LECTURE 12

Example 92. It is well known that for two normal random variables, zero covariance implies independence. Why does this not apply to the following situation: $X \sim N(0,1)$, $Cov(X, X^2) = EX^3 - EXEX^2 = 0 - 0 = 0$ but obviously X^2 is totally dependent on X? It is easy to show that independence of two random variables implies zero covariance:

$$Cov(X,Y) = E(XY) - EXEY = EXEY - EXEY = 0$$

The opposite is true only if X and Y are jointly normally distributed which can be checked by calculating the joint density and the product of the marginals. From above we see that, for standard normally distributed random variable X, we have $Cov(X, X^2) = 0$. In this example, zero covariance does not imply independence since the random variable X^2 is not normally distributed.

§55. LINEAR REGRESSION.

The main objective of many statistical investigations is to make predictions, preferably on the basis of mathematical equations. Usually, such predictions require that a formula be found which relates the dependent variable (whose value one wants to predict) to one or more independent variables.

§55-1. THE METHODS OF LEAST SQUARES.

Many engineering and scientific problems are concerned with determining a relationship between a set of variables. In many situations, there is a single response variable Y, also called the dependent variable, which depends on the value of a set of input, also called independent, variables. The simplest type of relationship between the dependent variable Y and the input variables η_1, η_2,..., η_r is a linear relationship. That is, for some constants α, β_1, β_2,..., β_r the equation

$$Y = \alpha + \beta_1 \eta_1 + ... + \beta_r \eta_r$$

would hold.

In this section we begin our study of the case where a dependent variable is to be predicted in terms of a single independent variables, that is we consider a simple case where $r = 1$, thus

$$Y = \alpha + \beta X.$$

a) Suppose that the responses Y_i corresponding to the input values X_i, $i = 1, ..., n$ are to be observed and used to estimate α and β in a simple linear regression model. To determine estimators of α and β we reason as follows: If A is the estimator of α and B of β, then the estimator of the response corresponding to the input variable X_i would be $A + B X_i$. Since the actual response is Y_i, the squared difference is $(Y_i - A - B X_i)^2$, and so if A and B are the estimators of α and β, then the sum of the squared differences between the estimated responses and the actual response values is given by

$$\varphi(A, B) = \sum_{i=1}^{n} (Y_i - A - B X_i)^2.$$

The method of least squares chooses as estimators of α and β the values of A and B that minimize $\varphi(A, B)$. To determine these estimators, we differentiate $\varphi(A, B)$ first with respect to A and then to B as follows:

$$\frac{\partial \varphi(A, B)}{\partial A} = -2 \sum_{i=1}^{n} (Y_i - A - B X_i),$$

$$\frac{\partial \varphi(A, B)}{\partial B} = -2 \sum_{i=1}^{n} X_i \, (Y_i - A - B X_i).$$

Setting these partial derivatives equal to zero yields the following equations for the minimizing values A and B:

$$\sum_{i=1}^{n} Y_i = n\, A + B \sum_{i=1}^{n} X_i, \tag{133}$$

$$\sum_{i=1}^{n} X_i \, Y_i = n\, A \sum_{i=1}^{n} X_i + B \sum_{i=1}^{n} X_i^2. \tag{134}$$

If we let

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i, \qquad \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

then we can write

$$A = \overline{Y} - B\, \overline{X}.$$

Substituting this value of A into equation (134) yields

$$\sum_{i=1}^{n} X_i Y_i = (\overline{Y} - B\,\overline{X})\, n\overline{X} + B \sum_{i=1}^{n} X_i^2$$

or

$$B = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}}{\sum_{i=1}^{n} X_i^2 - n\,\overline{X}^2}.$$

Theorem. The least squares estimators of β and α corresponding to the data set (X_i, Y_i), $i = 1, 2, ..., n$ are, respectively,

$$B = \frac{\sum_{i=1}^{n} X_i Y_i - \overline{X} \sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i^2 - n\,\overline{X}^2}, \tag{135}$$

$$A = \overline{Y} - B\,\overline{X}. \tag{136}$$

The straight line $A + Bx$ is called the estimated regression line.

b) We consider

$$\varphi(\alpha, \beta) = E[Y - \alpha - \beta X]^2$$

and we want to find the values of α and β which realized the minimum of φ.

$$\varphi(\alpha, \beta) = E[Y - EY - \beta(X - EX) + (EY - \beta EX - \alpha)]^2 =$$

$$= \mathrm{Var}(Y) + \beta^2 \mathrm{Var}(X) - 2\beta r_{XY} \sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)} + (EY - \beta EX - \alpha)^2,$$

where r_{XY} is the correlation coefficient between X and Y.

Therefore

$$\frac{\varphi(\alpha, \beta)}{\partial \alpha} = -2(EY - \beta EX - \alpha) = 0,$$

$$\frac{\varphi(\alpha, \beta)}{\partial \beta} = 2\beta \mathrm{Var}(X) - 2r_{XY} \sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)} - 2EX(EY - \beta EX - \alpha) = 0.$$

It follows that

$$\alpha = EY - r_{XY} \sqrt{\frac{\mathrm{Var}(Y)}{\mathrm{Var}(X)}}\, EX, \qquad \beta = r_{XY} \sqrt{\frac{\mathrm{Var}(Y)}{\mathrm{Var}(X)}}.$$

Hence

$$Y = EY + r_{XY} \sqrt{\frac{\mathrm{Var}(Y)}{\mathrm{Var}(X)}}(X - EX)$$