CISC 371 Class 25

Constrained Least Squares

Texts: [1] pp. 218–222; [2] pp. 305–317

Main Concepts:

- *Ordinary least squares: the normal equation*
- *Constrained least squares: restriction θ on OLS weight vector*
- *Two cases: OLS and scalar optimization of Lagrange multiplier*

Sample Problem, Data Analytics: How can we solve a simple least-squares problem that has a length constraint on the weights?

Ordinary least squares (OLS) is the conventional name in the area of optimization for a problem that is familiar from prerequisite material, where is is often called "linear least squares". The more general approximation problem is: given a set of m independent data vectors $\vec{x}_i$, and m dependent data readings y_i, to find a weight vector $\vec{w}$ that approximates the dependent data values as

$$\vec{x}_i \cdot \vec{w} \approx y_i$$

We can "vectorize" this approximation by gathering the independent data vectors $\vec{x}_i$ into a design matrix X, and gathering the dependent data readings y_i into a data vector $\vec{y}$. When we create the design matrix X, we have two choices: a data vector $\vec{x}_i$ can be the i^{th} column of X, or the transpose of the data vector $\vec{x}_i^T$ can be the i^{th} row of X. We will follow the convention in statistics that uses the second version, so we will define the design matrix as

$$X_{m\times n} \stackrel{\text{def}}{=} \begin{bmatrix} \vec{x}_1^T \\ \vec{x}_2^T \\ \vdots \\ \vec{x}_m^T \end{bmatrix} \tag{25.1}$$

Using Equation 25.1, the problem is to find a weight vector $\vec{w}$ so that

$$X\vec{w} \approx \vec{y} \tag{25.2}$$

We will assume that the data vectors $\vec{x}_i$ either exactly determine or over-determine the weight vector $\vec{w}$. Mathematically, we will assume that the design matrix X is full rank, so $\text{rank}(X) = n$.

In the first part of this class, we will explore the solution of Equation 25.2 as an unconstrained optimization problem. In the second part, we will impose a constraint on the weight vector $\vec{w}$. This constraint will be managed by forming a Lagrange equation and we will consider two specific types of solutions to the problem.

25.1 Ordinary Least Squares

We will simplify the approximation problem of Equation 25.2 by restricting the approximation. The vector of residual errors of the approximation is the difference between the approximated values $X\vec{w}$ and the dependent data readings $\vec{y}$, so

$$\vec{r}(\vec{w}) \stackrel{\text{def}}{=} X\vec{w} - \vec{y} \tag{25.3}$$

The OLS problem is to find the weight vector $\vec{w}^*$ that minimizes the sum of squares of the residual errors that are defined in Equation 25.3, which is

$$\begin{aligned} \vec{w}^* &= \operatorname*{argmin}_{\vec{w}\in\mathbb{R}^n} \sum_{i=1}^{m} (r_i(\vec{w}))^2 \\ &= \operatorname*{argmin}_{\vec{w}\in\mathbb{R}^n} \|\vec{r}(\vec{w})\|^2 \\ &= \operatorname*{argmin}_{\vec{w}\in\mathbb{R}^n} \|X\vec{w} - \vec{y}\|^2 \end{aligned} \tag{25.4}$$

We can formulate the OLS problem in Equation 25.4 by setting an objective function $f(\vec{w})$ to be

$$\begin{aligned} f(\vec{w}) &= \|X\vec{w} - \vec{y}\|^2 \\ &= [X\vec{w} - \vec{y}]^T [X\vec{w} - \vec{y}] \\ &= \vec{w}^T [X^T X]\vec{w} + [-2\vec{y}^T X]\vec{w} \end{aligned} \tag{25.5}$$

In Equation 25.5, we have dropped the $\vec{y}^T\vec{y}$ term because the term is constant with respect to $\vec{w}$. This gives us a quadratic optimization problem that has a single global minimum. The solution to OLS is the unconstrained problem

$$\vec{w}^* = \operatorname*{argmin}_{\vec{w}\in\mathbb{R}^n} f(\vec{w}) \tag{25.6}$$

Equation 25.6 can be solved by taking transposes and equating the stationary point of Equation 25.5 to the zero vector. Doing this gives us the normal equation

$$X^T X\vec{w}^* = X^T\vec{y} \tag{25.7}$$

Because X is assumed to be a full-rank matrix, $X^T X \succ 0$. The explicit solution to Equation 25.6, as a result of Equation 25.7, is

$$\vec{w}^* = [X^T X]^{-1} X^T\vec{y} \tag{25.8}$$

Practical Difficulties in OLS

In a later class, we will explore some of the practical shortcomings of the OLS solution. One foundational paper used the word "nonsensical" to describe the OLS solutions to real data that came from an industrial process.

Our primary concern is that OLS has a sensitivity to statistical outliers, which have quadratically increasing influence as they deviate from the optimal model. A secondary concern is that the error computed in an OLS estimate on training data are often not representative of the error that is subsequently computed on testing data. These concerns can be partially addressed by placing constraints on the solution vector $\vec{w}^*$ of OLS.

25.2 Constrained Least Squares

In constrained least squares (CLS), we place a single inequality constraint on the solution vector. One common constraint is to limit the magnitude of the solution. This requires the user to provide a *threshold value*, which we will write as the symbol θ.

The threshold value θ lets us impose a constraint on the magnitude of $\vec{w}$, such as $\|\vec{w}\|^2 \leq \theta$. Our CLS problem is

$$\begin{aligned} \vec{w}^* &= \underset{\vec{w}\in\mathbb{R}^n}{\operatorname{argmin}} [X\vec{w} - \vec{y}]^T [X\vec{w} - \vec{y}] \\ &= \underset{\vec{w}\in\mathbb{R}^n}{\operatorname{argmin}}\, \vec{w}^T [X^T X]\vec{w} + [-2\vec{y}^T X]\vec{w} \qquad (25.9) \\ \text{such that} \quad \|\vec{w}\|^2 &\leq \theta \end{aligned}$$

The inequality in Equation 25.9 is a quadratic constraint, so it is a convex constraint. CLS, described by Equation 25.9, is therefore a convex problem: it has a convex objective and a single convex inequality constraint.

We can begin our solution by forming the Lagrange function

$$\mathcal{L}(\vec{w}, \lambda) = f(\vec{w}) + \lambda(\|\vec{w}\|^2 - \theta) \qquad (25.10)$$

The KKT conditions require that, for $\vec{w}^*$ to be a solution to Equation 25.9, the Lagrange multiplier is non-negative; mathematically, this is the requirement

$$\lambda^* \geq 0$$

The three other KKT conditions are primal feasibility, stationarity, and complementary slackness:

$$\begin{aligned} \|\vec{w}^*\|^2 &\leq \theta & (25.11)\\ [\underline{\nabla}_{\vec{w}}\mathcal{L}(\vec{w}^*, \lambda^*)]^T = 2X^T[X\vec{w}^* - \vec{y}] + 2\lambda^*\vec{w}^* &= 0 & (25.12)\\ \lambda(\|\vec{w}^*\|^2 - \theta) &= 0 & (25.13) \end{aligned}$$

The condition on complementary slackness is Equation 25.13, has two cases: the OLS solution is feasible or the OLS solution is not feasible. We can introduce a temporary abbreviation for the OLS solution as

$$\vec{w}^*_{LS} = [X^T X]X^T\vec{y}$$

The two cases of complementary slackness, and thus of primal feasibility, are detailed in the extra notes for this class. Our conclusion is that the OLS solution is feasible if and only if $\lambda = 0$.

Comments on CLS

Constrained least squares can be summarized as having this structure:

- If the OLS solution is feasible, then use the OLS solution
- Otherwise, first estimate the Lagrange multiplier λ^* and then estimate the optimal weight vector $\vec{w}^*$

Example: 9 data with 2 outliers

We can test the effect of CLS on a simple data set. Suppose that, for the independent data, we use as x_i the integers from 1 to 9. For dependent data we will use a formula based on Euler's number e:

$$\begin{aligned} y_1 &= e\,x_1 + \pi - 5 \\ y_i &= e\,x_i + \pi \qquad\qquad \text{for } i = 2\ldots 8 \qquad (25.14)\\ y_9 &= e\,x_9 + \pi + 3 \end{aligned}$$

Our model of these data will be the first-order polynomial $X\vec{w} = \vec{c}$, in which X is the Vandermonde matrix, so the model is

$$\begin{bmatrix}\vec{x} & \vec{1}\end{bmatrix}\vec{w} = \vec{c}$$

If we fit a first-order polynomial model to all of the data, using ordinary least squares, we would get the weight vector

$$\vec{w}^*_{LS} \approx \begin{bmatrix} 3.1546 \\ 0.9780 \end{bmatrix}$$

If we use constrained least squares with $\theta = 8$ for a first-order polynomial model to all of the data, we would get the weight vector

$$\vec{w}^*_{CLS} \approx \begin{bmatrix} 2.7887 \\ 0.4724 \end{bmatrix}$$

The RMS errors of these are

$$RMS(\vec{w}^*_{LS}) = 1.3376 \qquad RMS(\vec{w}^*_{CLS}) = 2.7434 \tag{25.15}$$

The results of Equation 25.15 are to be expected, because the OLS is the optimal fit to all of the data and the CLS – which includes a constraint on the norm of the weight vector – must have a greater overall error. The data and fits are graphically illustrated in Figure 25.1.

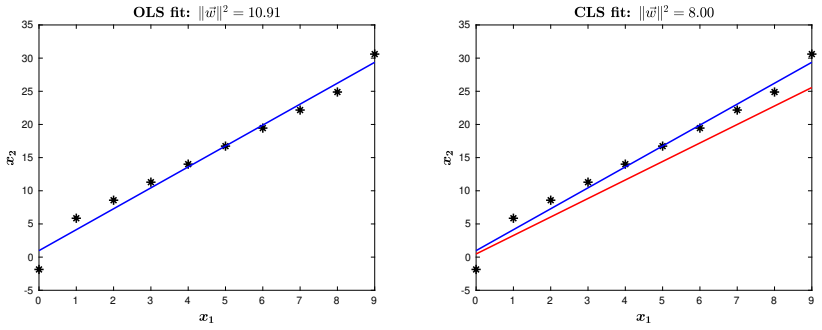

Figure 25.1: Example data and least-squares fits of a first-order polynomial model to the data. (A) The data, as black asterisks, and the ordinary least-squares regression model as a red line. (B) The constrained least-squares solution using $\theta = 8$, shown as a blue line.

The effect of CLS, if needed, is to move the unconstrained OLS solution to satisfy the constraint on the weight vector. The weight vectors for the above example are have two entries, so the weight vectors are plotted in 2D in Figure 25.2. The OLS vector, shown as a red asterisk, is altered to the CLS solution that is indicated by the blue asterisk. We can see that the CLS solution does not simply scale the OLS solution so that the constraint is satisfied.

In the next class, we will briefly explore a motivation for using a constraint that reduces to the problem of CLS, but for which the value λ^* is *provided* rather than being *estimated*.

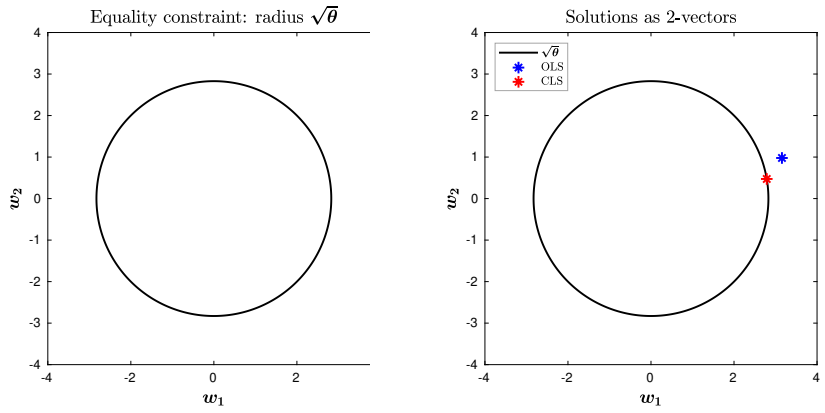

Figure 25.2: Effect of imposing a constraint on a least-squares solution, for 2D data. (A) The equality constraint describes a circle in the vector space. (B)The ordinary least-squares solution produces a weight vector that is show as a red asterisk and the constrained least-squares solution produces a weight vector that is show as a blue asterisk.

25.3 Cross-Validation of Linear Regression

Let us assess these solutions by performing 10 passes of 5-fold cross-validation. We will combine RMS errors by taking the RMS of the RMS errors, which is the square root of the mean variance. When we do this for OLS, we get the data in Table 25.1.

Table 25.1: RMS of fits for the training subset and the testing subset for 5-fold cross-validation of ordinary least squares. The data are from Equation 25.14.

	Train	**Test**
	1.2713	1.8791
	1.2707	1.8738
	1.2629	1.9321
	1.2664	1.9071
	1.2517	2.0417
	1.2742	1.8745
	1.2668	1.9023
	1.2801	1.8134
	1.3084	1.6437
	1.2252	2.2373
Mean:	1.2678	1.9105
Std:	0.0210	0.1523

Next, we will perform the same process using CLS with $\theta = 8$. These data are in Table 25.2.

Table 25.2: RMS of training and testing for 5-fold cross-validation of constrained least squares, using $\theta = 8$. The data are from Equation 25.14.

	Train	**Test**
	2.7394	2.7834
	2.7411	2.7643
	2.7422	2.7546
	2.7418	2.7582
	2.7398	2.7781
	2.7429	2.7481
	2.7421	2.7546
	2.7400	2.7786
	2.7412	2.7666
	2.7388	2.7884
Mean:	2.7409	2.7675
Std.:	0.0014	0.0139

Observations on Table 25.1:

- For OLS, the test errors appear to be substantially greater that the training errors
  - The mean of OLS training fits is low and the variance is high, suggesting a *poor model* of the data
  - The mean of OLS tests is higher than the mean of training, and the variance is also higher, suggesting a *poor model* and *high variance* in tests
- For CLS, the training errors and the test errors appear to be comparable
  - The mean of CLS training errors is higher than the mean of OLS training errors and the variance of CLS training is an order of magnitude less that the OLS training variance, suggesting a *better model* using CLS
  - The mean of CLS tests is comparable to the mean of CLS training, and the variance of CLS tests is an order of magnitude lower than variance of CLS training, suggesting a *better model* and *low variance*

This example suggests that constrained least squares acts to reduce the variance of the tests, at the cost of a lower RMS training error to the given data.

Statistical explanations of an estimator's bias and variance have a long history, so many thorough analyses are available. The interested student is encouraged to explore the relationship between bias, variance, and intrinsic error for a model that includes a random variate with a Gaussian distribution.

25.4 Extra Notes on Constrained Least Squares

As described in the main notes for this class, in CLS either the OLS solution is feasible in the primal formulation or the OLS solution is infeasible. We can reason about these two cases to determine what action we need to take.

Case 1: $\vec{w}^*_{LS}$ is feasible

If the OLS solution is feasible, then the constraint is inactive and we can set the Lagrange multiplier $\lambda = 0$. This is dual feasible and satisfies complementary slackness. Thus, if the OLS solution is feasible, then the Lagrange multiplier $\lambda = 0$. We can also reason that the converse is true.

If $\lambda = 0$, then the complementary slackness constraint of Equation 25.13 reduces to the Lagrange equation for the OLS problem. The solution to the CLS problem would then be the solution to the OLS problem, which is $\vec{w}^*$. If $\vec{w}^*$ satisfies Equation 25.11, then $\vec{w}^*$ is feasible.

Case 2: $\vec{w}^*_{LS}$ is not feasible

We have reasoned, above, that the OLS solution is feasible if and only if $\lambda^* = 0$. The second case is when the OLS solution is not feasible, which must – by our reasoning – be equivalent to $\lambda^* > 0$. We can write Equation 25.12, the constraint on stationarity, in a way that lets us find a solution for $\vec{w}^*$ in terms of λ^*. We can do this as

$$\begin{aligned} & 2X^T[X\vec{w}^* - \vec{y}] + 2\lambda^*\vec{w}^* &=& \ 0 \\ \equiv\ & X^TX\vec{w}^* - X^T\vec{y} + \lambda^* I\vec{w}^* &=& \ 0 \\ \equiv\ & [X^TX + \lambda^* I]\vec{w}^* &=& \ X^T\vec{y} \\ \equiv\ & \vec{w}^* &=& \ \left[X^TX + \lambda^* I\right]^{-1} X^T\vec{y} \\ \equiv\ & \vec{w}^*(\lambda^*) &=& \ \left[X^TX + \lambda^* I\right]^{-1} X^T\vec{y} \end{aligned} \tag{25.16}$$

According to Equation 25.16, the optimal weight vector $\vec{w}^*$ is a function of the optimal Lagrange multiplier λ^*. We can compute λ^* by requiring that both the optimal Lagrange multiplier and the optimal weight vector must satisfy the dual feasibility constraint, which is Equation 25.13. Because we are in the case where $\lambda > 0$, the constraint is satisfied if and only if $\|\vec{w}^*(\lambda^*)\|^2 - \theta = 0$.

We can write a new, temporary, function

$$g(\lambda^*) = \|\vec{w}^*(\lambda^*)\|^2 - \theta = 0 \tag{25.17}$$

We can solve Equation 25.17 numerically using many methods, including the builtin `MATLAB` function `fzero` using the hyper-parameter $\lambda^* = 0$ as an initial estimate. Another method for estimating λ^* is to bracket the root of $g(\lambda^*)$ and use a bisection search for the root of the function.

Using a zero-finding method, a bracketing method, or another appropriate method, from λ^* we can compute $\vec{w}^* = \vec{w}^*(\lambda^*)$. This completes the second case of CLS.

End of Extra Notes

References

[1] Beck A: Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB. Siam Press, 2014

[2] Boyd S, Vandenberghe L: Convex Optimization. Cambridge University Press, 2004