Optimization in Statistics

The goal in an optimization problem is to find the point at which the minimum (or maximum) of a real, scalar function \( f \) occurs and, usually, to find the value of the function at that point.

We use the term “optimum” or “extremum” to refer to a minimum or maximum. We commonly consider the minimization problem only.

Maximizing \( f(x) \) is equivalent to minimizing its negative, \(-f(x)\).

The general unconstrained optimization problem can be stated as the problem of finding the vector \( x_* \) where the minimum occurs, or of finding the minimum value of the function

\[
\min_x f(x) = f(x_*).
\]

The function \( f \) is called the objective function. The elements of \( x \) are often called decision variables.
Statistical Methods as Optimization Problems

Many common statistical methods are developed as solutions to optimization problems.

In a common method of statistical estimation, we maximize a likelihood, which is a function proportional to a probability density at the point of the observed data.

In another method of estimation and in standard modeling techniques, we minimize a norm of the residuals. The best fit of a model is often defined in terms of a minimum of a norm, such as least squares.
Other uses of optimization in statistical applications occur prior to collection of data, for example, when we design an experiment or a survey so as to minimize experimental or sampling errors.

When a statistical method is based on the solution of an optimization problem, to formulate that problem unambiguously helps us both to understand the method and to decide whether the method is appropriate to the purposes for which it is applied.
Fitting Statistical Models

A common type of model relates one variable to others in a form such as

\[ y = f(x; \theta) + \epsilon, \]

in which \( \theta \) is a vector of fixed parameters with unknown and unobservable values.

*Fitting* the model is mechanically equivalent to *estimating* \( \theta \).

The most familiar form of this model is the linear model

\[ y = x^T \beta + \epsilon, \]

where \( x \) and \( \beta \) are vectors. We also often assume that \( \epsilon \) is a random variable with a normal distribution.

There are two approaches, both optimization problems.

In both, the first step is to replace the fixed unknown parameter with a variable.
Fitted Residuals

For a given value of the variable used in place of the parameter and for each pair of observed values \((y_i, x_i)\), we form a fitted residual. In the case of the linear model, for example, we form the fitted residual

\[ r_i(b) = y_i - x_i^T b, \]

in terms of a variable \(b\) in place of the unknown estimand \(\beta\).

We should note the distinction between the fitted residuals \(r_i(b)\) and the unobservable residuals or errors, \(\epsilon_i = y_i - x_i^T \beta\). When \(\epsilon\) is a random variable, \(\epsilon_i\) is a realization of the random variable, but the fitted residual \(r_i(b)\) is not a realization of \(\epsilon\) (as beginning statistics students sometimes think of it).

Our objective in fitting the model is to choose a value of the variable, which when used in place of the parameter, minimizes the fitted residuals.

We note that the residuals, either fitted or those from the “true” model, are vertical distances.
The idea of minimizing the residuals from the observed data in the model is intuitively appealing.

Because there is a residual at each observation, however, “minimizing the residuals” is not well-defined without additional statements. When there are several things to be minimized, we must decide on some way of combining them into a single measure. It is then the single measure that we seek to minimize.

A useful type of overall measure is a norm of the vector of residuals. The measure, of course, is a function of $b$. The most obvious measure, perhaps, may just be the sum of the absolute values:

$$R_1(b) = \sum_{i=1}^{n} |y_i - x_i^T b|.$$  

This quantity is called the $L_1$ norm of the vector of residuals $r(b)$, and is denoted as $\|r(b)\|_1$. Another possible measure is the sum of the squares:

$$R_2(b) = \sum_{i=1}^{n} |y_i - x_i^T b|^2.$$
This quantity is called the $L_2$ norm of the vector of residuals $r(b)$, and is denoted as $\|r(b)\|_2$. 
For various reasons, the most common approach to fit the model with the given data is *least squares*; that is, to use $R^2(b)$ as the overall measure of the residuals to minimize. With $n$ observations, the ordinary least squares estimator of $\beta$ in the linear model is the solution to the *optimization problem*

$$\min_b \sum_{i=1}^{n} (y_i - x_i^T b)^2.$$  

This optimization problem is relatively simple, and its solution can be expressed in a closed form as a system of linear equations.

Sometimes we may know that $\beta$ must satisfy certain restrictions, and so we modify the optimization problem to impose constraints. For the case of $\beta \geq 0$, for example, we formulate the nonnegative least squares *optimization problem*

$$\min_{b \geq 0} \sum_{i=1}^{n} (y_i - x_i^T b)^2.$$  

This optimization problem is considerably more complicated than the unconstrained problem. Its solution cannot be expressed in a closed form.
Nonlinear Least Squares

When the original model is nonlinear, again, we form a norm of the residuals, but the least squares problem is much more difficult both computationally and conceptually than the linear least squares problem.

$$\min_t \sum_{i=1}^{n} (y_i - f(x_i; t))^2$$

In general, there is no closed-form solution to this optimization problem.
Minimizing Other Functions of the Residuals

For the general objective of minimizing the residuals we have alternatives. We may measure the overall size of the residuals by

$$R_\rho(t) = \sum_{i=1}^{n} \rho(y_i - f(x_i; t)),$$

where $\rho(\cdot)$ is some function of $r_i = y_i - f(x_i; t)$. Instead of minimizing the sum of the squares of the residuals, we fit the model by minimizing this measure; that is, by solving an optimization problem such as

$$\min_t \sum_{i=1}^{n} \rho(y_i - f(x_i; t)).$$

(1)

Depending on $\rho(\cdot)$, this problem is much more difficult both computationally and conceptually than the least squares problem, in which $\rho(r) = r^2$. One common choice of $\rho$ is just the absolute value itself, and the problem of fitting the model is the optimization problem

$$\min_t \sum_{i=1}^{n} |y_i - f(x_i; t)|.$$

(2)
There is no closed-form solution to this simple least-absolute-values problem, even in the linear case.
Treating Residuals Differently

In addition to choosing a function of the individual $r_i$, we might also reconsider how we choose to combine several individual residual values into a single measure. We may want to treat some residuals differently from others, resulting in the optimization problem

$$\min_t \sum_{i=1}^{n} w(y_i, x_i, t) \rho(y_i - f(x_i; t)),$$

where $w(y_i, x_i, t)$ is a nonnegative function. Because in practice, for this minimization problem, it is usually not explicitly a function of $y_i$, $x_i$, and $t$, we often write $w(y_i, x_i, t)$ as a simple fixed weight, $w_i$.

A common instance is the weighted linear least squares problem with fixed weights, in which the function to be minimized is

$$\sum_{i=1}^{n} w_i(y_i - x_i^T b)^2.$$

The weights do not materially change the complexity of this problem. It has a closed-form solution, just as the unweighted (or equally-weighted) problem.
Regularization of the Solution

We may also regularize the minimum residuals problem with additional criteria. We form a weighted linear combination of two functions of $t$,

$$
\sum_{i=1}^{n} w(y_i, x_i, t) \rho(y_i - f(x_i; t)) + \lambda g(t),
$$

where $g$ is some nonnegative function and $\lambda$ is some nonnegative scalar used to tune the optimization problem. The simplest instance of this kind of regularization is in ridge regression with a linear model, in which $w$ is constant, $\rho(z) = z^2$, $f(x_i; b) = x_i^T b$, and $g(b) = b^T b$. The optimization problem is

$$
\min_b \left( \sum_{i=1}^{n} (y_i - x_i^T b)^2 + \lambda b^T b \right).
$$

In ridge regression, we minimize a weighted combination of L$_2$ norms of the vector of residuals, $r(b)$, and of the coefficients, $b$. In lasso
regression with a linear model, an $L_2$ norm is applied to the residuals and an $L_1$ norm is applied to the coefficients, and the *optimization problem* is

$$\min_b (\|r(b)\|_2 + \lambda \|b\|_1).$$
Minimizing Residuals Nonparametrically

There are other ways of approaching the problem of fitting the model.

Instead of fixing the form of the function $f$ and determining a suitable value of $\theta$, we may assume the form of $f$ is unknown (or uninteresting) and approximate it in a way that the approximation $\tilde{f}(x)$ fits the data closely. This kind of approach is nonparametric.

The optimization problem that involves determination of $\tilde{f}(x)$ is a quite different problem from our previous examples. In any case, however, the first step is to be clear about our objective.

Just to minimize some function of the residuals is not sufficient. Unless we add some conditions on $\tilde{f}(x)$, there are infinitely many solutions that yield 0 residuals (assuming no two observations on $x$ have the same value).
Minimizing a Regularized Function of the Residuals

In a nonparametric approach, in addition to a requirement that the residuals be small, we may regularize the problem with other criteria for fitting the model.

For example, we may require that our approximation $\tilde{f}(x)$ be twice-differentiable and be “smooth”. If we measure the roughness or non-smoothness of a twice-differentiable function $f$ over a domain $D$ by the integral of the square of the second derivative,

$$R_{22}(f) = \int_D (f''(x))^2 \, dx,$$

we can include this expression in our optimization problem.

Our overall optimization would be a weighted combination of this expression and some measure of the residuals. In regularized least squares, the **optimization problem** is

$$\min_{\tilde{f}} \sum_{i=1}^{n} (y_i - \tilde{f}(x_i))^2 + \lambda R_{22}(\tilde{f}),$$

with the restriction that $\tilde{f}$ be twice-differentiable. In this formulation, $\lambda$ is a nonnegative smoothing parameter.
Maximum Likelihood Estimation

Another way of fitting the model \( y = f(x; \theta) + \epsilon \) is by maximizing the likelihood function that arises from the probability distribution of \( \epsilon \). Given the data, this is the optimization problem

\[
\max_t \prod_{i=1}^{n} p(y_i - f(x_i; t)),
\]

where \( p(\cdot) \) is the probability function or the probability density function of the random error.

This yields a maximum likelihood estimator (MLE).

Optimization problems of this type can be quite formidable computationally.

For a given probability density \( p(\cdot) \), the maximization problem for MLE may be equivalent to a minimization problem of fitting residuals.
Regularization in Maximum Likelihood Estimation

In a nonparametric formulation of the maximum likelihood approach, we are faced with the same kind of indeterminant problem as in a nonparametric approach to minimizing residuals.

Unless we regularize the problem with additional criteria, the problem is ill-posed. We can “penalize” the likelihood with a regularization measure that decreases (remember we are maximizing) as we move away from the desirable solution.

For example, if we require that the function be smooth (and, hence, twice-differentiable), we may form the optimization problem

$$\max_{\tilde{f}} \prod_{i=1}^{n} p(y_i - \tilde{f}(x_i)) e^{-\lambda \mathcal{R}_{22}(\tilde{f})},$$

where $\mathcal{R}_{22}(\cdot)$ is the same functional as before.
Clustering and Classification

Less formal statistical methods also use optimization. In k-means clustering, for example, we seek a partition of a dataset into a preset number of groups $k$ that minimizes the variation within each group. Each variable may have a different variation, of course. The variation of the $j^{th}$ variable in the $g^{th}$ group is measured by the within sum-of-squares:

$$s_{j(g)}^2 = \frac{\sum_{i=1}^{n_g}(x_{ij(g)} - \bar{x}_{j(g)})^2}{n_g - 1},$$

where $n_g$ is the number of observations in the $g^{th}$ group, and $\bar{x}_{j(g)}$ is the mean of the $j^{th}$ variable in the $g^{th}$ group. For data with $m$ variables there are $m$ such quantities. In k-means clustering the optimization problem is

$$\min \text{ all partitions } \sum_{g=1}^{k} \sum_{j=1}^{m} s_{j(g)}^2.$$

When groups or classes are known, the problem of determining to which group a given observation belongs is called “classification”. In
classification, we seek to determine optimal discriminators to define class membership.
Formulation of an Optimization Problem

The formulation of a statistical problem or any problem in data analysis as an optimization often helps us to understand the problem and to focus our efforts on the relevant aspects of the problem.

In calibration of tables or in data editing, for example, we seek adjustments that represent minimal changes from the original data.

If we do not think clearly about the problem, the resulting optimization problem may not be well-posed; that is, it may not have an unambiguous solution. Functional optimization problems often are not well-posed without the regularization component.

One of the most worrisome problems arises when the optimization problem has multiple points of optimality. The presence of multiple local minima should cause us to think about the problem more deeply.

The objective function should correspond to the objectives of the analysis.
Formulation of an Optimization Problem

In formulating a statistical problem as an optimization problem, we must be careful not to change the statistical objectives. The objective function and the constraints should reflect the desired statistical methodology.

An example in the literature of how available software can cause the analyst to reformulate the objective function began with the problem of fitting a linear regression model with linear constraints; that is, a problem in which the constraints on $b$ were of the form $Ab \leq c$.

It turns out that an optimization problem of least absolute values regression, with $f(x_i; b) = x_i^T b$, that is, linear regression, and with constraints of the form $Ab \leq c$, can be formulated as a linear programming problem with some additional constraints (see Charnes, Cooper, and Ferguson, 1955), and solved easily using available software. At the time, there was no readily available software for constrained least squares regression, so the reformulated problem was solved.
Formulation of an Optimization Problem

The solution to an optimization problem is in some sense “best” for that problem and its objective function. This fact may mean that the solution is considerably less good for some other optimization problem. It is often the case, therefore, that an optimal solution is not robust to assumptions about the phenomenon being studied.

Use of optimization methods is likely to magnify the effects of any assumptions.
Problem 3.27 in HTF

This problem requires familiarity with Lagrange multipliers in constrained optimization.

First of all, the Lagrange multiplier form of the objective function with equality constraints is quite simple.

If the objective function is $f(x)$ and the constraints are $g(x) = 0$, we have the Lagrange multiplier form of the objective function:

$$f(x) + \lambda g(x),$$

where the elements of $\lambda$ are called the Lagrange multipliers.
The so-called Karush-Kuhn-Tucker (KKT) conditions are quite simple.

They merely require that at the solution,

the gradients with respect to both $x$ and $\lambda$ are equal to 0 (this clearly requires that $g(x) = 0$, as required by the constraints),

and that the Hessian matrix with respect to both $x$ and $\lambda$ be either positive definite or negative definite (depending on whether the optimum is a minimum or a maximum).
Now, where do the constraints come from in a lasso problem and how do Lagrange multipliers fit in?

A complication in the lasso problem is that the expression $|\beta|$ is not differentiable.

A common way around this (that is used in least absolute values regression) is to form each element as the sum of two parts, each of which is constrained to be nonnegative: $\beta_j = \beta_j^+ - \beta_j^-$. 

Now we have a constrained optimization problem with objective function

$$f(x) = L(\beta) + \lambda \sum_j (\beta_j^+ - \beta_j^-)$$

and inequality constraints $\beta_j^+, \beta_j^- \geq 0$.

We form the Lagrangian function for this problem, and this is where equation (3.89) comes from.
I would simply call it the Lagrange function (not the Lagrange “dual” function).

In this equation, the Lagrange multipliers are the $\lambda_j^+$ and $\lambda_j^-$, not the $\lambda$, which is a fixed tuning parameter.

The expression itself is quite simple, and more-or-less follows from the simple ideas of Lagrangian functions.

The problem here is inequality constraints — rather than equality constraints, as in the simple case.