Regression - II

Prof. Dan A. Simovici

UMB

- When the number n of input variables is large, the assumption previously made concerning the linear independence of the columns $\mathbf{b}^1, \ldots, \mathbf{b}^n$ of the design matrix B may not hold and the rank of B may be smaller than n. In such a case, previous models are not applicable.
- The linear dependencies that may exist between the columns of B (reflecting linear dependencies among experiment variables) invalidate the assumptions previously made. These dependencies are known as *colinearities* among variables.

- One solution is to replace $B'B$ in the least-square estimate $\hat{\mathbf{r}} = (B'B)^{-1}B'\mathbf{y}$ by $B'B + \lambda I_n$ and to define the *ridge regression estimate* as $\mathbf{r}(\lambda) = (B'B + \lambda I_n)^{-1}B'\mathbf{y}$.
- The term *ridge regression* is justified by the fact that the main diagonal in the correlation matrix may be thought of as a ridge.
- We retrieve the ridge regression estimate as a solution of a regularized optimization problem, that is, as an optimization problem where the objective function is modified by adding a term that has an effect the shrinking of regression coefficients.

- Instead of minimizing the function $f(\mathbf{r}) = \| B\mathbf{r} - \mathbf{y} \|_2^2$ we use the objective function

$$g(\mathbf{r}, \lambda) = \| B\mathbf{r} - \mathbf{y} \|_2^2 + \lambda \| \mathbf{r} \|^2 .$$

  This approach is known as *Tikhonov regularization method* and g is known as the *ridge loss function*.
- Ridge regression imposes further constraints on the coefficients r_i by constraing the sum of their squares.

A necessary condition of optimality is $(\nabla g)_{\mathbf{r}} = \mathbf{0}_n$. This yields:

$$\begin{aligned}(\nabla g)_{\mathbf{r}} &= 2B'B\mathbf{r} - 2B'\mathbf{y} + 2\lambda\mathbf{r} \\ &= 2(B'B\mathbf{r} - B'\mathbf{y} + \lambda\mathbf{r}) \\ &= 2[(B'B + \lambda I_n)\mathbf{r} - B'\mathbf{y}] = \mathbf{0}_n,\end{aligned}$$

which yields the previous estimate of $\mathbf{r}$. The ridge estimator is therefore a stationary point of g.

The Hessian of g is the matrix $H_g(\mathbf{x}) = \left(\frac{\partial^2 f}{\partial r_j \, \partial r_k}\right)$, and it is easy to see that

$$H_g(\mathbf{x}) = 2(B'B + \lambda I_n).$$

This implies that H_g is positive definite, hence the stationary point is a minimum.
Note that the ridge loss function is convex, as a sum of two convex functions. Therefore, the stationary point mentioned above is a global minimum of this function.

If B is an unitary matrix (statisticians use the term "orthogonal covariates), we have $B'B = I_n$, so the equality

$$(B'B + \lambda I_n)\mathbf{r} - B'\mathbf{y} = \mathbf{0}_n,$$

implies

$$I_n(1 + \lambda)\mathbf{r} = B'\mathbf{y},$$

hence

$$\| \mathbf{r} \| \leqslant \frac{\| B'\mathbf{y} \|}{n(1 + \lambda)}.$$

Thus, large values of λ tend to control the number non-zero coefficients.

Despite its name *logistic regression* is essentially a classification technique. The term "regression" is justified by the use of a probabilistic approach involving the linear model defined for linear regression. The typical problem involves classifying objects into two classes, designated as C_1 and C_{-1}. Let $\mathbf{s}$ be a data sample of size m, that consists of the pairs of values of a random vector $\mathbf{X}$ ranging over $\mathbb{R}^n$ and a random variable Y ranging over $\{-1, 1\}$.

$$\mathbf{s} = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)),$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_m$ belong to $\mathbb{R}^n$ and $y_i \in \{-1, 1\}$ for $1 \leqslant i \leqslant m$.

In logistic regression we assume that the logarithmic ratio $\ln \frac{P(Y=1|\mathbf{X}=\mathbf{x})}{P(Y=-1|\mathbf{X}=\mathbf{x})}$ is an affine function $r_0 + r_1 x_1 + \cdots + r_n x_n$. If a dummy component x_0 that is set to 1 is added, as we did for linear regression, then the above assumption can be written as

$$\ln \frac{P(Y=1|\mathbf{X}=\mathbf{x})}{P(Y=-1|\mathbf{X}=\mathbf{x})} = \mathbf{r}'\mathbf{x}, \tag{1}$$

where $\mathbf{r}, \mathbf{x} \in \mathbb{R}^{n+1}$.

Let $\ell : (0,1) \longrightarrow \mathbb{R}$ be the *logit function* defined as

$$\ell(p) = \ln \frac{p}{1-p}$$

for $p \in (0,1)$ and let $f : \mathbb{R} \longrightarrow (0,1)$ be the logistic function $L(x) = \frac{e^x}{1+e^x}$. Note that $L(x) + L(-x) = 1$ for $x \in \mathbb{R}$ and the fact that L and ℓ are inverse functions.

Equality (1) can be written as

$$P(Y = 1|X = \mathbf{x}) = \frac{e^{\mathbf{r}'\mathbf{x}}}{1 + e^{\mathbf{r}'\mathbf{x}}} = L(\mathbf{r}'\mathbf{x}),$$

and

$$P(Y = -1|X = \mathbf{x}) = \frac{1}{1 + e^{\mathbf{r}'\mathbf{x}}} = 1 - L(\mathbf{r}'\mathbf{x}) = L(-\mathbf{r}'\mathbf{x}).$$

Both cases are captured by the equality

$$P(Y = y|X = \mathbf{x}) = L(y\mathbf{r}'\mathbf{x}).$$

Equivalently, we have $\ell(P(Y = y|X = \mathbf{x})) = y\mathbf{r}'\mathbf{x}$.

Since the example of $\mathbf{s}$ are independently generated the probability of obtaining the class y_i for each of the examples $\mathbf{x}_i$ is defined by the *likelihood function* $\prod_{i=1}^{m} P(Y = y_i|\mathbf{X} = \mathbf{x}_i)$. To simplify notations we denote this function of y_i and $\mathbf{x}_i$ as $\prod_{i=1}^{m} P(y_i|\mathbf{x}_i)$. Maximizing this function is equivalent to minimizing

$$\begin{aligned}\Lambda(\mathbf{r}) &= -\frac{1}{m}\ln\left(\prod_{i=1}^{m} P(y_i|\mathbf{x}_i)\right) = -\frac{1}{m}\sum_{i=1}^{m}\ln P(y_i|\mathbf{x}_i) \\ &= -\frac{1}{m}\sum_{i=1}^{m}\ln L(y_i\mathbf{r}\mathbf{x}_i) = \frac{1}{m}\sum_{i=1}^{m}\ln\frac{1}{L(y_i\mathbf{r}\mathbf{x}_i)} = \frac{1}{m}\sum_{i=1}^{m}\ln(1+e^{-y_i\mathbf{r}\mathbf{x}_i}),\end{aligned}$$

with respect to $\mathbf{r}$. Note that small values of this expression can be obtained when $y_i\mathbf{r}'\mathbf{x}_i$ is large, that is, when $\mathbf{r}'\mathbf{x}_i$ has the same sign as y_i.

To minimize $\Lambda(\mathbf{r})$ we need to impose the conditions $\frac{\partial \Lambda}{\partial r_j} = 0$ for $1 \leqslant j \leqslant n+1$, which amount to

$$\sum_{i=1}^{m} L'(y_i \mathbf{r} \mathbf{x}_i) \frac{\partial (y_i \mathbf{r} \mathbf{x}_i)}{\partial r_j} = 0,$$

or

$$\sum_{i=1}^{m} L(y_i \mathbf{r} \mathbf{x}_i)(1 - L(y_i \mathbf{r} \mathbf{x}_i)) y_j x_{ji} = 0,$$

for $1 \leqslant j \leqslant n+1$. This is a non-linear system in $\mathbf{r}$ which can be solved by approximation methods.