Linear programming: applications in statistics

2012-10-11

Hao Wu

Review

We have covered in previous two class:

- LP problem set up.
- Simplex method.
- Duality.
- Interior point algorithm.

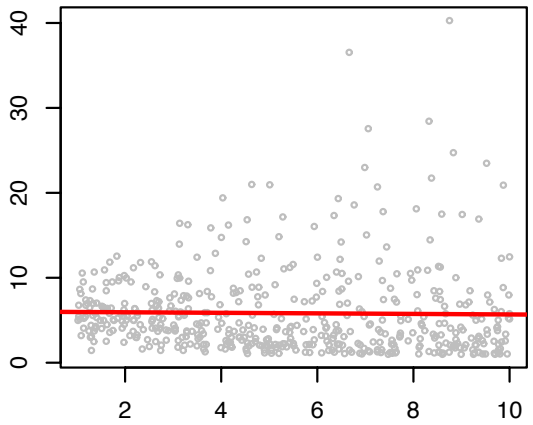
Now you should be able to formulate a LP problem and solve it. But how are these useful in statistics?

- Remember LP is essentially an optimization algorithm.
- There are plenty of optimization problems in statistics, e.g., MLE.
- It's just a matter of formulating the objective function and constraints.

LP in statistics I: quantile regression

Motivation:

- Goal of regression: to tease out the relationship between outcome and covariates. Traditional regression: mean of the outcome depends on covariates.
- Problem: data are not always well-behaved. Are mean regression methods sufficient in all circumstances?



Quantile regression:

- provides a much more exhaustive description of the data.
- The collection of regressions at all quantiles would give a complete picture of outcome-covariate relationships.

Quantile regression model

Regress conditional quantiles of response on the covariates. Assume the outcome Y is absolutely continous and that X is the vector of covariates.

- Classical model: $Q_\tau(Y|X) = X\beta_\tau$
- $Q_\tau(Y|X)$ is the τ^{th} conditional quantile of Y given X.
- β_τ is the parameter of interest.

The above model is equivalent to specifying

$$Y = X\beta_\tau + \epsilon,\ Q_\tau(\epsilon|X) = 0\ .$$

Pros and cons

Advantages:

- Regression at a sequence of quantiles provides a more complete view of data.
- Inference is robust to outliers.
- Estimation is more efficient when residual normality is highly violated.
- Allows interpretation in the outcome's original scale of measurement.

Disadvantages:

- Provides no additional info about covariate effects in the case of location-shift models.
- To be most useful, regress on a set of quantiles: computational burden.
- Solution has no closed form.
- Adaptation to non-continuous outcomes is difficult.

The loss function

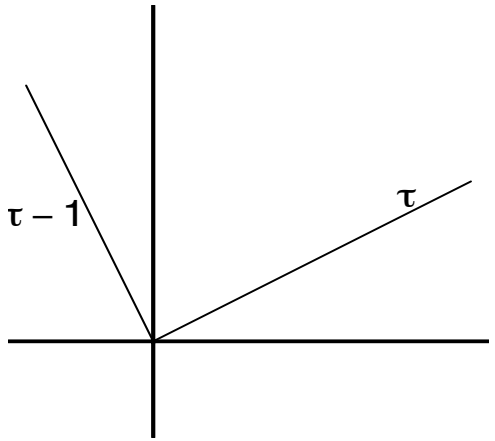
Link between estimands and loss functions.

- To obtain sample mean of $\{y_1, y_2, ..., y_n\}$, minimize $\sum_i (y_i - b)^2$.
- To obtain sample median of $\{y_1, y_2, ..., y_n\}$, minimize $\sum_i |y_i - b|$.

It can be shown that to obtain the sample τ^{th} quantile, one needs to minimize asymmetric absolute loss, that is, compute

$$\hat{Q}_\tau(\mathbf{Y}) = \underset{b}{\text{argmin}} \left\{ \sum_{i:y_i \geq b} \tau |y_i - b| + \sum_{i:y_i < b} (1 - \tau) |y_i - b| \right\} .$$

For convenience, defined $\rho_\tau(x) = x[\tau - 1(x < 0)]$.



Estimator

The classical linear quantile regression model is fitted by determining

$$\hat{\beta}_\tau = \underset{b}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(y_i - x_i b).$$

The estimator have all “expected” properties:

- Scale equivariance:

$$\hat{\beta}_\tau(ay, X) = a\hat{\beta}_\tau(y, X), \quad \hat{\beta}_\tau(-ay, X) = -a\hat{\beta}_{1-\tau}(y, X)$$

- Shift (or regression) equivariance:

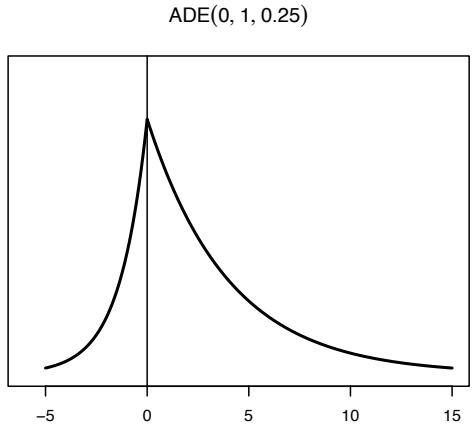
$$\hat{\beta}_\tau(y + X\gamma, X) = \hat{\beta}_\tau(y, X) + \gamma$$

- Equivariance to reparametrization of design:

$$\hat{\beta}_\tau(y, XA) = A^{-1}\hat{\beta}_\tau(y, X)$$

Asymmetric Double Exponential (ADE) distribution

- Least-squares estimator $\Leftrightarrow$ MLE if residuals are normal.
- QR estimator $\Leftrightarrow$ MLE if residuals are ADE.
- Density function for ADE: $f(y;\mu,\sigma,\tau) = \frac{\tau(1-\tau)}{\sigma}\exp\left\{-\rho_\tau\left(\frac{y-\mu}{\sigma}\right)\right\}$.



- If residuals are iid $ADE(0,1,\tau)$, then the log-likelihood for β_τ is

$$\ell(\beta_\tau;\mathbf{Y},\mathbf{X},\tau) = -\sum_{i=1}^{n}\rho_\tau\left(y_i - x_i\beta_\tau\right) + c_0$$

Model fitting

The quantile regression model is to find b to minimize $\sum_i \rho_\tau(y_i - x_i b)$. It can be framed into an LP problem. Define:

$$
\begin{aligned}
u_i &\equiv [y_i - x_i\beta]_+ \\
v_i &\equiv [y_i - x_i\beta]_- \\
b_+ &\equiv [b]_+ \\
b_- &\equiv [b]_-
\end{aligned}
$$

The the MLE problem can be formulated as:

$$
\begin{aligned}
\max \quad & -\sum_{i=1}^{n} [\tau u_i + (1-\tau) v_i] \\
s.t. \quad & y_i = x_i b_+ - x_i b_- + u_i - v_i \\
& u_i, v_i \geq 0, \quad i = 1, \ldots, n
\end{aligned}
$$

This is a standard LP problem can be solved by Simplex/Interior point method.

Written in matrix notation, and make u_i, v_i, b_+, b_- as unknowns, get

$$\begin{aligned} \max \quad & -[\mathbf{0}, \mathbf{0}, \tau, \mathbf{1}-\tau] \begin{bmatrix} \mathbf{b}_+ \\ \mathbf{b}_- \\ \mathbf{u} \\ \mathbf{v} \end{bmatrix} \\ s.t. \quad & [\mathbf{X}, -\mathbf{X}, \mathbf{I}, -\mathbf{I}] \begin{bmatrix} \mathbf{b}_+ \\ \mathbf{b}_- \\ \mathbf{u} \\ \mathbf{v} \end{bmatrix} = \mathbf{y} \\ & \mathbf{b}_+, \mathbf{b}_-, \mathbf{u}, \mathbf{v} \geq 0 \end{aligned}$$

The dual problem is:

$$\begin{aligned} \min \quad & \mathbf{y}^T \mathbf{d} \\ s.t. \quad & \begin{bmatrix} \mathbf{X}^T \\ -\mathbf{X}^T \\ \mathbf{I} \\ -\mathbf{I} \end{bmatrix} \mathbf{d} \geq -\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \tau \\ \mathbf{1}-\tau \end{bmatrix} \\ & \mathbf{d} \text{ is unrestricted} \end{aligned}$$

Manipulating the constraints, get

$$\mathbf{X}^T\mathbf{d} = \mathbf{0}$$
$$-\tau \leq \mathbf{d} \leq 1 - \tau$$

Define a new variable $\mathbf{a} = 1 - \tau - \mathbf{d}$, we get the following LP problem for the original QR problem:

$$\begin{aligned} \max \quad & \mathbf{y}^T\mathbf{a} \\ s.t. \quad & \mathbf{X}^T\mathbf{a} = \mathbf{0} \\ & \mathbf{a} \geq 0 \\ & \mathbf{a} \leq 1 \end{aligned}$$

Adding slack variables $\mathbf{s}$ for the $\leq$ constraints, the problem can be formulated in the standard form, and can be solving using either Simplex or Interior Point methods.

$$\begin{aligned} \max \quad & \mathbf{y}^T\mathbf{a} \\ s.t. \quad & \mathbf{X}^T\mathbf{a} = \mathbf{0} \\ & \mathbf{a} + \mathbf{s} = 1 \\ & \mathbf{a}, \mathbf{s} \geq 0 \end{aligned}$$

Consider usual regression settings with data $(\mathbf{x}^i, y_i)$, where $\mathbf{x}^i = (x_{i1}, \ldots, x_{ip})^T$ is the predictor and y_i is the response for the i^{th} object.
The ordinary linear regression setting:

- Find coefficient to minimize the residual sum of squares:

$$\hat{\beta} = \underset{b}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}^i \mathbf{b})^2 \right\}$$

- This solution happens to be the MLE assuming a normal model:

$$y_i = \mathbf{x}^i \mathbf{b} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

- This is not ideal when the number of predictors (p) are large, because
  1. it requires $p < n$, or there must be some degree of freedoms for error.
  2. you wants a small subset of predictors in the model, but OLS provides an estimated coefficient for each predictor.

The LASSO

LASSO stands for “least absolute shrinkage and selection operator”, which aims for model selection when p is large. It will drive some coefficient to 0, and shrink others.

The LASSO estimates are defined as:

$$\tilde{\beta} = \underset{b}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}^i \mathbf{b})^2 \right\}, s.t. \|\mathbf{b}\|_1 \leq t$$

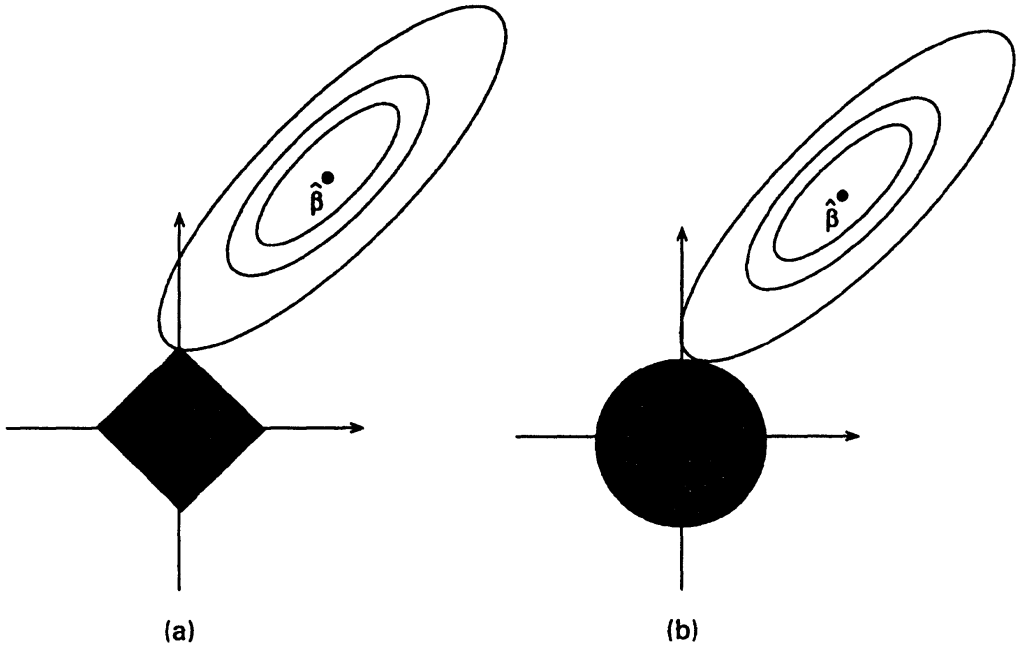
Here $\|\mathbf{b}\|_1 = \sum_j |b_j|$ is the L_1 norm, and $t \geq 0$ is a tuning parameter.

So LASSO tries to minimize the residual sum of square, with a constraint on the sum of the absolute values of the coefficients.

There are other types of “regularized” regressions. For example, regression with an $\boldsymbol{L_2}$ penalty, e.g., the constraint is $\sum_j \beta_j^2 \leq t$, is called “ridge regression”.

The feasible solution space for LASSO is linear (defined by the constraints), so often the optimal solution is at a corner point. The implication: at optimal, many coefficient will be 0 $\Rightarrow$ variable selection.

On the contrary, ridge regression cannot have coefficient being 0.



The LASSO problem can be solved by standard **quadratic programming algorithm**.

We have discussed **linear** programming, where both the objective function and constraints are linear functions of the unknowns.

The **quadratic** programming problem has quadratic objective function and linear constraints:

$$\begin{aligned} \max \quad & f(x) = \frac{1}{2} x^T B x + cx \\ s.t. \quad & Ax \leq b, x \geq 0 \end{aligned}$$

The algorithm for solving QP problem is very similar to that for LP. But first we need to introduce the KKT condition.

KKT conditions

The Karush-Kuhn-Tucker (KKT) conditions is a set of necessary conditions for a solution to be optimal in a general non-linear programming problem.
Consider the following problem :

$$\begin{aligned} \max \quad & f(x) \\ s.t. \quad & g_i(x) \leq 0, i = 1, \ldots, I \\ & h_j(x) = 0, j = 1, \ldots, J \end{aligned}$$

The Lagrangian is: $L(x, \mu, \lambda) = f(x) - \sum_i y_i g_i(x) - \sum_j z_j h_j(x)$. Then at the optimal solution, following **KKT** conditions must be satisfied:

- Primal feasibility: $g_i(x^*) \leq 0, h_j(x^*) = 0$.
- Dual feasibility: $y_i \geq 0$.
- Complementary slackness: $y_i g_i(x^*) = 0$.
- Stationary: $\nabla f(x^*) - \sum_i y_i \nabla g_i(x) - \sum_j z_j \nabla h_j(x) = 0$.

Optimal solution for QP

Following the same procedure, the Lagrangian for the QP problem can be expressed as : $L(x, \mu, \lambda) = \frac{1}{2}x^T Bx + cx - y^T(Ax - b) + z^T x$.

Then the KKT conditions for the QP problem is:

- Primal feasibility: $Ax \leq b, x \geq 0$.
- Dual feasibility: $y \geq 0, z \geq 0$.
- Complementary slackness: $Y(Ax - b) = \mathbf{0}$, $Zx = 0$.
- Stationary: $Bx + c - A^T y + z = 0$.

Y and Z are diagonal matrices with y and z at diagonal.

This can be solved using the interior-point method introduced in last lecture.

To be specific, add slack variable w, the optimality conditions become:

$$\begin{aligned} Ax + w - b &= 0 \\ Bx + c - A^T y + z &= 0 \\ Zx &= 0 \\ Yw &= 0 \\ x, y, z, w &\geq 0 \end{aligned}$$

The unknowns are x, y, z, w. We can then obtain the Jacobians, form the Newton equation and solve for the optimal solution iteratively.

LASSO model fitting

Now go back to LASSO, we need to solve the following optimization problem:

$$\begin{aligned} max \quad & -\sum_{i=1}^{n}(y_i - \sum_j b_j x_j)^2 \\ s.t. \quad & \sum_j |b_j| \leq t \end{aligned}$$

The trick is to convert the problem into the standard QP problem setting, e.g., remove the absolute value operator. The easiest way is to let $b_j = b_j^+ - b_j^-$, where $b_j^+, b_j^- \geq 0$. Then the problem can be written as:

$$\begin{aligned} max \quad & -\sum_{i=1}^{n}(y_i - \sum_j b_j^+ x_j + \sum_j b_j^- x_j)^2 \\ s.t. \quad & \sum_j (b_j^+ - b_j^-) \leq t, \\ & b_j^+, b_j^- \geq 0 \end{aligned}$$

This is a standard QP problem can be solved by standard QP solvers.

A little more on LASSO

The Lagrangian for the LASSO optimization problem is:

$$L(\mathbf{b}, \lambda) = -\sum_{i=1}^{n}(y_i - \sum_j b_j x_j)^2 - \lambda \sum_{j=1}^{p} |b_j|$$

This is equivalent to the likelihood function of a hierarchical model with a double exponential prior on b's:

$$b_j \sim DE(1/\lambda)$$
$$Y|X, \mathbf{b} \sim N(X\mathbf{b}, 1)$$

Here the DE density function is

$$f(x, \tau) = \frac{1}{2\tau} exp\left(\frac{-|x|}{\tau}\right).$$

As a side note, the ridge regression is equivalent with the hierarchical model with a **Normal** prior on b's (verify it).