Linear Regression[*]

Paul E Pfeiffer

Abstract

Consider a pair {X,Y} with a joint distribution. A value X(ω) is observed. It is desired to estimate the corresponding value Y(ω). The best that can be hoped for is some estimate based on an average of the errors, or on the average of some function of the errors. The most common measure of error is the mean (expectation) of the square of the error. This has two important properties: it treats positive and negative errors alike, and it weights large errors more heavily than smaller ones. In general, we seek a rule (function) r such that the estimate is r(X(ω)). That is, we seek a function r such that the expectation of the square of Y - r(X) is a minimum. The problem of determining such a function is known as the regression problem. LINEAR REGRESSION: we seek the best straight line function (the regression line of Y on X) of the form u = r(t) + b, such that the mean square of Y - r(X) is a minimum. Matlab approximation procedures are compared with analytic results. More general linear regression is considered

1 Linear Regression

Suppose that a pair $\{X, Y\}$ of random variables has a joint distribution. A value $X(\omega)$ is observed. It is desired to estimate the corresponding value $Y(\omega)$. Obviously there is no rule for determining $Y(\omega)$ unless Y is a function of X. The best that can be hoped for is some estimate based on an average of the errors, or on the average of some function of the errors.

Suppose $X(\omega)$ is observed, and by some rule an estimate $\hat{Y}(\omega)$ is returned. The error of the estimate is $Y(\omega) - \hat{Y}(\omega)$. The most common measure of error is the mean of the square of the error

$$E\left[\left(Y - \hat{Y}\right)^2\right] \tag{1}$$

The choice of the mean square has two important properties: it treats positive and negative errors alike, and it weights large errors more heavily than smaller ones. In general, we seek a rule (function) r such that the estimate $\hat{Y}(\omega)$ is $r(X(\omega))$. That is, we seek a function r such that

$$E\left[(Y - r(X))^2\right] \text{ is a minimum.} \tag{2}$$

The problem of determining such a function is known as the *regression problem*. In the unit on Regression[1], we show that this problem is solved by the conditional expectation of Y, given X. At this point, we seek an important partial solution.

[*] Version 1.6: Sep 18, 2009 1:15 pm -0500

[†] http://creativecommons.org/licenses/by/3.0/

[1] "Conditional Expectation, Regression": Section The regression problem <http://cnx.org/content/m23634/latest/#cid6>

The regression line of Y on X

We seek the best straight line function for minimizing the mean squared error. That is, we seek a function r of the form $u = r(t) = at + b$. The problem is to determine the coefficients a, b such that

$$E\left[(Y - aX - b)^2\right] \quad \text{is a minimum} \tag{3}$$

We write the error in a special form, then square and take the expectation.

$$\text{Error} = Y - aX - b = (Y - \mu_Y) - a(X - \mu_X) + \mu_Y - a\mu_X - b = (Y - \mu_Y) - a(X - \mu_X) - \beta \tag{4}$$

$$\text{Error squared} = (Y - \mu_Y)^2 + a^2(X - \mu_X)^2 + \beta^2 - 2\beta(Y - \mu_Y) + 2a\beta(X - \mu_X) - 2a(Y - \mu_Y)(X - \mu_X) \tag{5}$$

$$E\left[(Y - aX - b)^2\right] = \sigma_Y^2 + a^2\sigma_X^2 + \beta^2 - 2a\text{Cov}\,[X, Y] \tag{6}$$

Standard procedures for determining a minimum (with respect to a) show that this occurs for

$$a = \frac{\text{Cov}\,[X, Y]}{\text{Var}\,[X]} \qquad b = \mu_Y - a\mu_X \tag{7}$$

Thus the optimum line, called the *regression line of* Y *on* X, is

$$u = \frac{\text{Cov}\,[X, Y]}{\text{Var}\,[X]}(t - \mu_X) + \mu_Y = \rho\frac{\sigma_Y}{\sigma_X}(t - \mu_X) + \mu_Y = \alpha(t) \tag{8}$$

The second form is commonly used to define the regression line. For certain theoretical purposes, this is the preferred form. But for *calculation*, the first form is usually the more convenient. Only the covariance (which requres both means) and the variance of X are needed. There is no need to determine Var $[Y]$ or ρ.

Example 1: The simple pair of Example 3[2] from "Variance"

```
    jdemo1
jcalc
Enter JOINT PROBABILITIES (as on the plane)  P
Enter row matrix of VALUES of X  X
Enter row matrix of VALUES of Y  Y
 Use array operations on matrices X, Y, PX, PY, t, u, and P
EX = total(t.*P)
EX =   0.6420
EY = total(u.*P)
EY =   0.0783
VX = total(t.^2.*P) - EX^2
VX =   3.3016
CV = total(t.*u.*P) - EX*EY
CV =  -0.1633
a = CV/VX
a  =  -0.0495
b = EY - a*EX
b  =   0.1100            % The regression line is u = -0.0495t + 0.11
```

[2] "Variance", Example 3: $Z = g(X, Y)$ (Example 10[3] from "Mathematical Expectation: Simple Random Variables") <http://cnx.org/content/m23441/latest/#fs-id2579889>

Example 2: The pair in Example 6[4] from "Variance"

Suppose the pair $\{X,\,Y\}$ has joint density $f_{XY}\,(t,\,u) = 3u$ on the triangular region bounded by $u = 0$, $u = 1 + t$, $u = 1 - t$. Determine the regression line of Y on X.

ANALYTIC SOLUTION

By symmetry, $E\,[X] = E\,[XY] = 0$, so Cov $[X, Y] = 0$. The regression curve is

$$u = E\,[Y] = 3\int_0^1 u^2 \int_{u-1}^{1-u} dt du = 6\int_0^1 u^2\,(1-u)\,du = 1/2 \quad (9)$$

Note that the pair is uncorrelated, but by the rectangle test is not independent. With zero values of $E\,[X]$ and $E\,[XY]$, the approximation procedure is not very satisfactory unless a very large number of approximation points are employed.

Example 3: Distribution of Example 5[6] from "Random Vectors and MATLAB" and Example 12 [7] from "Function of Random Vectors"

The pair $\{X,\,Y\}$ has joint density $f_{XY}\,(t,\,u) = \frac{6}{37}\,(t + 2u)$ on the region $0 \le t \le 2$, $0 \le u \le max\{1,\,t\}$ (see Figure Figure 1). Determine the regression line of Y on X. If the value $X\,(\omega) = 1.7$ is observed, what is the best mean-square linear estimate of $Y\,(\omega)$?

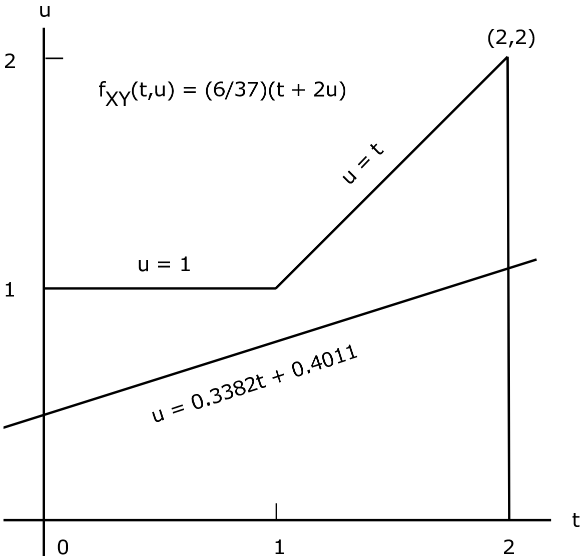

Figure 1: Regression line for Example 3 (Distribution of Example 5[9] from "Random Vectors and MATLAB" and Example 12 [10] from "Function of Random Vectors").

ANALYTIC SOLUTION

$$E\,[X] = \frac{6}{37}\int_0^1\int_0^1 \left(t^2 + 2tu\right)\,dudt + \frac{6}{37}\int_1^2\int_0^t \left(t^2 + 2tu\right)\,dudt = 50/37 \quad (10)$$

[4] "Variance", Example 6: A jointly distributed pair (Example 14[5] from "Mathematical Expectation; General Random Variables") <http://cnx.org/content/m23441/latest/#fs-id1169086148672>

[6] "Random Vectors and MATLAB", Example 5: Marginal distribution with compound expression <http://cnx.org/content/m23320/latest/#fs-id1169358726296>

[7] "Function of Random Vectors", Example 12: Continuation of Example 5[8] from "Random Vectors and Joint Distributions" <http://cnx.org/content/m23332/latest/#fs-id10748979>

The other quantities involve integrals over the same regions with appropriate integrands, as follows:

Quantity	Integrand	Value
$E\left[X^2\right]$	$t^3 + 2t^2u$	779/370
$E\left[Y\right]$	$tu + 2u^2$	127/148
$E\left[XY\right]$	$t^2u + 2tu^2$	232/185

Table 1

Then

$$\text{Var}\,[X] = \frac{779}{370} - \left(\frac{50}{37}\right)^2 = \frac{3823}{13690} \quad \text{Cov}\,[X,Y] = \frac{232}{185} - \frac{50}{37}\cdot\frac{127}{148} = \frac{1293}{13690} \tag{11}$$

and

$$a = \text{Cov}\,[X,Y]\,/\text{Var}\,[X] = \frac{1293}{3823} \approx 0.3382, \qquad b = E\left[Y\right] - aE\left[X\right] = \frac{6133}{15292} \approx 0.4011 \tag{12}$$

The regression line is $u = at + b$. If $X(\omega) = 1.7$, the best linear estimate (in the mean square sense) is $\hat{Y}(\omega) = 1.7a + b = 0.9760$ (see Figure 1 for an approximate plot).

APPROXIMATION

```
    tuappr
Enter matrix [a b] of X-range endpoints  [0 2]
Enter matrix [c d] of Y-range endpoints  [0 2]
Enter number of X approximation points  400
Enter number of Y approximation points  400
Enter expression for joint density  (6/37)*(t+2*u).*(u<=max(t,1))
Use array operations on X, Y, PX, PY, t, u, and P
EX = total(t.*P)
EX =  1.3517                  % Theoretical = 1.3514
EY = total(u.*P)
EY =  0.8594                  % Theoretical = 0.8581
VX = total(t.^2.*P) - EX^2
VX =  0.2790                  % Theoretical = 0.2793
CV = total(t.*u.*P) - EX*EY
CV =  0.0947                  % Theoretical = 0.0944
a = CV/VX
a  =  0.3394                  % Theoretical = 0.3382
b = EY - a*EX
b  =  0.4006                  % Theoretical = 0.4011
y = 1.7*a + b
y  =  0.9776                  % Theoretical = 0.9760
```

An interpretation of ρ^2

The analysis above shows the minimum mean squared error is given by

$$E\left[\left(Y - \hat{Y}\right)^2\right] = E\left[\left(Y - \rho\frac{\sigma_Y}{\sigma_X}\left(X - \mu_X\right) - \mu_Y\right)^2\right] = \sigma_Y^2 E\left[\left(Y^* - \rho X^*\right)^2\right] \tag{13}$$

$$= \sigma_Y^2 E\left[(Y^*)^2 - 2\rho X^* Y^* + \rho^2 (X^*)^2\right] = \sigma_Y^2 \left(1 - 2\rho^2 + \rho^2\right) = \sigma_Y^2 \left(1 - \rho^2\right) \tag{14}$$

If $\rho = 0$, then $E\left[\left(Y - \hat{Y}\right)^2\right] = \sigma_Y^2$, the mean squared error in the case of zero linear correlation. Then, ρ^2 is interpreted as the *fraction of uncertainty removed by the linear rule and X*. This interpretation should not be pushed too far, but is a common interpretation, often found in the discussion of observations or experimental results.

More general linear regression

Consider a jointly distributed class. $\{Y, X_1, X_2, \cdots, X_n\}$. We wish to deterimine a function U of the form

$$U = \sum_{i=0}^{n} a_i X_i, \ \text{ with } X_0 = 1, \text{ such that } \ E\left[(Y - U)^2\right] \ \text{ is a minimum} \tag{15}$$

If U satisfies this minimum condition, then $E\left[(Y - U)\, V\right] = 0$, or, equivalently

$$E\left[YV\right] = E\left[UV\right] \ \text{ for all } V \text{ of the form } \ V = \sum_{i=0}^{n} c_i X_i \tag{16}$$

To see this, set $W = Y - U$ and let $d^2 = E\left[W^2\right]$. Now, for any α

$$d^2 \le E\left[(W + \alpha V)^2\right] = d^2 + 2\alpha E\left[WV\right] + \alpha^2 E\left[V^2\right] \tag{17}$$

If we select the special

$$\alpha = -\frac{E\left[WV\right]}{E\left[V^2\right]} \quad \text{then} \quad 0 \le -\frac{2E[WV]^2}{E\left[V^2\right]} + \frac{E[WV]^2}{E[V^2]^2} E\left[V^2\right] \tag{18}$$

This implies $E[WV]^2 \le 0$, which can only be satisfied by $E\left[WV\right] = 0$, so that

$$E\left[YV\right] = E\left[UV\right] \tag{19}$$

On the other hand, if $E\left[(Y - U)\, V\right] = 0$ for all V of the form above, then $E\left[(Y - U)^2\right]$ is a minimum. Consider

$$E\left[(Y - V)^2\right] = E\left[(Y - U + U - V)^2\right] = E\left[(Y - U)^2\right] + E\left[(U - V)^2\right] + 2E\left[(Y - U)(U - V)\right] \tag{20}$$

Since $U - V$ is of the same form as V, the last term is zero. The first term is fixed. The second term is nonnegative, with zero value iff $U - V = 0$ *a.s.* Hence, $E\left[(Y - V)^2\right]$ is a minimum when $V = U$.

If we take V to be $1, X_1, X_2, \cdots, X_n$, successively, we obtain $n+1$ linear equations in the $n+1$ unknowns a_0, a_1, $\cdots$, a_n, as follows.

1. $E\left[Y\right] = a_0 + a_1 E\left[X_1\right] + \cdots + a_n E\left[X_n\right]$
2. $E\left[YX_i\right] = a_0 E\left[X_i\right] + a_1 E\left[X_1 X_i\right] + \cdots + a_n E\left[X_n X_i\right] \quad \text{for } \ 1 \le i \le n$

For each $i = 1, 2, \cdots, n$, we take $(2) - E\left[X_i\right] \cdot (1)$ and use the calculating expressions for variance and covariance to get

$$\text{Cov}\left[Y, X_i\right] = a_1 \text{Cov}\left[X_1, X_i\right] + a_2 \text{Cov}\left[X_2, X_i\right] + \cdots + a_n \text{Cov}\left[X_n, X_i\right] \tag{21}$$

These n equations plus equation (1) may be solved alagebraically for the a_i.

In the important special case that the X_i are uncorrelated (i.e., $\text{Cov}\,[X_i, X_j] = 0$ for $i \neq j$), we have

$$a_i = \frac{\text{Cov}\,[Y, X_i]}{\text{Var}\,[X_i]} \quad 1 \le i \le n \tag{22}$$

and

$$a_0 = E\,[Y] - a_1 E\,[X_1] - a_2 E\,[X_2] - \cdots - a_n E\,[X_n] \tag{23}$$

In particular, this condition holds if the class $\{X_i : 1 \le i \le n\}$ is iid as in the case of a simple random sample (see the section on "Simple Random Samples and Statistics[12]).

Examination shows that for $n = 1$, with $X_1 = X$, $a_0 = b$, and $a_1 = a$, the result agrees with that obtained in the treatment of the regression line, above.

Example 4: Linear regression with two variables.

Suppose $E\,[Y] = 3$, $E\,[X_1] = 2$, $E\,[X_2] = 3$, $\text{Var}\,[X_1] = 3$, $\text{Var}\,[X_2] = 8$, $\text{Cov}\,[Y, X_1] = 5$, $\text{Cov}\,[Y, X_2] = 7$, and $\text{Cov}\,[X_1, X_2] = 1$. Then the three equations are

$$\begin{array}{ccccccc} a_0 & + & 2a_2 & + & 3a_3 & = & 3 \\ 0 & + & 3a_1 & + & 1a_2 & = & 5 \\ 0 & + & 1a_1 & + & 8a_2 & = & 7 \end{array} \tag{24}$$

Solution of these simultaneous linear equations with MATLAB gives the results $a_0 = -1.9565$, $a_1 = 1.4348$, and $a_2 = 0.6957$.

[12] "Simple Random Samples and Statistics" <http://cnx.org/content/m23496/latest/>