Let input \( x \in \mathbb{R}^p \) and output \( y \in \mathbb{R} \). The linear regression model has the form

\[
y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \epsilon,
\]

where the noise \( \mathbb{E}(\epsilon | x) = 0, \mathbb{V}(\epsilon | x) = \sigma^2 \).

Typical goals of linear regression:
- Prediction: given \( x^* \), predict \( y^* \).
- Parameter estimation: find \( \beta \).
- Variable selection: identify variables with \( \beta_j \neq 0 \).

## 1 Ordinary Least Squares

Further assume \( \epsilon \sim N(0, \sigma^2) \). Then \( y|x \sim N(f(x), \sigma^2) \) where

\[
f(x) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j
\]

The conditional log likelihood function is

\[
\ell(\beta, \sigma) = \sum_{i=1}^{n} \log p(y_i|x_i^*) = -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}))^2
\]

Let \( X \) be a \( n \times (p+1) \) augmented feature matrix, with each row a feature vector, and the first column the constant 1 for bias. Let \( y \) be the \( n \)-vector of outputs. The Ordinary Least Squares (OLS) linear regression seeks the \((p+1)\)-vector \( \beta \) (the coefficients) such that

\[
\min_{\beta} (\mathbf{y} - X\beta)^\top (\mathbf{y} - X\beta).
\]

This is the MLE for \( \beta \). Assuming \( X \) has full column rank (which may not be true! Needed for matrix inversion below), there is a closed-form solution

\[
\hat{\beta} = (X^\top X)^{-1}X^\top \mathbf{y}.
\]

The fitted values at the input points are given by

\[
\hat{\mathbf{y}} = X\hat{\beta} = X(X^\top X)^{-1}X^\top \mathbf{y},
\]

where \( X(X^\top X)^{-1}X^\top \) is known as the hat matrix because it operates on \( \mathbf{y} \) to put a hat on it. Typically one estimates the noise variance as

\[
\hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.
\]

We also have

\[
\hat{\beta} \sim N(\beta, (X^\top X)^{-1}\sigma^2).
\]
2 Ridge Regression

Often we regularize the optimization problem. This practice is known as shrinkage in statistics. The classic regularizer is the squared $\ell_2$ norm of $\beta_{-1}$, where $\beta_{-1}$ is the $p$-vector of coefficients by removing the bias coefficient from $\beta$. This results in the familiar ridge regression problem:

$$\min_{\beta} (y - X\beta)^\top (y - X\beta) + \lambda \|\beta_{-1}\|_2^2.$$  \hfill (9)

However, now scaling of $X$ and $y$ matters. Furthermore, having to augment data with a constant feature for bias but then leave its coefficient out of regularization is a bit unwieldy. Therefore, one typically normalizes the data before running regression:

- standardize each feature (mean 0, variance 1);
- center the output by $y_i - \frac{1}{n} \sum_{k=1}^n y_k$

and then work on regression without the bias term ($X$ is $n \times p$, $\beta$ is a $p$-vector). Restating,

$$\min_{\beta} (y - X\beta)^\top (y - X\beta) + \lambda \beta^\top \beta.$$  \hfill (10)

The closed-form solution is

$$\hat{\beta} = (X^\top X + \lambda I)^{-1} X^\top y.$$  \hfill (11)

Unlike OLS, the matrix inversion is always valid for $\lambda > 0$.

3 Lasso Regression

Lasso stands for “Least Absolute Shrinkage and Selection Operator.” It replaces the 2-norm in ridge regression with a 1-norm:

$$\min_{\beta} (y - X\beta)^\top (y - X\beta) + \lambda \|\beta\|_1,$$  \hfill (12)

where $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$.

Lasso is widely regarded as a variable selection method, especially when $p \gg n$: its solution is sparse in the sense that many $\beta_j$’s are zero for large enough $\lambda$.

It is interesting to compare the coefficients of OLS, Ridge, Lasso, and Best subset (finding $k < p$ features whose OLS gives the smallest residual sum of squares, the objective in (4)). In general, these coefficients must be obtain via numerical methods. However, when the input matrix $X$ has orthonormal columns, we have explicit solutions for the $j$-th coefficient:

- OLS: $\hat{\beta}_j$
- Ridge: $\frac{\hat{\beta}_j}{1+\lambda}$
- Lasso: $\text{sign}(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda)_+$
- Best $k$ subset: $\begin{cases} \hat{\beta}_j, & \hat{\beta}_j \text{ among top } k \text{ in magnitude} \\ 0, & \text{otherwise} \end{cases}$

Note that the Lasso coefficients are biased. A standard practice is a two step procedure: first run Lasso to identify features with non-zero coefficients; then run OLS (if possible) on that subset of features to re-estimate the coefficients.

Assume $y = \beta^*^\top x + \epsilon$. Under the so-called compatibility condition, choosing $\lambda$ on the order of $\sqrt{\log(p)/n}$, asymptotically the prediction error

$$\frac{1}{n} \|\hat{\beta}^\top X - \beta^*^\top X\|_2^2 \leq \frac{s_0}{\sigma_0^2} O_P(\log(p)/n),$$  \hfill (13)
the estimation error
\[ \| \hat{\beta} - \beta^* \|_1 \leq \frac{s_0}{\phi_0} O_P(\sqrt{\log(p)/n}), \] (14)
where \( s_0 \) is the number of non-zero entries in \( \beta^* \), and \( \phi_0 \) is a constant in the compatibility condition.

4 Regularization vs. Constraints

The reason why Lasso tends to produce zero coefficients but Ridge does not can be understood by the shape of the \( \ell_q \) ball. The ball is pointy which tends to touch a quadratic (residual) contour at a corner, producing zero coefficients. The \( \ell_1 \) ball starts to be pointy while still be convex. Any \( q < 1 \) are pointy but non-convex. Best subset corresponds to \( q = 0 \), where the \( \ell_0 \)-norm is the cardinality of non-zero coefficients.

5 Elastic Net

If we have two identical and important features \( x_1, x_2 \), the Lasso solution is undetermined: as long as \( \hat{\beta}_1 + \hat{\beta}_2 = c \) and \( |\hat{\beta}_1| + |\hat{\beta}_2| = c' \) there is complete freedom in the values of \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \). In particular, it is possible \( \hat{\beta}_1 = c, \hat{\beta}_2 = 0 \) or vice versa. This is a serious problem from a variable selection perspective.

The effect of elastic net is to identify such groups of variables automatically (without knowing a priori what the groups are). It encourages sparsity at the group level. If a group is determined to be important, elastic net tends to spread the weight evenly to the coefficients in that group. In particular, the elastic net regularizer is

\[ \lambda_1 \| \beta \|_1 + \lambda_2 \| \beta \|_2^2. \] (15)

6 Group Lasso

Elastic net identifies unknown groups of important features automatically. What if we know which features are supposedly in one group, and we wish to control sparsity at the group level? Let the \( p \) features be divided into \( L \) groups, with \( p_l \) the number of features in group \( l \). The group lasso regularizer is

\[ \lambda \sum_{l=1}^{L} \sqrt{p_l} \| \beta_l \|_2. \] (16)

It can be viewed as the 1-norm at the group level, and the 2-norm at the features-in-group level. Notice the 2-norm is not squared.

7 Variable Selection Revisited: Stability Selection

This is a wrapper method, which needs a base learner such as (but not limited to) Lasso. The base learner has a parameter \( \lambda \) which controls sparsity.

- draw a subsample of size \( n/2 \) without replacement.
- run the base learner on the subsample.
- repeat the above many times, and compute the relative selection frequencies

\[ \hat{\Pi}_j^\lambda = \text{fraction of times feature } j \text{ has a non-zero weight, } j = 1 \ldots p \] (17)

- select variables with \( \hat{\Pi}_j^\lambda > \tau \), a threshold.
Consider a base learner which selects \( q \) variables. Denote by \( V \) the number of wrapper-selected variables that actually has \( \beta_j = 0 \), i.e., the number of false positives.

**Theorem 1 (Meinshausen & Bühlmann 2010)** Under appropriate conditions,

\[
E(V) \leq \frac{1}{2\tau - 1} q^2.
\] (18)

## 8 Sparse Classification: Logistic Regression with L1 Regularization

Logistic regression (two class \( y = -1, 1 \)) is defined as

\[
p(y \mid x) = \frac{1}{1 + e^{-y\beta^\top x}}.
\] (19)

Given \((x_1, y_1) \ldots (x_n, y_n)\) the conditional log likelihood is

\[
\ell(\beta) = -\sum_{i=1}^{n} \log(1 + e^{-y_i\beta^\top x_i}).
\] (20)

It is possible to add an L1 regularizer and optimize

\[
\min_{\beta} \sum_{i=1}^{n} \log(1 + e^{-y_i\beta^\top x_i}) + \lambda \|\beta\|_1,
\] (21)

where it is assumed that the data has been normalized and there is no intercept term in \( \beta \).