Regression

James D Emery

Last Addition or Modification 7/4/2014

Contents

1 Linear Regression

A good reference is H. D. Brunk, **Introduction to Mathematical Satistics**, chapter 12 2nd edition, 1965, Blaissdell.

Suppose we are given random variables x and y. Suppose the joint pdf is $f(x, y)$. If there is a dependent relationship between x and y then the probability mass will tend to be be located along a line or a curve in the xy plane. Let us check if there is a linear dependency so that y is given approximately by $y = \alpha x + \beta$. So we look at the expectation of the squared difference S,

$$S = (y - (\alpha x + \beta))^2$$

We attempt to determine α and β by minimizing the expected value of the random variable S.

$$E[S] = E((y - (\alpha x + \beta))^2)$$

$$= \int \int (y - (\alpha x + \beta))^2 f(x, y) dx dy.$$

This will be small if most of the probability mass lies near the line. Then the spread of the mass is said to regress toward the line.

Let μ_x be the mean of x, and σ_x^2 be the variance of x. Let μ_y be the mean of y, and σ_y^2 be the variance of y. We have

$$E[(y - (\alpha x + \beta))^2) = E[((y - \mu_y) - (\alpha(x - \mu_x) + (\mu_y - \alpha\mu_x - \beta))^2]$$

$$= E[(y - \mu_y)^2] + \alpha^2 E[(x - \mu_x)^2] + E[(\mu_y - \alpha\mu_x - \beta)^2]$$

$$-2\alpha E[(x-\mu_x)(y-\mu_y)] + 2E[(y-\mu_y)(\mu_y-\alpha\mu_x-\beta)] - 2\alpha E[(x-\mu_x)(\mu_y-\alpha\mu_x-\beta)]$$

We have

$$E[(y - \mu_y)^2] = \sigma_y^2$$

$$E[(x - \mu_x)^2] = \sigma_x^2$$

$$E[(y - \mu_y)] = 0$$

$$E[(x - \mu_x)] = 0.$$

and

$$E[(\mu_y - \alpha\mu_x - \beta)^2] = (\mu_y - \alpha\mu_x - \beta)^2.$$

So

$$E[S] =$$

$$\sigma_y^2 + \sigma_x^2 + (\mu_y - \alpha\mu_x - \beta)^2 - 2\alpha E[(x - \mu_x)(y - \mu_y)]$$

Whatever α is, for a minimum we must have

$$\beta = \mu_y - \alpha\mu_x$$

We define the correlation coefficient as

$$\rho = \frac{1}{\sigma_x\sigma_y} E[(x - \mu_x)(y - \mu_y)]$$

Then

$$E[S] =$$

$$\sigma_y^2 + \alpha^2\sigma_x^2 - 2\alpha\sigma_x\sigma_y\rho$$

$$= (\alpha\sigma_x - \sigma_y\rho)^2 + \sigma_y^2(1 - \rho^2).$$

This is minimized if

$$\alpha = \frac{\sigma_y\rho}{\sigma_x},$$

and so

$$E[S] = \sigma_y^2(1 - \rho^2)$$

Now $E[S] \geq 0$, so

$$-1 \leq \rho \leq 1.$$

If x and y are independent variables, then the pdf

$$f(x, y) = f_1(x)f_2(y)$$

$$E[(x - \mu_x)(y - \mu_y)] = \int\int(x - \mu_x)(y - \mu_y)f_1(x)f_2(y)dxdy$$

$$= \int(x - \mu_x)f_1(x)dx \int(y - \mu_y)f_2(y)dy = E[x - \mu_x]E[y - \mu_y] = 0$$

So $\rho = 0$. But the converse is not necessarily true. A counterexample is a random variable x with a uniform distribution, and a random variable

$$y = x^2.$$

See Brunk p211.

All of the above can be carried out with functions defined on a finite discrete domain of size n with a discrete Lebesgue measure (Lebesgue (luh beg) 1826-1941) of $\mu = 1/n$ on each point. So the integral above becomes a Lebesgue integral, that is here, the integral is just a summation. So for example the inner product of two functions in this finite domain is

$$(f, g) = \int f(x)g(x)dx = \frac{1}{n}\sum_{k=1}^{n} f(x_k)g(x_k).$$

2 The Covariance Matrix

Given a random variable vector

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \dots \\ x_n \end{bmatrix}$$

The n by n covariance matrix is a matrix with elements

$$C_{ij} = E[(x_i - \mu_i)(y_j - \mu_j)]$$

If the means μ_k, are zero for all k, the matrix is

$$E[x^T x]$$

3 The Correlation Coefficient

The Pearson correlation coefficient of two random variables X Y is defined to be the covariance divided by the product of the standard deviations.

$$\rho = \frac{E[(X - \mu_X)(Y - \mu_y)]}{\sigma_X \sigma_Y}.$$

This can be interpreted as an inner product of functions defined as an integral of a product of functions which are defined on a probability measure space, divided by the product of the norms of X and Y. That is,

$$\rho = \frac{((X-\mu_X),(X-\mu_X))}{\|X-\mu_x\|\|Y-\mu_Y\|},$$

where the inner product of two functions f and g is defined as

$$(f,g) = \int fgdm,$$

for some measure m, and

$$\|f\| = \sqrt{(f,f)}.$$

For a finite sample $\{(x_1,y_1),(x_2,y_2,...(x_n,y_n)\}$, the sample correlation coefficient is written as

$$r = \frac{(x-\bar{x})\cdot(y-\bar{y})}{\|x-\bar{x}\|\|y-\bar{y}\|},$$

where x is the n-dimensional x data vector, and y is the n-dimensional y data vector.

That is

$$\begin{aligned} r &= \frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}\sqrt{\sum_{i=1}^n (y_i-\bar{y})}} \\ &= \frac{1}{n-1}\frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{s_x s_y}, \end{aligned}$$

where

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

and

$$s_y = \frac{1}{n-1}\sum_{i=1}^n (y_i-\bar{y})^2.$$

The term $n-1$ occurs rather than n in order to make s_x and s_y unbiased estimators of the variances of X and Y.

By the Cauchy-Schwarz inequality

$$|(x-\bar{x})\cdot(y-\bar{y})| \le \|x-\bar{x}\|\|y-\bar{y}\|,$$

so

$$-1 \le r \le 1$$

and $|r| = 1$ if and only if $x-\bar{x}$ and $y-\bar{y}$ are dependent.

4 The Kalman Filter

The Kalman Filter is an optimal recursive data processing algorithm. This is a quite mathematically sophisticated algorithm, and is based on probability theory, functional analysis, and the theory of Stochastic Processes. Since the technique depends upon the process being modelled, there does not exist a general filter for all processes. This algorithm is usually used with a process defined by a system of differential equations. The state variable method is used in this analysis, that is the system is reduced to a system of n first order differential equations. So the state variables might consist of the derivatives of an nth order differential equation. The original example of use had in mind the control of a spacecraft in orbit with position measurements made by earth based stations. The measurement errors were assumed to have zero mean and gaussian distribution. The filter is optimal in a certain sense and solves the problem first posed by Wiener.

R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems", ASME, 1960.

Peter S Maybeck, "Stochastic Models, Estimation, and Control", Academic Press, Volume 1,1979.

5 Kalman

Classic Kalman filter paper, 1960. Received the presidential medal in October 2009. Reference: See the Wikipedia article on Kalman.

6 Norbert Wiener

Norbert Wiener was a child prodigy. His father was a famous linguist. Norbert was born in Columbia Missouri when his father, a Polish emigrant, was at the University of Missouri. Norbert founded much of modern Control Theory, and started the field called Cybernetics. He spent most of his career at MIT. He studied philosophy under Bertrand Russell and mathematics under G. H. Hardy. He was also a social activist, a pacifist, and a quite interesting character. He was a pacifist, but made many contributions during the second world war. Some of his famous books are: Cybernetics, and Time Series. See the Wikipedia article on Norbert Wiener and his two autobiographies.

7 ARIMA Models, (AutoRegressive Integrated Moving Average)

The Box-Jenkins models are ARIMA Models.

8 Box-Jenkins Models

These models are very widely used for smoothing and prediction. They do not necessarily need information about the process that generates the data. See the famous book: **Time Series Analysis: Forecasting and Control**, (3rd Edition) George Box, Gwilym M. Jenkins, Gregory Reinsel, 1994.

9 Multiple Regression

10 Regression in Excel

11 Appendix A, The Cauchy-Schwarz Inequality

Cauchy-Schwarz Inequality. If u and v are two vectors in an inner product space, then

$$|(u,v)| \leq \|u\|\|v\|,$$

and we have equality if and only if u and v are dependent.
Proof. Given u and v, let us perform a Gram-Schmidt orthogonalization, to get a vector w that is orthogonal to v

$$w = u - \frac{(u,v)v}{\|v\|^2}$$

Then

$$u = \frac{(u,v)v}{\|v\|^2} + w$$

is an orthogonal decomposition of u. By the pythagorean theorem

$$\|u\|^2 = \|\frac{(u,v)v}{\|v\|^2}\|^2 + \|w\|^2 \geq \|\frac{(u,v)v}{\|v\|^2}\|^2 = \frac{|(u,v)|^2}{\|v\|^2},$$

with equality iff $w = 0$, that is iff u and v are dependent, that is iff u is a multiple of v.

See the document **Topics In Linear Algebra and Its Applications** in the bibliography for an extended treatment.

12 Bibliography

[1] Greg Welch, Gary Bishop, **An Introduction to the Kalman Filter**, The University of North Carolina at Chapel Hill Department of Computer Science, 2001.

[2] M.S. Grewal, A.P. Andrews, **Kalman Filtering - Theory and Practice Using MATLAB**, Wiley, 2001.

[3] Peter S Maybeck, **Stochastic Models, Estimation and Control**, Volume one.

[4] H. D. Brunk **Introduction to Mathematical Statistics**, Chapter 12 Regression, 2nd edition, 1965, Blaissdell.

[5] Edwin L. Crowe, Francis A Davis, Margaret W. Maxfield, **Statistics Manual**, Fitting a function, Chapter 6, U. S. Navy Ordinace Test Station, Dover Publications, 1960.

[6] Emery James D **Topics In Linear Algebra and Its Applications** stem2.org lineara.pdf file.

[7] Emery James D **Probability Theory** stem2.org, probabilitytheory.pdf, probabilitytheory.tex files.

[8] Emery James D **Least Squares Approximation** stem2.org, lsq.pdf, lsq.tex files.

[9] Emery James D, **Air Permeance** airperm.tex airperm.pdf.

[10] Hogg Robert V, Craig Allen T **Introduction to Mathematical Statistics**, Second Edition, Macmillan, 1966. This book gives a derivation for

the Student's T Distribution and some of its uses.