LINEAR PREDICTION OF A RANDOM VARIABLE

Some Important Definitions

Suppose $X$ and $Y$ are random variables. For example, $X =$ Today's Price, $Y =$ Tomorrow's Price.

The **mean**, or expectation of $X$ is denoted by $E[X]=\mu_x$. $E[X]$ is the average value of $X$ over a large number of (hypothetical) repeated observations. Similarly, we write the expectation of $Y$ as $E[Y]=\mu_y$.

The **variance** of $X$ is $var(X)=E[(X-\mu_x)^2]$.

The **covariance** between $X$ and $Y$ is $cov(X,Y)=E[(X-\mu_x)(Y-\mu_y)]$.

The **correlation** between $X$ and $Y$ is

$$
corr(X,Y) = \frac{cov(X,Y)}{\sqrt{var(X)\cdot var(Y)}} .
$$

Some Basic Facts

If $a$ and $b$ are constants (i.e., nonrandom numbers), then

$$
E[aX+bY]=aE[X]+bE[Y]
$$

$$
var[aX+b]=a^2\cdot var(X)
$$

$$
var[X+Y]=\cdot var(X)+\cdot var(Y)+2\cdot cov(X,Y)
$$

$$
var[X-Y]=\cdot var(X)+\cdot var(Y)-2\cdot cov(X,Y)
$$

$$
cov(aX,bY)=ab\cdot cov(X,Y)
$$

$$
cov(X,Y+Z)=cov(X,Y)+cov(X,Z)
$$

$$
corr(aX,bY)=corr(X,Y) .
$$

$$
-1\leq corr(X,Y)\leq 1 .
$$
The Linear Prediction Problem

Given random variables \(X, Y\), suppose we want to predict \(Y\) based on \(X\). For example, \(X\) and \(Y\) might be today’s and tomorrow’s high temperature. Using data from the last month, we would find that warm days tend to be followed by warm days, but not in a completely predictable way. To better understand the relationship, we could fit the least squares line, \(\hat{y} = \hat{a} + \hat{b}x\), where

\[
\hat{b} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\text{Sample cov}(X, Y)}{\text{Sample var}(X)}.
\]

We can measure the strength of the linear relationship between \(X\) and \(Y\) by

\[
R = \frac{\text{Sample cov}(X, Y)}{\sqrt{\text{Sample var}(X) \text{Sample var}(Y)}}.
\]

On the other hand, if we think of \((X, Y)\) as random variables, and we know the means, variances and covariance of \((X, Y)\), we don’t need to collect any data to find the best linear predictor of \(Y\) based on \(X\). The problem is to find \(a\) and \(b\) such that the linear predictor \(\hat{Y} = a + bX\) has the smallest possible mean squared prediction error, \(E[Y - \hat{Y}]^2\). For simplicity, suppose \(\mu_x = \mu_y = 0\). Then it’s not hard to prove that the solution to the problem is \(a = 0\), and

\[
b = \frac{E[XY]}{\text{var}(X)} = \frac{\text{cov}(X, Y)}{\text{var}(X)}.
\]

We can measure the strength of the linear relationship between the two random variables \((X, Y)\) by

\[
corr(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \text{var}(Y)}}.
\]

We see that the solution to the (theoretical) linear prediction problem is just a population version of the usual linear regression formulas described earlier.

A Warning

The best predictor may not be linear. For example, suppose \(X\) is standard normal, and define \(Y = X^2 - 1\). The best predictor of \(Y\) is \(X^2 - 1\). (It’s a perfect predictor!) But \(\mu_x = \mu_y = 0\), and
\begin{align*}
cov(X, Y) &= E[X(X^2 - 1)] = E[X^3 - X] = 0.
\end{align*}

Since \( X, Y \) are uncorrelated, the best \textit{linear} predictor is \textit{zero}. So \( X, Y \) are uncorrelated, but they are \textit{not} independent.
Today’s vs Yesterday’s NASDAQ DAILY INDEX
1 Sept to 30 Nov 1987

Yesterday
Least Squares Line: Today = 1.65 + .990 Yesterday

Today’s vs Yesterday’s GM Returns
1 Sept to 30 Nov 1987