(3) X has MGF $\psi_X(t)$, finite in an open interval around $t=0$

Y has MGF $\psi_Y(t)$,

then $\psi_X(t) = \psi_Y(t) \overset{\text{iff}}{\longleftrightarrow}$ X, Y have identical probability distributions

So the MGF (if it exists) uniquely characterizes a random variable.

Mean versus median	we've already made some contrasts between the mean and the median of a distribution;

here are 2 more things worth saying.

(1) X rv with values in an interval I; (CDF F_X)

$h(x)$ 1-1 function on I, $Y = h(X)$;

if m_X is a median of X (i.e., if $m_X = F_X^{-1}(\frac{1}{2})$), then $h(m_X)$ is a median of $Y = h(X)$. This is not in general true of the mean; as we have already seen:

$$E[h(X)] \neq h[E(X)]$$

unless $h(x) = ax + b$

Prediction

X rv with mean μ_X, SD σ_X

Before X is observed, suppose your job is to <u>predict</u> what its value will be; what should you do? How can you tell if a prediction is <u>good</u>?

Let's say you pick the number $\hat{x}$ ← x-hat (a fixed known constant) before X is observed.

Then, after X arrives, your prediction error would be $(\hat{x} - X)$, which might be either positive or negative. One possible criterion for goodness would be to find $\hat{x}$ such that $E(\hat{x} - X) = 0$.

Def The *bias* of $\hat{x}$ as a prediction for X is $\text{bias}(\hat{x}) \triangleq E(\hat{x} - X)$.

Def Your prediction $\hat{x}$ is *unbiased* if $\text{bias}(\hat{x}) = 0$.

Clearly, to achieve this just choose $\hat{x} = E(X)$.

Another possible criterion for goodness would be to find $\hat{x}$ such that $E(\hat{x}-X)^2$ is small. (Gauss)

Def. $E\left[(\hat{x}-X)^2\right]$ is called the <u>mean squared error</u> (MSE) of $\hat{x}$ as a prediction for X.

Small theorem: The $\hat{x}$ that minimizes MSE is $\hat{x} = E(X)$.

Small proof

$$E\left[(\hat{x}-X)^2\right] = E\left(\hat{x}^2 - 2\hat{x}X + X^2\right)$$

$$= \hat{x}^2 - 2\hat{x}\,E(X) + E(X^2)$$

This is a quadratic function of $\hat{x}$;

$$\frac{d}{d\hat{x}} E\left[(\hat{x}-X)^2\right] = 2\hat{x} - 2E(X) = 0$$

iff $\boxed{\hat{x} = E(X)}$

$$\frac{d^2}{d\hat{x}^2} = 2 > 0$$

so $E(X)$ is a minimum

Also easy to show

$$MSE(\hat{x}) = E(\hat{x} - X)^2$$

$$= V(X) + [\text{bias}(\hat{x})]^2$$

So the choice $\hat{x} = E(X)$ both minimizes $MSE(\hat{x})$ and achieves $\hat{0}$ bias, and with this choice $MSE(\hat{x}) = V(X)$
$= \sigma_X^2$

A different criterion

Yet another possible criterion for a good prediction $\hat{x}$ would be to find $\hat{x}$ such that $E[|\hat{x} - X|]$ is small. (Laplace)

Definition

$E|\hat{x} - X|$ is called the <u>mean absolute error</u> (MAE) of $\hat{x}$ as a prediction for X

Another small theorem

X rv with finite mean μ_X; let m_X be (a/the) <u>median</u> of X;

$\rightarrow$ the $\hat{x}$ that minimizes $MAE(\hat{x})$ is (a/the) <u>median</u> m_X.

Reminder: why a/the?

Careful definition of median

X rv $\rightarrow$ every number m such that

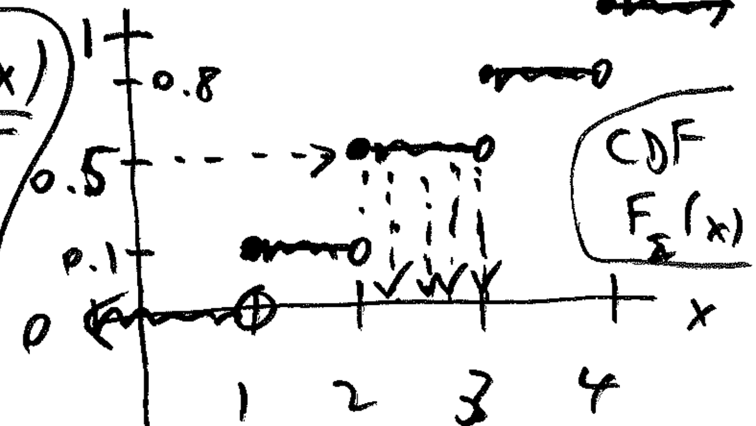
$$P(X \le m) \ge \frac{1}{2} \text{ and } P(X \ge m) \ge \frac{1}{2}$$

is <u>a median</u> of the dist. of X

Example of nonunique median

X discrete on $\{1, 2, 3, 4\}$

x	$P(X=x)$
1	0.1
2	0.4
3	0.3
4	0.2



All $2 \le x < 3$ have $F_X(x) = \frac{1}{2}$

Most people would say median $= 2.5$

Which is a better criterion, MSE or MAE?

There is no universal right answer (210) to this question: it depends on the real-world consequences of your prediction errors $(\hat{X} - X)$; quantifying these consequences involves the creation of a utility function, which we'll briefly examine later.

Covariance & correlation

Independence of 2 or more RVs is a special case of a more general reality, in which (your uncertainty about something) and (your uncertainty about something else) are related.

Let's see how to quantify such relationships.

Def. X, Y rv with finite means $\mu_X = E(X)$ and $\mu_Y = E(Y)$. The covariance of X and Y, written $C(X, Y)$, is defined as

DS use $\mathrm{Cov}(X, Y)$

$$C(X, Y) = E\left[(X - \mu_X)(Y - \mu_Y)\right],$$

as long as this expectation exists

Consequence of this definition

① $(X - \mu_X)\cdot(Y - \mu_Y) =$

$$X \cdot Y - \mu_X \cdot Y - \mu_Y \cdot X + \mu_X \mu_Y$$

so $C(X, Y) = E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X \mu_Y$

$$= E(XY) - \mu_X \mu_Y - \mu_X \mu_Y + \mu_X \mu_Y$$

$$\boxed{C(X, Y) = E(XY) - \mu_X \mu_Y}$$

much easier formula to compute with

(expectation of product − product of expectations)

(2) Sufficient condition for $C(X, Y)$ to exist: $\sigma_X^2 < \infty$ and $\sigma_Y^2 < \infty$.

(3) Covariance is a good start at measuring strength of relationship, but it has a big flaw: its value depends on the units of measurement of X and Y.

Example: X = education level (years of schooling completed)

Y = yearly income (\$)

$C(X, Y)$ comes out in (years) $\cdot$ (\$) ??

Example:

X = max temperature in °C

Y = max relative humidity (%)

If you change your mind & measure temperature X' in $°F = \frac{9}{5}°C + 32°$,

$$C(X', Y) = C\left(\tfrac{9}{5}X + 32, Y\right) \neq C(X, Y)$$

Easy to show that if a, b are fixed constants then $C(a\underline{X}+b, \underline{Y}) = a\,C(\underline{X}, \underline{Y})$ so

$C(\underline{X}', \underline{Y}) = 1.8 \cdot C(\underline{X}, \underline{Y})$, i.e. you can

°F °C

make the association between temperature & relative humidity seem larger just by switching from °C to °F (???)

Easy fix:

Def The process of converting a rv $\underline{X}$ to standard units (SU) is achieved with the linear transformation $\underline{X}' = \dfrac{\underline{X} - E(\underline{X})}{SD(\underline{X})}$

(as long as $\sigma_{\underline{X}} < \infty$, this is a meaningful definition)

$= \dfrac{\underline{X} - \mu_{\underline{X}}}{\sigma_{\underline{X}}}$

$E(\underline{X}') = 0$, $V(\underline{X}') = 1 = SD(\underline{X}')$

Def. X, Y rv with finite variances σ_X^2 and σ_Y^2 (and therefore finite means μ_X and μ_Y) $\rightarrow$ the <u>correlation</u> of X and Y is

$$\rho(X, Y) = E\left[\left(\frac{X - \mu_X}{\sigma_X}\right) \cdot \left(\frac{Y - \mu_Y}{\sigma_Y}\right)\right]$$

$$= \frac{C(X, Y)}{\sigma_X \cdot \sigma_Y}$$

With this definition, the correlation is <u>invariant to linear transformation</u> of either variable (or both): for any constants $a, c > 0$ and b, d,

$$\rho(aX + b, cY + d) = \rho(X, Y).$$

(If $a < 0$, $\rho(aX + b, Y) = -\rho(X, Y)$.)

Consequences of the correlation definition

① Cauchy–Schwarz inequality:

For all rv X, Y for which $E(XY)$ exists, $\left[E(XY)\right]^2 \le \left[E(X)\right]^2 \cdot \left(E(Y)\right)^2$,

Karl Schwarz (1843–1921) German mathematician

from which $\left[C(X,Y)\right]^2 \le \sigma_X^2 \cdot \sigma_Y^2$

and $-1 \le \rho(X, Y) \le +1$.

Def $\rho(X,Y) > 0 \leftrightarrow X, Y$ positively correlated (associated)

$\rho(X,Y) < 0 \leftrightarrow X, Y$ negatively correlated

$\rho(X,Y) = 0 \leftrightarrow X, Y$ uncorrelated

② X, Y independent rv with $\left\{ \begin{array}{l} 0 < \sigma_X^2 < \infty \\ 0 < \sigma_Y^2 < \infty \end{array} \right\}$

$\rightarrow C(X, Y) = \rho(X, Y) = 0$

So independence implies 0 correlation, but (interestingly) not the converse:

Example: $X \sim \text{Uniform}\{-1, 0, +1\}$, $Y \triangleq X^2$.
$E(X) = 0$

→ X, Y clearly dependent since X completely determines Y, but $E(XY) = E(X^3)$
$= E(X) = 0$

(since X and X^3 are identically distributed) and thus

$$C(X, Y) = \underbrace{E(XY)}_{0} - \underbrace{E(X)}_{0} \cdot E(Y) = 0$$

so $\rho(X, Y) = \dfrac{C(X, Y)}{\sigma_X \sigma_Y} = 0$ and X, Y are uncorrelated.

③ X rv with $0 < \sigma_X^2 < \infty$, $Y = aX + b$ for $\left\{ \begin{matrix} a \neq 0 \\ b \end{matrix} \right\}$ constants → $(a > 0)$ $\rho(X, Y) = +1$

$(a<0)\ \rho(X, Y) = -1$ | So $\rho(X, Y)$ (21)

measures the strength of linear association between X and Y.

④ Important:

if X, Y rv, $\sigma_X^2 < \infty$, $\sigma_Y^2 < \infty$ $\longrightarrow$ then

$$V(X+Y) = V(X) + V(Y) + 2\,C(X, Y)$$

⑤ (a, b, c any constants) $\quad C(aX, bY) = ab\,C(X, Y)$

$$\sigma_X^2 < \infty,\ \sigma_Y^2 < \infty \rightarrow V(aX + bY + c) = a^2 V(X) + b^2 V(Y) + 2ab\,C(X, Y)$$

Special case:

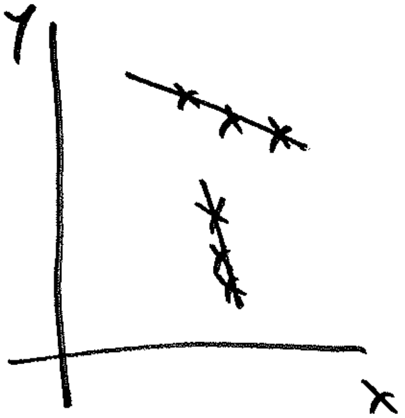
$$V(X-Y) = V(X) + V(Y) - 2C(X, Y).$$

(6) (If) $X_1, \dots, X_n$ such that (X_i, X_j) uncorrelated for all $1 \le i \ne j \le n \rightarrow$ (then) $V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i)$

---

(7)

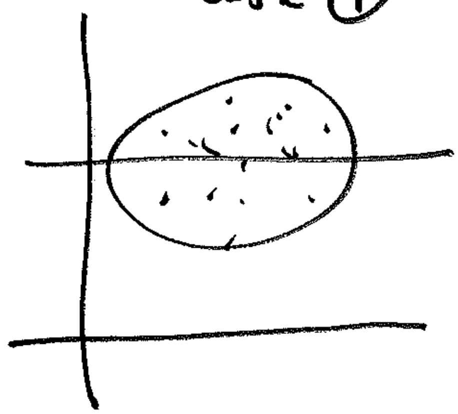
$\rho(X, Y) = -1$



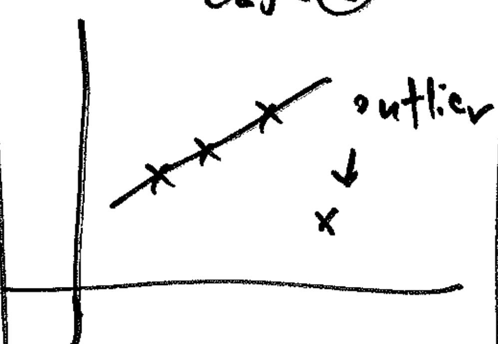
points in scatterplot sample from $f_{X,Y}(x,y)$ all fall on line with negative slope (not necessarily -1)

$\rho(X, Y) = 0$

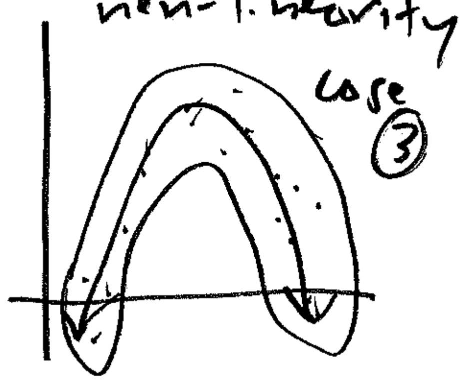
case (1)



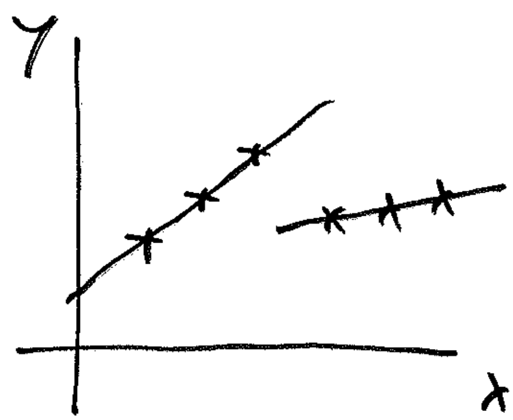
case (2)



non-linearity case (3)



$\rho(X, Y) = +1$

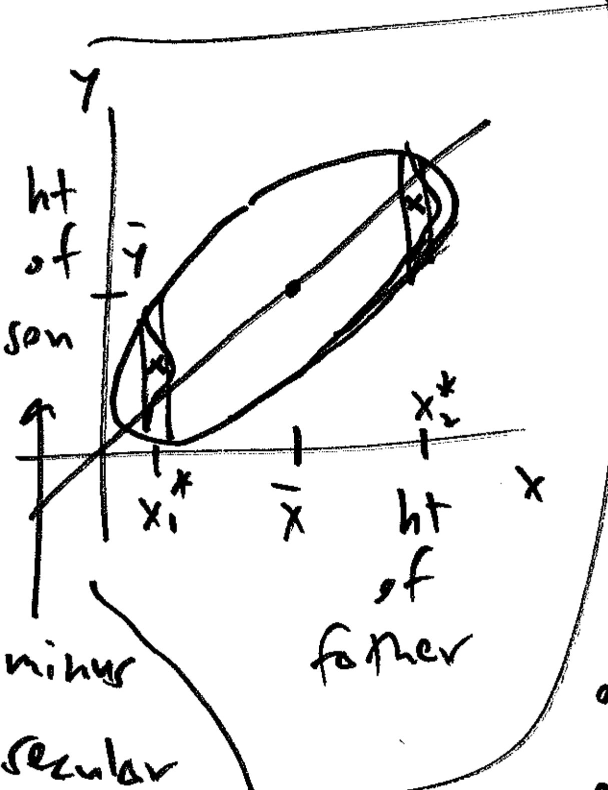


points in scatterplot sample from $f_{X,Y}(x,y)$ all fall on line with positive slope (not necessarily $+1$)

(21 Aug 17)

Conditional Expectation

X, Y related rvs (not independent): then there is information in X for predicting Y; i.e., we should be able to find some function $d: \mathbb{R} \to \mathbb{R}$ such that $d(X)$ is "close" in some sense to Y — what is the optimal d?



Galton example:

Galton divided the elliptical scatterplot up into a bunch of narrow vertical strips, e.g., the one over x_1^* or the other one over x_2^*. The points in the vertical strip over x_2^* are a random sample from the conditional

<u>distribution</u> of Y given $X = x_2^*$, $f_{Y|X}(y \mid x = x_2^*)$

Galton knew about the small theorem

back on p. 207: the number $\hat{w}$ that minimizes the mean squared error (MSE) $E\left[(\hat{w} - W)^2\right]$ of $\hat{w}$ as a prediction for W is $\hat{w} = E(W)$.

So he adopted MSE as his measure of "close" and concluded that the $\hat{y}$ that minimizes the MSE $E\left[(\hat{y} - Y)^2\right]$ in the vertical strip defined by $x \doteq x_2^*$ must be the <u>conditional mean</u>, or <u>conditional expectation</u>, of the rv $(Y \mid X = x_2^*)$

$\boxed{\text{Def.}}$ X, Y rv, Y finite mean →

$$\left\{ \begin{array}{l} \underline{\text{conditional expectation}} \\ \text{(mean) of } Y \text{ given } X = x \end{array} \right\} = E(Y \mid x) \text{ is just}$$

the expectation of the conditional distribution

$f_{Y|X}(y|x)$ of Y given $X = x$,

namely $E(Y|x) = \int_{\mathbb{R}} y\, f_{Y|X}(y|x)\, dy$

for continuous $(Y|X=x)$

and $E(Y|x) = \sum_{\text{all } y} y\, f_{Y|X}(y|x)$

for discrete $(Y|X=x)$

So far, $E(Y|x)$ is just a constant,

equal to the conditional mean of Y

when X is x. the constant

$\boxed{\text{Def.}}$ $h(x) \stackrel{\Delta}{=} E(Y|X=x)$

then the rv $E(Y|X) \stackrel{\Delta}{=} h(X)$ is the

conditional expectation of Y given X.

Clinical trial example, continued

$(n_C + n_T)$ people (a) who are similar in all relevant ways to $\begin{pmatrix} \text{population} \\ \underline{P} \end{pmatrix} = \{$ all adult patients with disease A $\}$

and (b) who consent to participate in your clinical trial are randomized, n_C to $\begin{pmatrix} \text{the control} \\ \text{group} \end{pmatrix}$ C

and n_T to $\begin{pmatrix} \text{the treatment} \\ \text{group} \end{pmatrix}$ T.

outcome of interest is dichotomous:

(success)	1 = disease went into remission
(failure)	0 = didn't

let θ be the proportion of successes you *would* have seen if you *could* have put (everybody in P) into your treatment group; θ is unknown.

---

let $S_i = \begin{cases} 1 & \text{if patient } i \text{ in the actual (T) group had a success} \\ 0 & \text{otherwise} \end{cases}$

Then the rvs $(\mathcal{S}_i \mid \theta)$ are IID Bernoulli(θ)

and the rv $\mathcal{S} = \sum_{i=1}^{n_T} \mathcal{S}_i$ has a conditional

binomial dist: $(\mathcal{S} \mid \theta) \sim$ Binomial (n_T, θ)

It's meaningful to talk about the conditional

expectation r.v. $E(\mathcal{S} \mid \theta) = n_T \theta$ (a linear function of θ),

and – via Bayes's Theorem – it's even more

meaningful to talk about the conditional

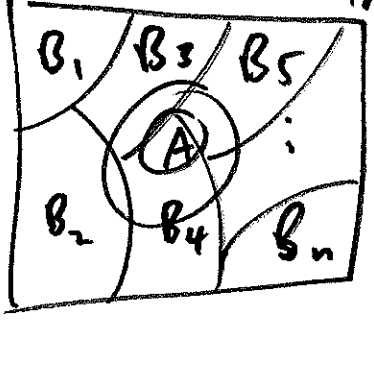
expectation <u>r.v.</u> $E(\theta \mid \mathcal{S})$ (more about this later)

---

and the <u>constant</u> $E(\theta \mid \mathcal{S} = s)$.

Important consequence of the def. of conditional expectation

Remember the Law of Total Prob.!



$$P(A) = \sum_{i=1}^{n} P(B_i)\, P(A \mid B_i)$$

(LTP)

Continuous version of LTP

X, Y continuous r.v. for which all named densities exist $\rightarrow$

$$f_Y(y) = \int_{-\infty}^{\infty} f_X(x) \cdot f_{Y|X}(y|x)\, dx$$

Earlier we agreed that, by definition,

$$E(Y|x) = \int_{-\infty}^{\infty} y\, f_{Y|X}(y|x)\, dx$$

So watch the following slightly magical calculation:

$$E(Y) = \int_{-\infty}^{\infty} y\, f_Y(y)\, dy$$

$$= \int_{-\infty}^{\infty} y \left[\int_{-\infty}^{\infty} f_X(x)\, f_{Y|X}(y|x)\, dx \right] dy$$

if ok to interchange order of integration

$$= \int_{-\infty}^{\infty} f_X(x) \left[\int_{-\infty}^{\infty} y\, f_{Y|X}(y|x)\, dy \right] dx$$

$$= \int_{-\infty}^{\infty} f_X(x) \cdot E(Y \mid x) \, dx \quad (*)$$

, and this is of the form $\left\{ \begin{array}{l} \text{weighted average of } E(Y \mid x), \\ \text{with } f_X(x) \text{ as the weights} \end{array} \right\}$

Recall that for any continuous r.v. W,

$$E(W) = \int_{-\infty}^{\infty} w \, f_W(w) \, dw$$

and

(LOTUS)

$$E\left[h(W)\right] = \int_{-\infty}^{\infty} h(w) \, f_W(w) \, dw$$

so $(*)$ is just

$$E_X\left[E(Y \mid X)\right]$$

and we have shown that (Adam)

$$E(Y) = E_X\left[E(Y \mid X)\right]$$

This is referred to as part 1 of the double expectation theorem; strangely, DS don't even mention that name, calling it instead the LTP for expectations.

I need to postpone examples of these conditional expectation calculations until we've covered more standard distributions.

$\boxed{}$ X, Y r.v. such that $f_{Y|X}(y|x)$ exists → it makes sense to speak not only of $E(Y|x)$, the <u>mean</u> of $f_{Y|X}(y|x)$, but also of the <u>variance</u> of that dist.

$\boxed{\text{Def}}$ $$\overbrace{V(Y|x)}^{\text{the number}} \triangleq E_{X}\left\{\left[Y - E(Y|x)\right]^2 \mid x\right\} = g(x)$$

is called the <u>conditional variance</u> of Y given $X = x$, and the <u>rv</u> $V(Y|X)$ is just $g(X)$, the conditional variance of Y given X.

The payoff from all of this

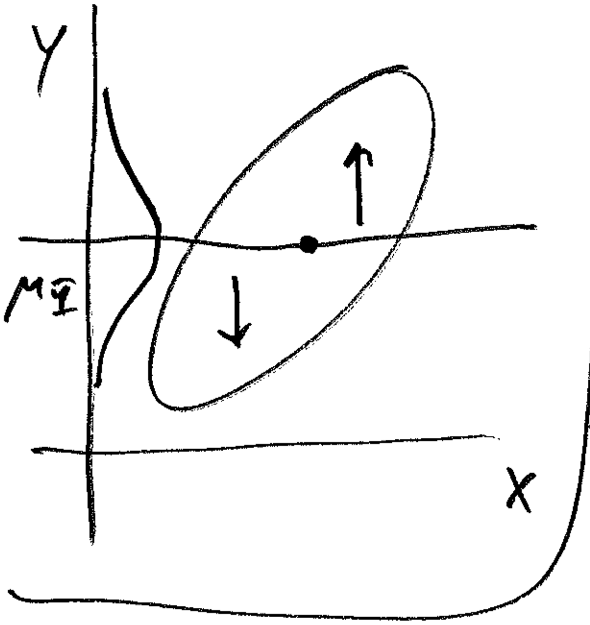
(formalizing Galton's intuition)

Theorem X, Y related rv; want to use some function $\hat{Y} = d(X)$ to predict Y from X $\rightarrow$ the prediction $\hat{Y} = d(X)$ that minimizes the MSE $E(Y - \hat{Y})^2 = E\{[Y - d(X)]^2\}$ is $\hat{Y} = d(X) = E(Y \mid X)$, the conditional expectation of Y given X.

Part 2 of the double expectation theorem

X, Y r.v. such that all of the following expressions exist, $\rightarrow$

$$V(Y) = E_X\left[V(Y \mid X)\right] + V_X\left[E(Y \mid X)\right]. \qquad \text{(Eve)}$$



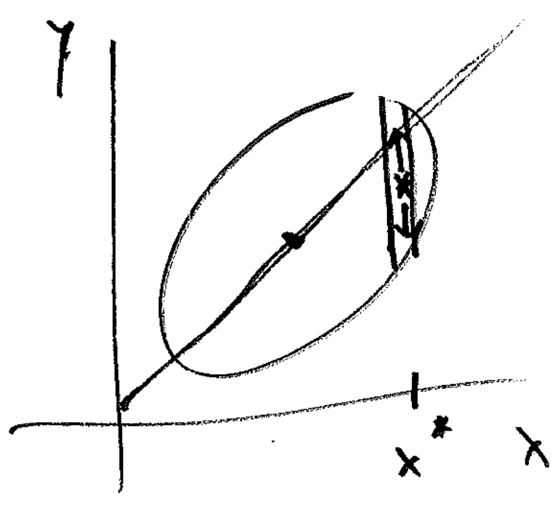
Imagine a 2-part game!

Stage 1 Predict Y without knowing X. Well, if you buy into MSE as your measure of "goodness" of a prediction, we know that you should predict $\hat{Y}_{\text{no } X} = \mu_Y = E(Y)$

and your resulting MSE will be

$$E\left[(Y - \mu_Y)^2\right] = V(Y) = \sigma_Y^2.$$

Stage 2 Observe X, now predict Y



Let's say $X = x^*$. Then we know the MSE-optimal prediction is $\hat{Y}_{X = x^*} = E(Y \mid X = x^*)$

and your resulting MSE will be

$$E\left\{\left[Y - E(Y \mid X = x^*)\right]^2\right\} = \underbrace{V(Y \mid x^*)}_{**} .$$

From the vantage point of someone thinking about stage 2 before it happens, X is not yet known, so the expected value of $**$, namely $E_X\left[V(Y \mid X)\right]$, is the best you can do to guess at how good the stage 2 prediction will be.

The second part of the double expectation theorem says

$$\underbrace{V(Y)}_{\substack{\uparrow \\ \text{MSE of} \\ \hat{Y}_{\text{no } X}}} = \underbrace{E_X\left[V(Y \mid X)\right]}_{\substack{\text{"E(MSE)" of} \\ \hat{Y}_X = E(Y \mid X)}} + \underline{V_X\left[E(Y \mid X)\right]}$$

But since variances are always non-negative,

$$V_X\left[E(Y \mid X)\right] \ge 0, \text{ so}$$

$$\underbrace{E_X\left[V(Y \mid X)\right] + V_X\left[E(Y \mid X)\right]}_{\substack{V(Y) \\ \text{MSE of } \hat{Y}_{\text{no } X}}} \ge \underbrace{E_X\left[V(Y \mid X)\right]}_{\substack{\text{"E(MSE)"} \\ \text{of } \hat{Y}_X}}$$

Thus you always expect your predictive accuracy to get better (or at least stay the same) when you use $E(Y \mid X)$ to predict Y.

Another complete switch in subject!

Utility

Q: How to take action sensibly when the consequences are uncertain?