Perceptrons

- linear discriminant functions for classification

. linear discriminant functions:

$$g(\underline{x}) = w_0 + \sum_{i=1}^{d} w_i x_i$$

Let

$\underline{x} = [1, x_1, \cdots, x_d]^T$ and $\underline{w} = [w_0, w_1, \cdots, w_d]^T$.

Then,

$g(\underline{x}) = \underline{w}^T \underline{x}$.

. binary classification:

two classes, c_1 and c_2.

if $\underline{w}^T \underline{x}_i > 0$, output of classifier = c_1

otherwise, output of classifier = c_2

if $\underline{x}_i$ belongs to c_1, $\underline{w}^T \underline{x}_i > 0$.

if $\underline{x}_i$ belongs to c_2, $\underline{w}^T \underline{x}_i < 0$ or $\underline{w}^T(-\underline{x}_i) > 0$.

So, let's change $\underline{x}_i$ into $-\underline{x}_i$, if $\underline{x}_i$ belongs to c_2.

Then,

$\underline{w}^T \underline{x}_i > 0$ if $\underline{x}_i$s are classified correctly.

. Perceptron criterion function:

$$J_P(\underline{w}) = \sum_{\underline{x} \in X(\underline{w})} -\underline{w}^T \underline{x}$$

where

$X(\underline{w})$ represents a set of samples misclassified by $\underline{w}$.

. gradient descent algorithm

The weights are updated as follows:

$$\underline{w}_{k+1} = \underline{w}_k - \rho_k \nabla J_P(\underline{w}_k)$$

where ρ_k represents a positive scaling factor (learning rate) and

$$\nabla J_P(\underline{w}_k) = \frac{\partial J_P}{\partial \underline{w}}|_{\underline{w}=\underline{w}_k}.$$

(1) batch mode

$$\underline{w}_{k+1} = \underline{w}_k - \rho_k \sum_{\underline{x} \in X_k} \underline{x}$$

where X_k represents a set of samples misclassified by $\underline{w}_k$.

(2) on-line mode (or incremental mode)

Training samples $\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_n$ are cyclically applied.

Training Rule:

1) arbitrary set w_1

2) $\underline{w}_{k+1} = \underline{w}_k - \rho_k \underline{x}_k$ for $k \geqq 1$

where $\underline{w}_k^T \underline{x}_k \leqq 0$

cf. $\rho_k = 1$ in Rosenblatt's Perceptron.

. convergence of on-line mode

Let w^* is a solution vector. Then,

$w_{k+1} - w^* = w_k - w^* + \rho_k x_k$ and

$$\| w_{k+1} - w^* \|^2 = \| w_k - w^* \|^2 + 2\rho_k (w_k - w^*)^T x_k + \rho_k^2 \| x_k \|^2.$$

Since $w_k^T x_k \leqq 0$,

$$\| w_{k+1} - w^* \|^2 \leqq \| w_k - w^* \|^2 - 2\rho_k w^{*T} x_k + \rho_k^2 \| x_k \|^2.$$

The term $-2\rho_k w^{*T} x_k + \rho_k^2 \| x_k \|^2$ is negative when

$0 < \rho_k < \dfrac{2 w_k^T x_k}{\| x_k \|^2}$. (convergence condition for ρ_k)

Therefore, the optimal learning rate is $\rho_k^* = \dfrac{w_k^{*T} x_k}{\| x_k \|^2}$.

By substituting the optimal learning rate, we get

$$\| w_{k+1} - w^* \|^2 \leqq \| w_k - w^* \|^2 - \frac{(w^{*T} x_k)^2}{\| x_k \|^2}.$$

. The upper bound of the number of updates

Let $\beta^2 = \max_i \| x_i \|^2$ and $\gamma = \min_i w^{*T} x_i$.

Then, after k updates,

$$\| w_{k+1} - w^* \|^2 \leqq \| w_1 - w^* \|^2 - k\frac{\gamma^2}{\beta^2}.$$

If $w_1 = 0$, k can be bounded by

$$k_0 = \frac{\beta^2 \| w^* \|^2}{\gamma^2}.$$

That is, Perceptron converges within the finite number of steps.

. adaptation of ρ_k

It can be shown if the samples are linearly separable and ρ_k satisfies the following conditions:

$$\lim_{k \to \infty} \rho_k = 0,$$

$$\lim_{m \to \infty} \sum_{k=1}^{m} \rho_k = \infty, \text{ and } \lim_{m \to \infty} \sum_{k=1}^{\infty} \rho_k^2 < \infty,$$

w_k converges to a solution vector satisfying $w^T x_i > 0 \quad \forall i$.

example.

$$\rho_k = \frac{1}{k}$$

. optimal choice of ρ_k

The weight update rule:

$$w_{k+1} = w_k - \rho_k \nabla J(w_k).$$

Taylor series expansion of $J(w)$ around w_k:

$$J(w) \approx J(w_k) + \nabla J^T (w - w_k) + \frac{1}{2}(w - w_k)^T H (w - w_k)$$

where $H = \frac{\partial^2 J}{\partial w_i \partial w_j}|_{w = w_k}$.

Then, $J(w_{k+1}) \approx J(w_k) - \rho_k \| \nabla J \|^2 + \frac{1}{2}\rho_k^2 \nabla J^T H \nabla J$.

Therefore, the optimal choice of ρ_k is $\rho_k = \frac{\| \nabla J \|^2}{\nabla J^T H \nabla J}$.

Reference: Pattern Classification, and Scene Analysis, chapter 5.

Chapter 9. Regression

a linear regression equation:

$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + e$ where

Y = a response variable (dependent variable)

$x_1, \cdots, x_r$ = input variables (independent variables)

$\beta_0, \beta_1, \cdots, \beta_r$ = regression coefficients

e = the random error assumed to be a random variable having mean zero.

If $r = 1$, a linear regression equation is called a simple regression equation, and

if $r > 1$, a linear regression equation is called a multiple regression equation.

Another way of expressing a linear regression equation:

$$E[Y|\, x] = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r$$

where $x = (x_1, \cdots, x_r)$ is the set of independent variables, and $E[Y|\, x]$ is the expected response given the inputs x.

- least squares estimators of the regression parameters

Suppose that the responses Y_i corresponding to the input values x_i, $i = 1, \cdots, n$ are to be observed and consider the following simple linear regression equation:

$$Y_i = \alpha + \beta x_i + e \quad \text{for } i = 1, \cdots, n$$

Let A and B are the estimators of α and β respectively.

Then, the sum of squared differences between the estimated responses and the actual response values - called it SS - given by

$$SS = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$$

the least square error estimator:

$$\frac{\partial SS}{\partial A} = -2\sum_{i=1}^{n} (Y_i - A - Bx_i) = 0,$$

$$\frac{\partial SS}{\partial B} = -2\sum_{i=1}^{n} x_i (Y_i - A - Bx_i) = 0$$

$\rightarrow \quad A = \overline{Y} - B\bar{x}$ and

$$B = \frac{\sum_{i=1}^{n} x_i Y_i - n\bar{x}\overline{Y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})\, Y_i}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

- distribution of the estimators

Let us assume that

$$e \sim N(0, \sigma^2).$$

Then,

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2).$$

the mean and variance of B:

$$E[B] = \frac{\sum_{i=1}^{n}(x_i - \bar{x})E[Y_i]}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\alpha + \beta x_i)}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$= \frac{\alpha\sum_{i=1}^{n}(x_i - \bar{x}) + \beta\sum_{i=1}^{n} x_i(x_i - \bar{x})}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \beta$$

that is, B is an unbiased estimator of β.

$$Var(B) = \frac{Var(\sum_{i=1}^{n}(x_i - \bar{x})Y_i)}{(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2)^2}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 Var(Y_i)}{(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2)^2}$$

$$= \frac{\sigma^2\sum_{i=1}^{n}(x_i - \bar{x})^2}{(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2)^2}$$

$$= \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

the mean and variance of A:

$$A = \frac{1}{n}\sum_{i=1}^{n} Y_i - B\bar{x}$$

$$E[A] = \frac{1}{n}\sum_{i=1}^{n} E[Y_i] - \bar{x}E[B] = \frac{1}{n}\sum_{i=1}^{n}(\alpha + \beta x_i) - \bar{x}\beta = \alpha$$

that is, A is also an unbiased estimator α.

$$Var(A) = \frac{\sigma^2\sum_{i=1}^{n} x_i^2}{n(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2)}$$

residuals: $Y_i - A - Bx_i, \quad i = 1, \cdots, n$

the sum of squares of the residuals:

$$SS_R = \sum_{i=1}^{n}(Y_i - A - Bx_i)^2$$

It can be shown that

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2}$$

because

$\sum_{i=1}^{n}\frac{(Y_i - E[Y_i])^2}{Var(Y_i)} = \sum_{i=1}^{n}\frac{(Y_i - \alpha - \beta x_i)^2}{\sigma^2} \sim \chi^2_n$ and

if we substitute the estimators A and B for α and β, then 2 degrees of freedom are lost.

$$E[\frac{SS_R}{\sigma^2}] = n - 2 \quad \text{or} \quad E[\frac{SS_R}{n-2}] = \sigma^2$$

Thus, $SS_R/(n-2)$ is an unbiased estimator of σ^2.

In addition, SS_R is independent of the pair A and B.

some notations:

$$S_{xY} = \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2$$

Then, the least square estimators can be expressed as

$$B = \frac{S_{xY}}{S_{xx}} \quad \text{and} \quad A = \overline{Y} - B\bar{x}.$$

Also note that

$$B \sim N(\beta, \frac{\sigma^2}{S_{xx}}) \quad \text{and} \quad A \sim N(\alpha, \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{nS_{xx}})$$

The sum of squares of the residuals can be expressed as

$$SS_R = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}}$$

- statistical inferences about the regression parameters

(1) inferences concerning β

the simple linear regression model:

$$Y = \alpha + \beta x + e$$

the problem of testing

$$H_0 : \beta = 0 \quad \text{versus} \quad H_1 : \beta \neq 0$$

that is, Y is not dependent upon x.

Recall

$$\frac{B-\beta}{\sqrt{\sigma^2/S_{xx}}} = \frac{\sqrt{S_{xx}}(B-\beta)}{\sigma} \sim N(0,1) \quad \text{and}$$

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2}.$$

Hence, from the definition of a t distribution it follows that

$$\frac{\sqrt{S_{xx}}(B-\beta)/\sigma}{\sqrt{SS_R/(\sigma^2(n-2))}} = \sqrt{\frac{(n-1)S_{xx}}{SS_R}}(B-\beta) \approx t_{n-2}$$

Therefore, if H_0 is true (and so $\beta = 0$), then

$$\sqrt{\frac{(n-2)S_{xx}}{SS_R}}B \sim t_{n-2}$$

the test statistic:

$$TS = \sqrt{\frac{(n-2)S_{xx}}{SS_R}}|B|$$

a significance level γ test of H_0 is to

reject H_0 if $TS > t_{\gamma/2,\, n-2}$

accept H_0 otherwise.

If $TS = v$, the p-value is given by

$$p\text{-value} = P\{|T_{n-2}| > v\} = 2P\{T_{n-2} > v\}$$

eg. An individual claims that the fuel consumption of his automobile does not depend on how fast the car is driven. To test the plausibility of this hypothesis, the car was tested at various speeds between 45 and 70 miles per hour.

The miles per gallon attained at each of these speeds was determined, with the following data resulting

speed	45	50	55	60	65	70	75
miles per gallon	24.2	25.0	23.3	22.0	21.5	20.6	19.8

Do these data refuse the claim at the 1 percent level of significance?

(Sol)

A simple linear regression model:

$Y = \alpha + \beta x + e$

where Y = the miles per gallon of the car and x = the speed at which it is being driven.

The problem of testing

$H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$.

The test statistic:

$$TS = \sqrt{\frac{(n-2)S_{xx}}{SS_R}}\,|B|$$

where $SS_R = (S_{xx}S_{YY} - S_{xY}^2)/S_{xx}$ and $B = S_{xY}/S_{xx}$.

From the data, $S_{xx} = 700$, $S_{YY} = 21.757$, and $S_{xY} = -119$.

Hence, $SS_R = (700 \cdot 21.757 - (119)^2)/700 = 1.527$,

$B = -119/700 = -0.17$, and

$TS = \sqrt{5 \cdot 700/1.527}\,|-0.17| = 8.139$.

Since $t_{0.005,\,5} = 4.032$, the hypothesis H_0 is rejected at the 1 percent level of significance.

A confidence interval of the estimator for β:

From

$$\sqrt{\frac{(n-2)S_{xx}}{SS_R}}(B-\beta) \sim t_{n-2},$$

for any a, $0 < a < 1$,

$$P\{-t_{a/2,n-2} < \sqrt{\frac{(n-2)S_{xx}}{SS_R}}(B-\beta) < t_{a/2,n-2}\} = 1-a.$$

or equivalently,

$$P\{B - \sqrt{\frac{SS_R}{(n-2)S_{xx}}}t_{a/2,n-2} < \beta < B + \sqrt{\frac{SS_R}{(n-2)S_{xx}}}t_{a/2,n-2}\} = 1-a.$$

Therefore, a 100(1-a) percent confidence interval of the estimator for β is

$$(B - \sqrt{\frac{SS_R}{(n-2)S_{xx}}}t_{a/2,n-2},\ B + \sqrt{\frac{SS_R}{(n-2)S_{xx}}}t_{a/2,n-2}).$$

eg. Derive a 95 percent confidence interval of the estimator for β.

(Sol)

Since $t_{0.025,\,5} = 2.571$, the 95 percent confidence interval is determined by

$$-0.170 \pm 2.571\sqrt{\frac{1.527}{5 \cdot 700}} = -0.170 \pm 0.054.$$

- regression to the mean

If we assume a linear regression relationship between the characteristic of the offspring (Y), and that of the parent (x), then a regression to the mean will occur when the regression parameter β is between 0 and 1. That is, if

$E[Y] = \alpha + \beta x$ and $0 < \beta < 1$,

then $E[Y]$ will be smaller than x when x is large, and greater than x when x is small.

eg. To illustrate Galton's thesis of regression to the mean, the British statistician Karl Pearson plotted the heights of 10 randomly chosen sons versus that of their fathers. The resulting data (in inches) were as follows:

Father's height	60	62	64	65	66	67	68	70	72	74
Son's height	63.6	65.2	66	65.5	66.9	67.1	67.4	68.3	70.1	70

We will determine whether the preceding data are strong enough to prove that there is a regression toward the mean by taking this statement as the alternative hypothesis. That is, we will use the above data to test

$$H_0 : \beta \geq 1 \quad \text{versus} \quad H_1 : \beta < 1$$

which is equivalent to a test of

$$H_0 : \beta = 1 \quad \text{versus} \quad H_1 : \beta < 1.$$

When $\beta = 1$, the test statistic is

$$TS = \sqrt{\frac{(n-2)S_{xx}}{SS_R}}(B-\beta) = \sqrt{\frac{8S_{xx}}{SS_R}}(B-1) \sim t_8$$

The significance level α test is to

reject H_0 if $TS < -t_{\alpha,8}$

$$TS = \sqrt{\frac{8S_{xx}}{SS_R}}(B-1) = 30.2794\,(0.4646-1) = -16.21$$

Since $t_{0.01,8} = 2.869$, the hypothesis H_0 is rejected at the 1 percent level of significance.

Therefore, a regression toward mean is established.

(2) inferences concerning α

The determination of confidence intervals and hypothesis tests for α is accomplished in exactly the same manner as was done for β.

Note that

$$\sqrt{\frac{n(n-2)S_{xx}}{\sum_{i=1}^{n} x_i^2 SS_R}}(A-\alpha) \sim t_{n-2}.$$

Therefore, the 100(1-a) percent confidence interval for α is the interval

$$A \pm \sqrt{\frac{\sum_{i=1}^{n} x_i^2 SS_R}{n(n-2)S_{xx}}}\, t_{a/2,\,n-2}.$$

(3) inferences concerning the mean response of $\alpha + \beta X$

It is often of interest to use the data pairs (x_i, Y_i), $i = 1, \cdots, n$, to estimate $\alpha + \beta x_0$,

the mean response for a given input level x_0.

The estimator $A + Bx_0$ is an unbiased estimator since

$$E[A + Bx_0] = E[A] + x_0 E[B] = \alpha + \beta x_0.$$

We see that

$$A + Bx_0 = \bar{Y} - B\bar{x} + Bx_0 = \bar{Y} - B(\bar{x} - x_0)$$

$$= \sum_{i=1}^{n} Y_i \left(\frac{1}{n} - \frac{(x_i - \bar{x})(\bar{x} - x_0)}{S_{xx}}\right)$$

and

$$Var(A + Bx_0) = \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{(x_i - \bar{x})(\bar{x} - x_0)}{S_{xx}}\right)^2 Var(Y_i)$$

$$= \sigma^2 \left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}\right)$$

Hence,

$$A + Bx_0 \sim N(\alpha + \beta x_0, \sigma^2 \left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}\right)).$$

Since

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2},$$

it follows that

$$\frac{A + Bx_0 - (\alpha + \beta x_0)}{\sqrt{\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}}\sqrt{\frac{SS_R}{n-2}}} \sim t_{n-2}.$$

Therefore, the 100(1-a) percent confidence interval for $\alpha + \beta x_0$ is the interval

$$A + Bx_0 \pm \sqrt{\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}}\sqrt{\frac{SS_R}{n-2}}\, t_{a/2,\, n-2}.$$

eg. In the example of the regression to the mean, determine a 95 percent confidence interval for *the average height of all males* whose fathers are 68 inches tall.

(Sol)

the observed values are

$$n = 10, \quad x_0 = 68, \quad \bar{x} = 66.8,$$
$$S_{xx} = 171.6, \quad SS_R = 1.49721$$

we see that

$$\sqrt{\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}}\sqrt{\frac{SS_R}{n-2}} = 0.1424276$$

$$t_{0.025,\,8} = 2.306, \quad A + Bx_0 = 67.56751.$$

Therefore, the 95 percent confidence interval:

$$\alpha + \beta x_0 \in (67.568 - 2.306 \cdot 0.142, 67.568 + 2.306 \cdot 0.142)$$
$$\in (67.239, 67.896)$$

(4) prediction interval of a future response

Let us now suppose that rather than being concerned with determining *a single value* to predict a response, we are interested in finding a prediction interval that, with a given degree of confidence, will contain the response.

$$Y \sim N(\alpha + \beta x_0, \sigma^2)$$

$$A + Bx_0 \sim N(\alpha + \beta x_0, \sigma^2 \left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}\right))$$

Y is independent of $A + Bx_0$ and so

$$Y - A - Bx_0 \sim N(0, \sigma^2 \left(1 + \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}\right)).$$

This implies that

$$\frac{Y - A - Bx_0}{\sigma\sqrt{\frac{n+1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}}} \sim N(0,1).$$

Since

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2},$$

$$\frac{Y - A - Bx_0}{\sigma\sqrt{\frac{n+1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}}\sqrt{\frac{SS_R}{n-2}}} \sim t_{n-2}.$$

For any value a, $0 < a < 1$,

$$P\{-t_{a/2,\,n-2} < \frac{Y - A - Bx_0}{\sigma\sqrt{\frac{n+1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}}\sqrt{\frac{SS_R}{n-2}}} < t_{a/2,\,n-2}\} = 1 - a.$$

Based on the response values Y_i corresponding to the input values x_i, $i = 1, \cdots, n$:

With 100(1-a) percent confidence, the response Y at the input level x_0 will be contained in the interval

$$A + Bx_0 \pm t_{a/2,\,n-2}\sqrt{(\frac{n+1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}})\frac{SS_R}{n-2}}$$

(prediction interval)

Note: A confidence interval is an interval that does contain, with a given confidence, a fixed parameter of interest.

A prediction interval, on the other hand, is an interval that will contain, again with a given degree of confidence, a random variable of interest.

eg. In the example of the regression to the mean, determine an interval that we can be 95 percent certain will contain *the height of a given male* whose fathers are 68 inches tall.

That is, the prediction interval for the value of $x = 68$.

(Sol)

$$t_{a/2,\,n-2}\sqrt{(\frac{n+1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}})\frac{SS_R}{n-2}} = 1.050$$

Therefore, the prediction interval with 95 percent confidence:

$Y(68) \in 67.568 \pm 1.050$

- the coefficient of determination and the sample correlation coefficient

Suppose we wanted to measure the amounts of variation in the set of response values $Y_1, \cdots, Y_n$ corresponding to the set of input values $x_1, \cdots, x_n$.

A standard measure in statistics of the amount of variation in a set of values $Y_1, \cdots, Y_n$ is given by the quantity

$$S_{YY} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2.$$

Let us consider now the question as to how much of the variation in the values of the response variable is due to the different input values, and how much is due to the inherent variance of the responses even when the input values are taken into account.

Note that the quantity

$$SS_R = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$$

measures the remaining amount of variation in the response values after the different input values have been taken into account.

The quantity R^2 defined by

$$R^2 = \frac{S_{YY} - SS_R}{S_{YY}} = 1 - \frac{SS_R}{S_{YY}}$$

(coefficient of determination)

represents the proportion of the variation in the response variables that is explained by the different input values.

$\rightarrow$ an indicator of how well the regression model fits data.

The coefficient of determination R^2 will have a value between 0 and 1.

the value of R^2 near 1: most of the variation of response data is explained by the different input values

$\rightarrow$ a good fit in the regression model.

the value of R^2 near 0: little of the variation of response data is explained by the different input values

$\rightarrow$ a poor fit in the regression model.

eg. In the previous example which relates the height of a son to that of his father,

$S_{YY} = 38.521, \quad SS_R = 1.497$

Thus, the coefficient of determination

$$R^2 = 1 - \frac{1.497}{38.531} = 0.961.$$

In other words, 96 percent of the variation of the heights of the 10 individuals is explained by the heights of their fathers.

The remaining (unexplained) 4 percent of the variation of a son's height, that is, it is due to σ^2, the variance of the error random variable.

the sample correlation coefficient r of the set of data pairs (x_i, Y_i), $i = 1, \cdots, n$ by

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (Y_i - \overline{Y})^2}} = \frac{S_{xY}}{\sqrt{S_{xx} S_{YY}}}.$$

Since

$$SS_R = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}},$$

we see that

$$r^2 = \frac{S_{xY}^2}{S_{xx}S_{YY}} = \frac{S_{xx}S_{YY} - SS_R S_{xx}}{S_{xx}S_{YY}} = 1 - \frac{SS_R}{S_{YY}} = R^2.$$

That is,

$$|r| = \sqrt{R^2}.$$

Here, the sign of r is the same as that of B.

When the simple linear regression model is correct, the standardized residuals are approximately independent standard normal random variables, and thus should be randomly distributed about 0 with about 95 percent of their values being between -1.96 and +1.96.

In addition, a plot of the standardized residuals should not indicate any distinct pattern.

- analysis of residuals: assessing the model

a simple linear regression model:

$$Y = \alpha + \beta x + e, \quad e \sim N(0, \sigma^2).$$

To test this model is appropriate in a given situation, the scatter diagram is investigated. The analysis begins by normalizing, or standardizing, the residuals, that is,

$$\frac{Y_i - (A + Bx_i)}{\sqrt{SS_R/(n-2)}}, \quad i = 1, \cdots, n.$$

(standardized residuals)

- Multiple Linear Regression

. A sample of size n:

$(\underline{x}_i, Y_i),\ i=1, \cdots, n,$ where $\underline{x}_i$ represents the ith $k(>1)$ dimensional input vector and the ith output (random variable) is given by

$$Y_i = \sum_{j=0}^{k} \beta_j x_{ij} + e,\ i=1, \cdots, n,$$

where $e \sim N(0, \sigma^2)$.

Let

$\underline{x}_i = [1, x_{i1}, \cdots, x_{ik}]^T$, $\underline{Y} = [Y_1, \cdots, Y_n]^T$, $\underline{\beta} = [\beta_0, \cdots, \beta_k]^T$,

$\underline{e} = [e_1, \cdots, e_n]^T$, and

$$X = \begin{bmatrix} \underline{x}_1 \\ \vdots \\ \underline{x}_n \end{bmatrix} = \begin{bmatrix} 1\ x_{11} \cdots x_{1k} \\ \vdots \\ 1\ x_{n1} \cdots x_{nk} \end{bmatrix}.$$ Then, $\underline{Y} = X\underline{\beta} + \underline{e}$.

. An estimator of Y_i:

$$\hat{Y}_i = \sum_{j=0}^{k} B_j x_{ij},\ i=1, \cdots, n.$$

Let $\underline{\hat{Y}} = [\hat{Y}_1, \cdots, \hat{Y}_n]^T$ and $\underline{B} = [B_0, \cdots, B_k]^T$. Then, $\underline{\hat{Y}} = X\underline{B}$

. The sum of square residuals:

$$SS_R = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

By differentiating SS_R with respect to β_is' and making them zero; that is,

$$\frac{\partial SS_R}{\partial \beta_i}|_{\beta_i = B_i} = 0,\ i = 1,\ \cdots, k,$$

The parameter vector $\underline{B}$ is determined by

$$X^T X\underline{B} = X^T \underline{Y}.$$

This implies that $\underline{B} = (X^T X)^{-1} X^T \underline{Y}$.

Here, $\underline{B}$ is an unbiased estimator since

$$\begin{aligned} E[\underline{B}] &= E[(X^T X)^{-1} X^T \underline{Y}] \\ &= E[(X^T X)^{-1} X^T (X\underline{\beta} + \underline{e})] \\ &= E[(X^T X)^{-1}(X^T X)\underline{\beta} + (X^T X)^{-1} X^T \underline{e}] \\ &= \underline{\beta}. \end{aligned}$$

. Covariance of $\underline{B}$

Let $C = (X^T X)^{-1} X^T$. Then,

$$\underline{B} = C\underline{Y} = \begin{bmatrix} c_{11} \cdots c_{1n} \\ \vdots \quad \vdots \\ c_{p1} \cdots c_{pn} \end{bmatrix} \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix},\ \text{where}\ p = k+1.$$

The covariance $\underline{B}$ is determined by

$$\begin{aligned} Cov(B_{i-1,}, B_{j-1}) &= Cov\left(\sum_{l=1}^{n} c_{il} Y_l, \sum_{r=1}^{n} c_{jr} Y_r\right) \\ &= \sum_{r=1}^{n}\sum_{l=1}^{n} c_{il} c_{jr} Cov(Y_l, Y_r) \end{aligned}$$

Here,

$$Cov(Y_l, Y_r) = \begin{cases} 0 & \text{if } l \neq r \\ Var(Y_r) = \sigma^2 & \text{if } l = r \end{cases}$$

That is,

$$Cov(B_{i-1,}B_{j-1}) = \sigma^2 \sum_{r=1}^{n} c_{ir}c_{jr} = \sigma^2 \left[CC^T \right]_{ij}.$$

Let

$$Cov(B) = \begin{bmatrix} Cov(B_0, B_0) \cdots Cov(B_0, B_k) \\ \vdots \qquad\qquad \vdots \\ Cov(B_k, B_0) \cdots Cov(B_k, B_k) \end{bmatrix}.$$

Then,

$$Cov(B) = \sigma^2 CC^T = \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}.$$

Therefore,

$$\underline{B} \sim N(\underline{\beta}, \sigma^2 (X^T X)^{-1}).$$

. Estimation of noise variance σ^2

The sum of square residuals is given by

$$SS_R = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

Here,

$$\frac{SS_R}{\sigma^2} = \chi^2_{n-(k+1)}.$$

Then,

$$E\left[\frac{SS_R}{\sigma^2} \right] = n-k-1; \text{ that is, } E\left[\frac{SS_R}{n-k-1} \right] = \sigma^2.$$

Therefore, $\hat{\sigma}^2 = \frac{SS_R}{n-k-1}$ is an unbiased estimator of σ^2.

. Coefficient of multiple determination

The coefficient of multiple determination is defined by

$$R^2 = 1 - \frac{SS_R}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2},$$

where

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

That is, the sum of square residuals is compared with the variation of Y_i to measure the fitness to the given data (or performance of regression).

. Confidence Interval for the mean of output Y

For the given input $\underline{x} = [x_1, \cdots, x_k]^T$, the mean of Y is

$$E[Y|_{\underline{x}}] = E\left[\sum_{i=0}^{k}\beta_i x_i + e\right] = \sum_{i=0}^{k}\beta_i x_i.$$

Then, the mean response of multiple regression model is

$$E\left[\hat{Y}|_{\underline{x}}\right] = E\left[\sum_{i=0}^{k} B_i x_i\right] = \sum_{i=0}^{k} E[B_i]x_i = \sum_{i=0}^{k}\beta_i x_i$$

and the variance of $\hat{Y}|_{\underline{x}}$ is

$$Var\left(\sum_{i=1}^{k} B_i x_i\right) = Cov\left(\sum_{i=0}^{k} B_i x_i, \sum_{j=0}^{k} B_j x_j\right)$$
$$= \sum_{i=0}^{k}\sum_{j=0}^{k} x_i x_j Cov(B_i, B_j) = \underline{x}^T (X^T X)^{-1}\underline{x}\sigma^2.$$

That is, $\hat{Y}|_x \sim N\left(\sum_{i=0}^{k}\beta_i x_i, \underline{x}^T(X^TX)^{-1}\underline{x}\sigma^2\right)$. Then,

$\frac{\sum_{i=0}^{k}B_i x_i - \sum_{i=0}^{k}\beta_i x_i}{\sigma\sqrt{\underline{x}^T(X^TX)^{-1}\underline{x}}} \sim N(0,1)$. Here, replacing σ with

$\hat{\sigma} = \sqrt{\frac{SS_R}{n-k-1}}$ yields that $\frac{\sum_{i=0}^{k}B_i x_i - \sum_{i=0}^{k}\beta_i x_i}{\hat{\sigma}\sqrt{\underline{x}^T(X^TX)^{-1}\underline{x}}} \sim t_{n-k-1}$.

Therefore, a $100(1-\alpha)\%$ confidence interval for the mean of $Y|_{\underline{x}}$ is determined by

$$Y|_{\underline{x}} = \sum_{i=0}^{k}\beta_i x_i \in \left(\sum_{i=0}^{k}B_i x_i \pm t_{\alpha/2,n-k-1}\hat{\sigma}\sqrt{\underline{x}^T(X^TX)^{-1}\underline{x}}\right).$$

. Prediction Interval for the output Y

For the given input $\underline{x} = [x_1, \cdots, x_k]^T$, the output $Y|_{\underline{x}}$ is

$$Y|_{\underline{x}} = \sum_{i=0}^{k}\beta_i x_i + e \sim N\left(\sum_{i=0}^{k}\beta_i x_i, \sigma^2\right) \text{ and}$$

the estimated output $\hat{Y}|_{\underline{x}}$is

$$\hat{Y}|_x = \sum_{i=0}^{k}B_i x_i \sim N\left(\sum_{i=0}^{k}\beta_i x_i, \underline{x}^T(X^TX)^{-1}\underline{x}\sigma^2\right).$$

Then,

$$Y|_{\underline{x}} - \hat{Y}|_{\underline{x}} = Y|_{\underline{x}} - \sum_{i=0}^{k}B_i x_i \sim N\left(0, \sigma^2 + \underline{x}^T(X^TX)^{-1}\underline{x}\sigma^2\right)$$

and

$$\frac{Y|_{\underline{x}} - \sum_{i=0}^{k} B_i x_i}{\sigma\sqrt{1+\underline{x}^T(X^TX)^{-1}\underline{x}}} \sim N(0,1).$$

Here, replacing σ with $\hat{\sigma} = \sqrt{\frac{SS_R}{n-k-1}}$ yields that

$$\frac{Y|_{x} - \sum_{i=0}^{k} B_i x_i}{\hat{\sigma}\sqrt{1+\underline{x}^T(X^TX)^{-1}\underline{x}}} \sim t_{n-k-1}.$$

Therefore, a $100(1-\alpha)\%$ confidence interval for the output $Y|_{\underline{x}}$ is determined by

$$Y|_{\underline{x}} \in \left(\sum_{i=0}^{k} B_i x_i \pm t_{\alpha/2,n-k-1}\hat{\sigma}\sqrt{1+\underline{x}^T(X^TX)^{-1}\underline{x}} \right).$$

- linear regression models

. linear estimator: the output for the kth input $\underline{x}_k$ is given by

$$y_k = y(\underline{x}_k) = w_0 + \sum_{i=1}^{d} w_i x_{ik}.$$

Let

$\underline{x}_k = [1, x_{1k}, \cdots, x_{dk}]^T$ and $\underline{w} = [w_0, w_1, \cdots, w_d]^T$.

Then,

$$y_k = \underline{w}^T \underline{x}_k.$$

Let d_k be the desired response for the kth input pattern.

Then, the error for the kth pattern is

$$\epsilon_k = d_k - y_k = d_k - \underline{w}^T \underline{x}_k.$$

the error square:

$$\epsilon_k^2 = d_k^2 + \underline{w}^T \underline{x}_k \underline{x}_k^T \underline{w} - 2 d_k \underline{x}_k^T \underline{w}.$$

the mean square error (MSE):

$$E[\epsilon_k^2] = E[d_k^2] + \underline{w}^T E[\underline{x}_k \underline{x}_k^T] \underline{w} - 2E[d_k \underline{x}_k^T] \underline{w}$$

Let

$R = E[\underline{x}_k \underline{x}_k^T]$ (input correlation matrix) and

$P = E[d_k \underline{x}_k]$.

Then, MSE becomes

$E[\epsilon_k^2] = E[d_k^2] + \underline{w}^T R \underline{w} - 2P^T \underline{w}$. (quadratic form)

the derivative of MSE with respect to $\underline{w}$:

$$\nabla E[\epsilon_k^2] = \frac{\partial E}{\partial \underline{w}} = 2R\underline{w} - 2P.$$

The optimal weight vector $\underline{w}^*$ can be determined by the condition of

$\nabla E[\epsilon_k^2]|_{\underline{w} = \underline{w}^*} = 0$.

That is, $2R\underline{w}^* - 2P = 0$.

This implies that the optimal weight is described by

$\underline{w}^* = R^{-1}P$.

In this case, the minimum mean square error (MMSE) becomes

$MMSE = E[d_k^2] + \underline{w}^{*T} R \underline{w}^* - 2P^T \underline{w} = E[d_k^2] - P^T \underline{w}^*$.

Let $\underline{v} = \underline{w} - \underline{w}^*$ and rewrite the MSE as follows:

$$\begin{aligned} E[\epsilon_k^2] &= E[d_k^2] + \underline{w}^T R \underline{w} - 2P^T \underline{w} \\ &= E[d_k^2] - P^T \underline{w}^* + \underline{w}^{*T} R \underline{w}^* + \underline{w}^T R \underline{w} - 2\underline{w}^T R \underline{w}^* \\ &= MMSE + (\underline{w} - \underline{w}^*)^T R (\underline{w} - \underline{w}^*) \\ &= MMSE + \underline{v}^T R \underline{v} \end{aligned}$$

To analyse the MSE, let us set

$$F(\underline{v}) = \underline{v}^T R \underline{v}.$$

That is,

$$E[\epsilon_k^2] = MMSE + F(\underline{v}).$$

The MSE depends on $F(\underline{v})$.

Here, the input correlation matrix R can be decomposed by

$$R = Q \Lambda Q^{-1}, \quad Q = \begin{pmatrix} q_1 \, q_2 \cdots q_d \end{pmatrix}, \text{ and } \Lambda = \begin{pmatrix} \lambda_0 & 0 & \cdots & 0 \\ 0 & \lambda_1 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & \lambda_d \end{pmatrix}$$

where q_i is the column vector representing the eigenvector of R and λ_i is the eigenvalue corresponding to q_i.

Since R is the symmetric matrix, q_is are orthogonal and

$$R = Q \Lambda Q^{-1} = Q \Lambda Q^T.$$

Here, $F(\underline{v})$ can be rewritten as

$$F(\underline{v}) = \underline{v}^T R \underline{v} = \underline{v}^T Q \Lambda Q^T \underline{v}.$$

Let $\underline{v}' = Q^T \underline{v}$, that is, rotation of $\underline{v}$ toward eigenvector axes. Here, the eigenvectors of R define the principal axes of the error surface. Then,

$$F(\underline{v}') = \underline{v}'^T \Lambda \underline{v}'. \text{ (ellipsoid)}$$

Let $F(\underline{v}') = c$ in 2 dimensional space.

Then,

$\lambda_0 v_0'^2 + \lambda_1 v_1'^2 = c.$

the length of v_0' axis = $\sqrt{c/\lambda_0}$, the length of v_1' axis = $\sqrt{c/\lambda_1}$

cf. the equation of ellipse: $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$

. gradient descent method (batch mode)

The weight vector is updated by

$$\begin{aligned}\underline{w}_{k+1} &= \underline{w}_k + \mu(-\nabla F|_{\underline{v}=\underline{v}_k}) \\ &= \underline{w}_k + 2\mu R(\underline{w}^* - \underline{w}_k) \\ &= (I - 2\mu R)\underline{w}_k + 2\mu R\underline{w}^*\end{aligned}$$

This can be redescribed by

$\underline{v}_{k+1} = (I - 2\mu R)\underline{v}_k = (I - 2\mu Q\Lambda Q^T)\underline{v}_k.$

Let $\underline{v} = Q\underline{v}'$.

Then,

$Q\underline{v}'_{k+1} = (I - 2\mu Q\Lambda Q^T)Q\underline{v}'_k$, that is, $\underline{v}'_{k+1} = (I - 2\mu\Lambda)\underline{v}'_k$.

The iterative form of $\underline{v}_k^{'}$ is

$$\underline{v}_k^{'} = (I - 2\mu\Lambda)^k \underline{v}_0^{'}.$$

Therefore, the gradient descent algorithm is stable and convergent when

$$\lim_{k\to\infty}(I - 2\mu\Lambda)^k = 0.$$

This implies that the convergence condition is

$$0 < \mu < \frac{1}{tr[R]} \leqq \frac{1}{\lambda_{\max}}$$

where $\lambda_{\max}$ is the maximum eigenvalue of R.

Note that

$$\lambda_{\max} \leqq tr[\Lambda] = tr[R].$$

In this case, the learning curve of MSE is determined as follows:

$$\begin{aligned} E[\epsilon_k^2] &= MMSE + \underline{v}_k^T R \underline{v}_k \\ &= MMSE + \underline{v}_k^{'} \Lambda \underline{v}_k^{'} \\ &= MMSE + [(I - 2\mu\Lambda)^k \underline{v}_0^{'}]^T \Lambda [(I - 2\mu\Lambda)^k \underline{v}_0^{'}] \\ &= MMSE + \underline{v}_0^{'T} (I - 2\mu\Lambda)^{2k} \Lambda \underline{v}_0^{'} \\ &= MMSE + \sum_{n=0}^{d} v_{0n}^{'2} \lambda_n (1 - 2\mu\lambda_n)^{2k} \end{aligned}$$

The learning curve of MSE is dependent upon the geometric ratio

$$r_n^2 \equiv (1 - 2\mu\lambda_n)^2$$

. least mean square (LMS) method (on-line mode)

The weight vector is updated by

$$\begin{aligned}\underline{w}_{k+1} &= \underline{w}_k - \mu\widehat{\nabla}_k \\ &= \underline{w}_k + 2\mu\epsilon_k\underline{x}_k\end{aligned}$$

where

$\epsilon_k = d_k - y_k = d_k - \underline{x}_k^T\underline{w}_k$ and

$$\widehat{\nabla}_k \equiv [\frac{\partial\epsilon_k^2}{\partial w_0}, \frac{\partial\epsilon_k^2}{\partial w_1}, \cdots, \frac{\partial\epsilon_k^2}{\partial w_d}]^T = -2\epsilon_k\underline{x}_k.$$

That is, $\widehat{\nabla}_k$ is not a true gradient, but an estimated gradient from ϵ_k and $\underline{x}_k$.

Is $\widehat{\nabla}_k$ an unbiased estimator?

$$\begin{aligned}E[\widehat{\nabla}_k] &= -2E[\epsilon_k\underline{x}_k] \\ &= -2E[d_k\underline{x}_k - \underline{x}_k\underline{x_k}^T\underline{w}_k] \\ &= 2(R\underline{w}_k - P) \\ &= \nabla E|_{\underline{w}=\underline{w}_k}\end{aligned}$$

The answer is yes. To check the convergence condition of the LMS method, let us consider the expectation of $\underline{w}_{k+1}$:

$$\begin{aligned}E[\underline{w}_{k+1}] &= E[\underline{w}_k] + 2\mu E[\epsilon_k\underline{x}_k] \\ &= E[\underline{w}_k] + 2\mu(E[d_k\underline{x}_k] - E[\underline{x}_k\underline{x}_k^T\underline{w}_k] \\ &= E[\underline{w}_k] + 2\mu(P - RE[\underline{w}_k]) \\ &= (1-2\mu R)E[\underline{w}_k] + 2\mu R\underline{w}^*\end{aligned}$$

This can be redescribed by

$$E[\underline{v}_{k+1}] = (I - 2\mu R)E[\underline{v}_k],$$

that is,

$$E[\underline{v}_k^{'}] = (I - 2\mu \Lambda)^k \underline{v}_0^{'}.$$

This implies that the convergence condition is

$$0 < \mu < \frac{1}{tr[R]} \leqq \frac{1}{\lambda_{\max}}$$

where $\lambda_{\max}$ is the maximum eigenvalue of R,

that is, the convergence condition is same as the batch mode.

To check the performance of LMS method, let us consider the following definition of misadjustment:

$$M \equiv \frac{excess\, MSE}{MMSE} = \frac{E[MSE - MMSE]}{MMSE}$$

where the excess MSE is given by

$$\begin{aligned} excess\, MSE &= E[\underline{v}_k^T R \underline{v}_k] \\ &= E[\underline{v}_k^{'T} \Lambda \underline{v}_k^{'}] \\ &= \sum_{n=0}^{d} \lambda_n E[v_{nk}^{'2}] \end{aligned}$$

To analyse the excess MSE, let us set

$$\widehat{\nabla}_k = \nabla_k + \underline{n}_k$$

where $\underline{n}_k$ represents the noise vector associated with $\widehat{\nabla}_k$.

Near $\underline{w}^*$, the noise vector can be approximated as

$$\underline{n}_k \approx \widehat{\nabla}_k = -2\epsilon_k \underline{x}_k.$$

Then, the covariance of $\underline{n}_k$ becomes

$$Cov[\underline{n}_k] = E[\underline{n}_k \underline{n}_k^T] \approx 4E[\epsilon_k^2 \underline{x}_k \underline{x}_k^T].$$

Near $\underline{w}^*$, ϵ_k and $\underline{x}_k$ is uncorrelated, that is,

$$Cov[\underline{n}_k] \approx 4E[\epsilon_k^2]E[\underline{x}_k \underline{x}_k^T] \approx 4MMSE \cdot R \quad \text{and}$$

$$\begin{aligned} Cov[\underline{n}_k'] &= Cov[Q^{-1}\underline{n}_k] \\ &= E[Q^{-1}\underline{n}_k (Q^{-1}\underline{n}_k)^T] \\ &= Q^{-1} Cov[\underline{n}_k] Q \\ &\approx 4MMSE \cdot \Lambda \end{aligned}$$

From the weight update rule,

$$\begin{aligned} \underline{v}_{k+1} &= \underline{v}_k - \mu \widehat{\nabla}_k \\ &= \underline{v}_k - \mu(2R\underline{v}_k + \underline{n}_k) \\ &= (I - 2\mu R)\underline{v}_k - \mu \underline{n}_k \end{aligned}$$

This implies that

$$\underline{v}_{k+1}' = (I - 2\mu\Lambda)\underline{v}_k' - \mu \underline{n}_k'.$$

From this equation, the covariance of $\underline{v}_k^{'}$ is determined by

$$Cov[\underline{v}_k^{'}] = (I-2\mu\Lambda)^2 Cov[\underline{v}_k^{'}] + \mu^2 Cov[\underline{n}_k^{'}]$$
$$= \frac{\mu}{4}(\Lambda - \mu\Lambda^2)^{-1} Cov[\underline{n}_k^{'}]$$

By substituting the result of $Cov[\underline{n}_k^{'}]$, we get

$$Cov[\underline{v}_k^{'}] \approx \mu MMSE(\Lambda - \mu\Lambda^2)^{-1}\Lambda \approx \mu MMSE \cdot I.$$

This implies that, the excess MSE becomes

$$excess\,MSE = \sum_{n=0}^{d} \lambda_n E[v_{nk}^{'2}] \approx \mu MMSE \sum_{n=0}^{d} \lambda_n.$$

Therefore, the misadjustment becomes

$$M \equiv \frac{excess\,MSE}{MMSE} \approx \mu \cdot tr[R].$$

Here, let us investigate the learning curve when we use the LMS method.

The time constant is defined by the time interval in which the given signal $f(t)$ is reduced by $1/e$.
example. $f(t) = e^{-t/\tau}$, time constant = τ.

the geometric ratio $r^2 \equiv e^{-1/\tau}$, that is, $r^2 \approx 1 - 1/\tau$.
From the previous result,

$r_n^2 = (1-2\mu\lambda_n)^2 \approx 1 - 1/\tau_n$, that is, $1/\tau_n \approx 1/(4\mu\lambda_n)$.

Here, the trace of R can be redescribed by

$$tr[R] = \sum_{n=0}^{d} \lambda_n \approx \frac{1}{4\mu} \sum_{n=0}^{d} \frac{1}{\tau_n} = \frac{d+1}{4\mu} \frac{1}{\tau_{av}}$$

where $\frac{1}{\tau_{av}} = \frac{1}{d+1} \sum_{n=0}^{d} \frac{1}{\tau_n}$.

This implies that

$\tau_{av} \approx \frac{d+1}{4\mu \cdot tr[R]}$ and

$M \approx \frac{d+1}{4\tau_{av}}$.

These results show that

(1) large μ -> small τ_{av} (fast convergence) -> large M and

(2) small μ -> large τ_{av} (slow convergence) -> small M.

Reference: Adaptive Signal Processing, chapters 3, 4, 5, and 6.