Linear Methods for Regression

Juan Carlos Calvo Jackson

November 10, 2017

Abstract

This is an overview of Chapter 3 from *The Elements of Statistical Learning* [1]. For an application of this methods, please see this example.

1 Linear Regression Model and Least Squares

The linear regression model has the form

$$f(X) = X^T\beta. \tag{1}$$

By abuse of notation X is the imput vector whose first component is 1, to account for an intercept, and has p other components.

Let $\mathbf{X}$ denote the $N \times (p+1)$ matrix whose rows are N data entries for the imput vector above. Similarly, let $\mathbf{y}$ denote the N-vector of corresponding outputs.

The game is the following, given the data $(\mathbf{X}, \mathbf{y})$ we want to find an estimate $\hat{\beta}$ of β such that $\mathbf{X}\hat{\beta}$ *approximates* $\mathbf{y}$. By *approximation*, we will mean close respect to the distance squared.

Let's define the residual sum of squares as

$$RSS(\beta) = ||\mathbf{y} - \mathbf{X}\beta||^2.$$

We say that a least squares estimator of β, denoted $\hat{\beta}$, is a value of β that minimizes the residual sum of squares. It's easy to see then, that

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Basically, the components of $\hat{\beta}$ are the coefficients of the $(p+1)$ columns of $\mathbf{X}$ in $\mathbb{R}^N$ so that $\hat{\mathbf{y}} := \mathbf{X}\hat{\beta}$ is the proyection of $\mathbf{y}$ onto the hyperplane spanned by the columns of $\mathbf{X}$.

We now assume that at each X, (1) is the correct model of the mean of the ouput Y, i.e., $f(X) = E(Y|X)$; we also assume that deviations from this mean are additive and gaussian with standar deviation σ

$$Y \sim E(Y|X) + \epsilon$$

with $\epsilon \sim \mathcal{N}(0, \sigma)$.

Assuming the $\mathbf{X}$ fixed (non-random), we have

$$\mathrm{Var}(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2.$$

Therefore,

$$\hat{\sigma}^2 = \frac{||\mathbf{y} - \hat{\mathbf{y}}||^2}{N-p-1}$$

implies that $E(\hat{\sigma}^2) = \sigma^2$, which is to say, $\hat{\sigma}^2$ is an umbiased estimator of σ^2.

We also get

$$\hat{\beta} \sim \mathcal{N}(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2) \tag{2}$$

and

$$(N-p-1)\hat{\sigma}^2 \sim \sigma^2\chi^2_{N-p-1},$$

a chi-squared distribution with $N-p-1$ degrees of freedom. Note that $\hat{\beta}$ and $\hat{\sigma}^2$ are statistically independent.

We now use these distributions to estimate confidence intervals and test hypothesis of β.

Under the hypothesis that $\beta_j = 0$ for some $j = 0, \dots, p$,

$$z_j := \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}} \sim t_{N-p-1},$$

where t_{N-p-1} is a t distribution with $(N-p-1)$ degrees of freedom and $v_j = \mathrm{diag}(\mathbf{X}^T\mathbf{X})^{-1}_j$. Hence, a large value of z_j will lead to arejection of this hypothesis.

To test the hypothesis of setting to zero k coefficients, we define

$$F = \frac{(RSS_0 - RSS_1)/k}{RSS_1/(N-p-1)},$$

where RSS_0 is the residual sum of squares for the least squares fit of the smaller model, and RSS_1 the one for the original model. Under the Gaussian assumption, and the null hypothesis,

$$F \sim F_{k,N-p-1}.$$

Note that for $k=1$ $F_{1,N-p-1} = t_{N-p-1}$.

An estimation of the confidence interval for β_j is easily obtained from (2) as

$$(\hat{\beta}_j - z^{(1-\alpha)}\sqrt{v_j}\hat{\sigma}, \hat{\beta}_j + z^{(1-\alpha)}\sqrt{v_j}\hat{\sigma}).$$

Here, $z^{(1-\alpha)}$ is the $1-\alpha$ percentile of the standard normal distribution.

2 The Gauss-Markov Theorem

Let's consider the linear combination $\theta = a^T\beta$, e.g., a prediction $x_0^T\beta$. It's least squares estimate is

$$\hat{\theta} = a^T\hat{\beta} = a^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

For fixed $\mathbf{X}$ this is linear in $\mathbf{y}$. If we assume the linear model to be correct, then

$$E(a^T\hat{\beta}) = a^T\beta$$

and thus $a^T\hat{\beta}$ is an unbiased estimator.

The Gauss-Markov theorem states that for any other linear unbiased estimator $\hat{\theta} = c^T\mathbf{y}$,

$$\text{Var}(a^T\hat{\beta}) \leq \text{Var}(c^T\mathbf{y}).$$

For the proof is enough to note that

$$\begin{aligned} E((c^T\mathbf{y} - a^T\hat{\beta})a^T\hat{\beta}) &= a^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^Tc - a^T(\mathbf{X}^T\mathbf{X})^{-1}a \\ &= a^T(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^Tc - a) \\ &= 0 \end{aligned}$$

as $0 = E(c^Ty - a^T\beta) = (c^T\mathbf{X} - a^T)\beta \quad \forall\beta$. This means that $(c^T\mathbf{y} - a^T\hat{\beta}) \perp a^T\hat{\beta}$ and hence by Pitagoras,

$$\begin{aligned} E((c^Ty)^2) &= E((c^T\mathbf{y} - a^T\hat{\beta})^2) + E((a^T\hat{\beta})^2) \\ &\Rightarrow \text{Var}(a^T\hat{\beta}) \leq \text{Var}(c^Ty). \end{aligned}$$

In general, for an estimator $\tilde{\theta}$ of θ, its mean squared error satisfies

$$MSE(\tilde{\theta}) = \text{Var}(\tilde{\theta}) + (E(\tilde{\theta} - \theta))^2$$

so for among the unbiased estimators, the mean squared estimator gives the one with the smallest mean squared error. However, thre might be a biased estimator with smaller mean squared error.

The mean squared error is intimately related to the rpediction accuracy. Let $Y_0 = f(x_0) + \epsilon_0$, be the response at x_0. Then, the espected error of an estimate $\tilde{f}(x_0) = x_0\tilde{\beta}$ is

$$E(Y_0 - \tilde{f}(x_0)) = \sigma^2 + MSE(\tilde{f}(x_0)).$$

Hence, they only differ by the constant σ^2.

3 Shrinkage Methods

Consider the estimate

$$\tilde{\beta} = \arg\min_{\beta_0,\beta} \left(\mathbf{y} - \beta_0\mathbf{1} - \mathbf{X}\beta)^2 + \lambda|\beta|^q\right), \tag{3}$$

where now $\mathbf{X}$ is a $N \times p$ matrix (no column of ones), β is a p-vector, and $|\beta|^q := \sum_{j=1}^p |\beta_j|^q$. It can be shown that the solution to (3) is equivalent to set $\beta_0 = \bar{y}$ and solve,

$$\tilde{\beta} = \arg\min_{\beta} \left(||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda|\beta|^q\right). \tag{4}$$

where both $\mathbf{X}$ and $\mathbf{y}$ have been reparametrized so as to have zero mean, i.e., their components have been replaced by $x_{ij} - \bar{x}_{.j}$ and $y_i - \bar{y}$.

Note that (4) is equivalent to

$$\tilde{\beta} = \arg\min_{\beta} \left(||\mathbf{y} - \mathbf{X}\beta||^2 \right)$$
$$\text{subject to } |\beta|^q \leq t.$$

For $q = 0$ this is just subset selection, it imposes a penalty on the number of parameters. For $q = 1$ (Lasso), if $t \geq |\hat{\beta}^{\text{ls}}|$, then the estimator is just least squares. For smaller t it enforzes the average shrinkage of coefficients in a linear manner. Due to β being non-smooth, some coefficients of the estimator can become exactly zero as t gets smaller. If $q = 2$ (Ridge), it can be solved exactly, giving

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X} + \lambda\,\mathrm{id})^{-1}\mathbf{X}\mathbf{y}.$$

A *singular value decomposition* (SVD) of $\mathbf{X}$ reveals that $\hat{\beta}^{\text{ridge}}$ favours a shrinkage of the coefficients of the previctors with the lowest variance compared to the high-variance ones.

References

[1] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.