbb{R}$ and $p : \mathbb{R}^n \to \mathbb{R}$ where f is the objective function and p is the constraint. For the optimal point $\vec{w}^* \in \mathbb{R}^n$ where f and p intersect only at $\vec{w}^*$, the gradients of each function are scalar multiples of each other

$$\underline{\nabla} f(\vec{w}^*) = -\mu \underline{\nabla} p(\vec{w}^*) \tag{1}$$

The scalar value that relates the two gradients is known as the Lagrange multiplier. The Lagrange equation is formulated by subtracting $-\mu\underline{\nabla}p(\vec{w}^*)$ from both sides and setting the result to 0. This can then be made into an equation with argument $\vec{w}$ and a vector of Lagrange multipliers $\vec{\mu}$ to extend to the case for multiple constraints $p : \mathbb{R}^n \to \mathbb{R}^m$. For any local minimizer $\vec{w}_0$ of f, where $\vec{w}(\vec{w}_0) = 0$, the Lagrange function is defined as

$$\mathcal{L}(\vec{w}, \vec{\mu}) := f(\vec{w}) + \vec{\mu}^T \vec{p}(\vec{w}) \tag{2}$$

The stationary point(s) of the Lagrange function $\vec{w}^*$ and $\vec{\mu}^*$ can be written as

$$\begin{bmatrix} \left[\frac{\partial \mathcal{L}}{\partial \vec{w}}\right]^T \\ \left[\frac{\partial \mathcal{L}}{\partial \vec{\mu}}\right]^T \end{bmatrix} (\vec{w}^*, \vec{\mu}^*) = \vec{0} \tag{3}$$

1.2 The KKT Conditions

The KKT conditions outline a set of conditions that are necessary and sufficient for constrained convex optimization problems with inequality constraint(s) written as $A\vec{w} \leqslant \vec{b}$. They must all be true for a solution to be viable (a KKT point). For a problem with a Lagrange function $\mathcal{L}(\vec{w}, \vec{\lambda}) = f(\vec{w}) + \vec{\lambda}^T[A\vec{w} - \vec{b}]$, a solution $\hat{w}$, and its corresponding Lagrange multiplier $\hat{\lambda}$, the KKT conditions state that

1. The solution must satisfy primal feasibility, meaning it must satisfy the inequality such that

$$A\hat{w} - \vec{b} \leqslant 0 \tag{4}$$

2. The solution must satisfy dual feasibility, meaning that

$$\exists \hat{\lambda} \in \mathbb{R}^m \geqslant 0 \tag{5}$$

3. The solution must satisfy stationarity meaning it must be a stationary point of the Lagrange function.

4. The solution must satisfy complementary slackness, meaning that

$$\hat{\lambda}^T [A\vec{w} - \vec{b}] = 0 \tag{6}$$

The KKT conditions have corresponding geometric interpretations which provide some insight into what the value of a certain $\hat{\lambda}$ means. Primal feasibility simply that the solution must fall within the boundary of the constraint(s). Stationary states that the solution must be as close as possible to the minimizer of the objective function. Complementary slackness states that each constraint must be satisfied as an equality meaning the solution falls exactly on the boundary of a constrained such that $A\vec{w} - \vec{b} = 0$ and thus $\hat{\lambda} \geqslant 0$ or the solution does not lie on the boundary and $\hat{\lambda} = 0$. If the solution is not on the boundary that means the constraint is entirely inactive either because the solution is not within the boundary of the constraint or the solution lies in an interior point of the constraint. These insights can be used to interpret a $\hat{\lambda} = 0$. Lastly, dual feasibility states that a solution must match one of the previously described cases meaning that $\hat{\lambda} < 0$ implies inconsistent constraint(s).

1.3 Comparing OLS and CLS

The scientific question was: "How did the CLS regression on this data set compare with that of the OLS regression in terms of influence by outliers and differences in training and testing error means and standard deviations?" Ten repetitions of 5-fold cross-validation were performed for each regression model and the results were compared based on the mean and standard deviation of the training and testing errors along with how the training and testing statistic differed from one model to the other. The regression models were also plotted along with the data for visual interpretation of the results.

2 Methods

2.1 Ordinary Least Squares (OLS)

Given a full-rank data matrix $X \in \mathbb{R}^{m \times n}$, and an dependent data vector of $\vec{y}$. Consider the goal to find a weight vector $\vec{w}$ such that

$$X\vec{w} \approx \vec{y} \tag{7}$$

This goal can be formulated as an ordinary least squares (OLS) regression problem. The residual error between the approximated values and the dependent variable. This is defined as

$$\vec{r}(\vec{w}) := X\vec{w} - \vec{y} \tag{8}$$

The objective function can be formulated as the squared norm of the residual error vector

$$f(\vec{w}) = ||X\vec{w} - \vec{y}||^2 \tag{9}$$

The solution weight vector $\vec{w}^*$ that minimizes the objective function is defined as

$$\vec{w}^* = \underset{\vec{w} \in \mathbb{R}^n}{\operatorname{argmin}} \ f(\vec{w}) \tag{10}$$

An explicit solution can be found by setting the function to 0 and solving for $\vec{w}^*$

$$0 = ||X\vec{w}^* - \vec{y}||^2 \tag{11}$$

$$0 = [X\vec{w}^* - \vec{y}]^T [X\vec{w}^* - \vec{y}] \tag{12}$$

$$0 = \vec{w}^{*T} [X^T X]\vec{w}^* + [-2\vec{y}^T X]\vec{w}^* \tag{13}$$

$$\vec{w}^* = [X^T X]^{-1} X^T \vec{y} \tag{14}$$

$[X^T X]$ is invertible since X is full rank. The solution vector $\vec{w}^*$ for ordinary least squares will be denoted as $\vec{w}^*_{OLS}$. This solution was implemented the same way in the code. The data used was a 10×2 design matrix where the first column contains the integers from 0 to 9 and the second column contains a column of 1s for the bias terms. The dependent data is a vector with entries $y_i = ex_i + \pi$ where the first and last entries were subtracted by 5 and added 3 respectively for the purpose of creating two outliers. The OLS regression model was put through ten repetitions of 5-fold cross-validation. The results from the OLS model were evaluated based on the training and testing mean and the training and test standard deviation. These statistics and the plotted regression line for the OLS model were compared with that of the CLS model.

2.2 Constrained Least Squares (CLS)

OLS in practice can give solutions that are "nonsensical". This assignment will concern itself with two problems about OLS. The first is that OLS is sensitive to outliers in the data, the influence of which scales quadratically with their deviance from the optimal model. The second is that there is often a large difference between the training and testing errors. Constrained least squares (CLS) attempts to address these by constraining the squared norm of the solution vector $\vec{w}^*$. The constraint value will be denoted as θ where $||\vec{w}^*||^2\theta$. This can be formulated as a Lagrange equation

$$\mathcal{L}(\vec{w}, \lambda) = [X\vec{w} - \vec{y}]^T[X\vec{w} - \vec{y}] + \lambda(||\vec{w}||^2 - \theta) \tag{15}$$

The solution $\vec{w}^*$ and λ^* for this equation must satisfy the KKT conditions which can be explicitly written for this problem as

$$\lambda^* \geqslant 0 \tag{16}$$
$$||\vec{w}||^2 \leqslant \theta \tag{17}$$
$$[\nabla_{\vec{w}}\mathcal{L}]^T = 2X^T[X\vec{w}^* - \vec{y}] + 2\lambda^*\vec{w}^* = \vec{0} \tag{18}$$
$$\lambda(||\vec{w}||^2 - \theta) = 0 \tag{19}$$

If the ordinary least squares solution $\vec{w}^*_{OLS}$ is feasible according to Equation (17), the constraint is not active and $\lambda = 0$. This will also satisfy dual feasibility and complementary slackness. Thus, in such a case, the OLS solution is the CLS solution. If the OLS solution is not feasible, then the equation from the stationary condition in Equation (18) can be used to express $\vec{w}^*$ as a function with λ^* as its input which can be written as

$$2X^T[\vec{X}\vec{w}^* - \vec{y}] + 2\lambda^*\vec{w}^* = 0 \tag{20}$$
$$X^T X\vec{w}^* - X^T\vec{y} + \lambda^* I\vec{w}^* = 0 \tag{21}$$
$$[X^T X + \lambda^* I]\vec{w}^* = X^T\vec{y} \tag{22}$$
$$\vec{w}^* = [X^T X + \lambda^* I]^{-1}X^T\vec{y} \tag{23}$$
$$\vec{w}^*(\lambda^*) = [X^T X + \lambda^* I]^{-1}X^T\vec{y} \tag{24}$$

Now the problem is shifted to find a λ^* that minimizes $\vec{w}^*$. This can be done by defining an intermediate function

$$g(\lambda^*) := ||\vec{w}^*(\lambda^*)||^2 - \theta \tag{25}$$

The value of $g(0) > 0$ since $||\vec{w}^*(\lambda^*)||^2 > \theta$. Consider the limit for some real scalar u

$$\lim_{u\to\infty} g(u) = -\theta \tag{26}$$

The limit will be $-\theta$ since as the uI component will overwhelm $X^T X$ and the inverse of uI will approach 0. This means that there is some λ^* between 0 and ∞ such that $g(\lambda^*) = 0$. The aforementioned equality for $g(\lambda^*)$ was solved in code by using the MATLAB function "fzero()" which solves for the root of any inputted function. It also requires an initial estimate which was selected to be 0. The code also implements a conditional statement before solving $g(\lambda^*) = 0$ that checks if the OLS solution is feasible. If

the OLS solution is not feasible, λ^* will be found and it will be substituted into the function $\vec{w}^*(\lambda^*)$. The data used in CLS is the same as the data used in the OLS problem and the constraint θ was chosen to be 8. The CLS regression model was also put through ten repetitions of 5-fold cross-validation. The results from the CLS model were evaluated based on the training and testing mean and the training and test standard deviation. These statistics and the plotted regression line for the CLS model were compared with that of the OLS model.

3 Results

Table 1: Solution vectors for linear regression of simple data using an ordinary least squares (OLS) solution and a constrained least squares (CLS) solution. Values are reported to four decimal places.

OLS $\vec{w}^*$	**CLS** $\vec{w}^*$
$\begin{bmatrix} 3.1546 \\ 0.97880 \end{bmatrix}$	$\begin{bmatrix} 2.7887 \\ 0.4724 \end{bmatrix}$

Table 2: Statistical summary for linear regression of simple data using an ordinary least squares (OLS) solution and a constrained least squares (CLS) solution. For the trained folds and the tested folds, means and standard deviations are reported separately. Values are reported to four decimal places

	OLS		CLS	
	Train	Test	Train	Test
Mean	1.2712	1.8928	2.7394	2.7821
Std. Dev.	0.0301	0.2188	0.0014	0.0137

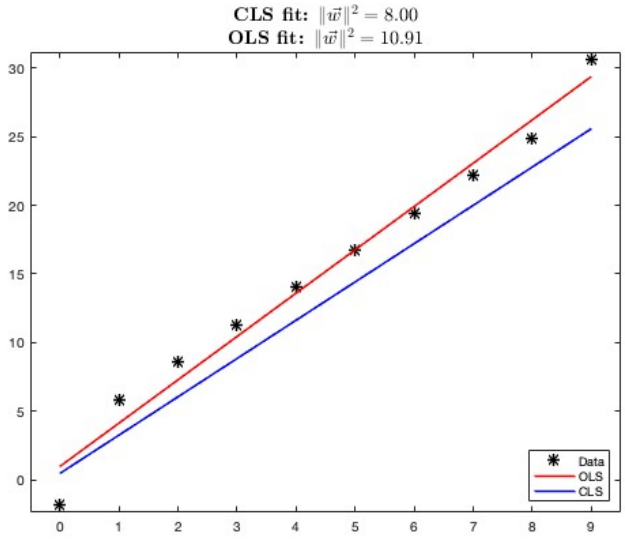

Figure 1: Plots of the OLS (red) regression line and CLS (blue) regression line with a constraint value of 8 alongside a scatter plot of the data.

4 Discussion

For the given data set, the OLS solution vector was [3.1546 ; 0.9780] and the CLS solution vector was [2.7887 ; 0.4724]. The 5-fold cross-validation of the OLS regression model had a mean training error of 1.2712, a training standard deviation of 0.0301, a mean testing error of 1.8928, and a testing standard deviation of 0.2188. The 5-fold cross-validation of the CLS regression model had a mean training error of 2.7394, a training standard deviation of 0.0014, a mean testing error of 2.7821, and a testing standard deviation of 0.0137. Note that since the k-fold cross-validation used randomly shuffles the data, the reported statistics will vary from run to run which should be kept in mind when interpreting the numbers. However, from observation, the difference from run to run was not significant enough to affect the interpretation of the results.

It can be observed from the reported statistics that the OLS model consistently had lower mean training and testing errors than those of the CLS model. So if one were to assess the performance of the model strictly based on mean errors, the OLS model would be considered the better performer. On the other hand, the OLS model had consistently higher training and testing standard deviations than those of the CLS model. This can be interpreted as the CLS model producing errors in the training and testing data that are more closely grouped together near the mean. This provides some strong evidence that the CLS regression generally performs more consistently than the OLS model since this was found for both the training and testing data. The reason for this difference in standard deviation most likely has to do with the OLS model's inclusion (and the CLS model's neglect) of outliers which means that when the model is put through cross-validation where the data is shuffled, this behaviour will lead to more inconsistent errors since the shuffling of the outliers will change the solution vector for OLS more than it does for CLS. It was also observed that for both models, the testing data mean error and standard deviation were consistently higher than those of the training data which is not a notable finding itself. However, the difference between the training and testing statistics for the CLS model was noticeably smaller than the difference for the OLS model. Thus, it can be said that for this problem, the training statistics are much more representative of the testing statistics. For this problem, it seems that CLS was able to address one of the concerns about OLS stated earlier in Section 2.2.

There is a temptation to say that the OLS model is overfitted to the training data compared to the CLS model - which is true if the performance criteria were strictly concerned with differences between training and testing statistics. However, the graphical interpretation of these models plotted beside the data reveals a more nuanced situation. It can be observed that the CLS regression line minimizes its distance from the data better than the CLS line. However, this includes the outliers as well meaning that the OLS line is heavily influenced by the outliers and this sensitivity to outliers does not necessarily mean OLS regression has overfitted to this training data but nonetheless, this still contributes to the robustness and generalizability of the model. On the other hand, the CLS regression line mostly ignores the two outliers, which contributes to its higher mean errors for both the training and testing data. But because of its neglect of the outliers, the CLS regression line fits the trend of the eight data points that are not outliers better than the OLS line which is more desired in most cases. This highlights the limitation of measuring performance using mean errors without accounting for outliers since without looking at the inner workings of OLS and CLS, the lower mean errors of OLS are misleading. The chosen constraint value was $\theta = 8$. It was also found that the squared norm of the solution vector for OLS, to two decimal places, was 10.91. This means that had $\theta \approx 10.91$, the OLS and CLS solution would be the same. Further exploration was done to see how adjusting the constraint such that $0 < \theta < 10.91$ would affect the CLS regression line. It was observed that as θ was decreased from 10.91 to 0.01, the slope of the CLS regression line decreased predictably with each decrease of θ until it was close enough to 0 that the regression line appeared to be flat. $\theta = 8$ seems close to being an optimal constraint value for the CLS regression line, however, it was observed that a slightly higher θ value consistently gave slightly improved training and testing mean errors. It was also observed that the difference between training and testing statistics remained minimal for different θ values. In addition, the standard deviation for different values of θ remained about the same meaning that even for values of θ that perform poorly, the consistency of errors did not change significantly. The differences in the model as θ changed show that there is merit tuning θ to find some optimal constraint value.

The problem of the influence of outliers on models is an almost ubiquitous one in data analysis. A 2013 paper [Wen+13] discusses how OLS's bias toward outliers negatively affects a net-benefit regression model. Similar to this assignment, the authors discuss how OLS can be ineffective because it often fails to capture the linear relationship present in most of the variables even in the presence of just a few outliers. The authors used simulated and real antiplatelet treatment data and applied a variety of regression methods to compare how they handled outliers. They found that the more robust methods were able to perform just as well in terms of accuracy of classification, unlike in this assignment, especially for data with a higher proportion of outliers. This naturally proposes that the problem in this assignment could be expanded to a larger data set with more outliers and see if the difference in the mean errors for OLS and CLS improves.

To conclude, it was found that the OLS regression model performed better in terms of mean training and testing errors. However, the consistency of mean errors, measured by the training and testing standard deviations, was lower for the OLS model (higher standard deviation). It was also found that the training statistics of the OLS model were not as representative of the testing statistics as the CLS model's were. In addition to this, it was found that the OLS model's bias toward outliers negatively impacted its ability to capture the linear relationship in the data that were not outliers, unlike the CLS model that was able to.

References

[Wen+13] Yue Wen et al. "The impact of outliers on Net-Benefit regression model in Cost-Effectiveness Analysis". In: *PLOS ONE* 8.6 (June 2013), e65930. DOI: 10.1371/journal.pone.0065930. URL: https://doi.org/10.1371/journal.pone.0065930.