Regression/Correlation Analysis

interdependence of variables

---

statistics $\longleftrightarrow$ probability

takes distributions

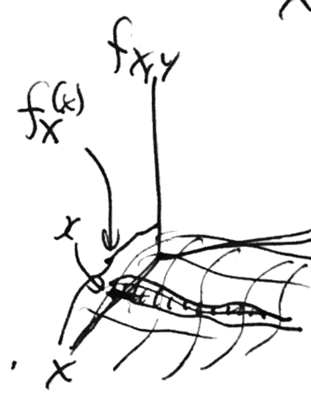
$\downarrow$

Outcomes

Recall: $f_Y(y; X=x)$ = conditional density of Y knowing value of X.

$$= \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

$f_{X,Y}(x,y)$ = joint prob density



marginal density $f_X(x) = \int_Y f_{X,Y}(x,y)\,dy$.

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)}.$$

Conditional expectation $\mathbb{E}(Y|X=x) = \int_Y y\, f_Y(y|X=x)\,dy$

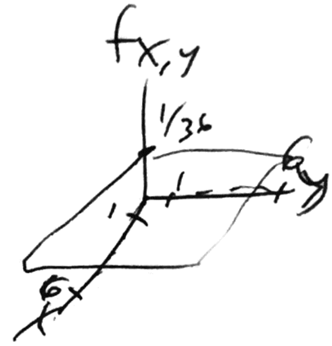
E.g.: X, Y iid dice

If X & Y independent, $\mathbb{E}(Y|X=x) = \mathbb{E}(Y)$.

$f_{X,Y}$

$\mathbb{E}(X) = 3.5$

$X \backslash Y$	1	2	3	4	5	6
1	1/36	.	.	.	.	.
2	–	.	.	.	–	–
3	–	–		–		.
4	–					.
5	–					
6	–				–	



Let $Z = X + Y$ Look at $f_{X,Z}(x,z)$.

$X \backslash Z$	2	3	4	5	6	7	8	9	10	11	12	$f_X(x)$
1	1/36	1/36	.	.	.	.						1/6
2		.	.	.	.	.	.					1/6
3			.	.	.	.	.	.				1/6
4				.	.	.	.	.	.			1/6
5					.	.	.	.	.	.		1/6
6						.	.	.	.	.	.	1/6
$f_Z(z)$	1/36	2/36	3/36	4	5	6	5	4	3	2	1	

36

$$\mathbb{E}(X \mid Z = z) = \int_X x \cdot \frac{f_{X,Z}(x,z)}{f_Z(z)}\, dx = \frac{z}{2}$$

Z	2	3	4	5	6	7	8	9	10	11	12	z
$\mathbb{E}(X \mid Z = z)$	1	$\frac{\frac{1}{36}+\frac{2}{36}}{2/36}$ 1.5	2	2.5	3	3.5					6	$\frac{z}{2}$

$$1 \cdot \frac{\frac{1}{36}}{3/36} + 2 \cdot \frac{\frac{1}{36}}{3/36} + 3 \cdot \frac{\frac{1}{36}}{3/36}$$

2

$\mathbb{E}(Z \mid X=x)$ Exercise: work this out explicitly.

$$\mathbb{E}(Z \mid X=x) = \mathbb{E}(X+Y \mid X=x) = \underbrace{\mathbb{E}(X \mid X=x)}_{x} + \underbrace{\mathbb{E}(Y \mid X=x)}_{3.5}$$

$= x+3.5$

Note: Conditional expectation is a RV!

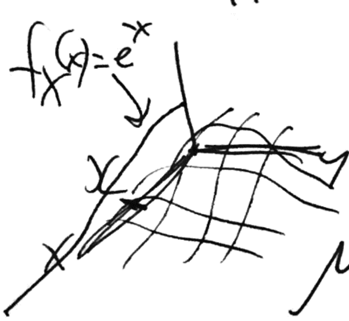
$$\mathbb{E}(Z \mid X) = X + 3.5.$$

Def: ~~RVs~~ A regression (i.e. conditional expectation) is linear if $\mathbb{E}(Y \mid X=x) = \alpha + \beta x$.

Find Bivariate regression equation for Y on X for

$$f_{X,Y}(x,y) = \begin{cases} x e^{-x(1+y)}, & x,y>0 \\ 0 & \text{else.} \end{cases}$$

$x e^{-x} \cdot e^{-xy}$. Not indep.



$$\mu_{Y|X} = \mathbb{E}(Y \mid X) = \int_Y y \, \frac{f_{X,Y}(x,y)}{f_X(x)} \, dy.$$

Step 1: compute marginal density $f_X(x)$.

$$f_X(x) = \int_Y f_{X,Y}(x,y)\,dy = \int_0^\infty x e^{-x} e^{-xy}\,dy =$$

$$= x e^{-x} \left. \frac{e^{-xy}}{-x} \right|_0^{\infty} = + x e^{-x} \frac{1}{x} \quad (x>0) = e^{-x}.$$

$$\mu_{Y|x} = \int_0^{\infty} y \cdot \frac{x \cancel{e^{-x}} e^{-xy}}{\cancel{e^{-x}}} \, dy = x \int_0^{\infty} y e^{-xy} \, dy.$$

$$= \frac{1}{x},$$

Regression curve: $\mu_{Y|x} = 1/x$

<u>E.g.:</u> Find multiple regression of X_2 on X_1, X_3 where

$$f_{X_1,X_2,X_3}(x_1,x_2,x_3) = \begin{cases} (x_1+x_2)e^{-x_3} & 0<x_1,x_2<1 \\ & x_3>0 \\ 0 & \text{else}. \end{cases}$$

$$\mu_{X_2|X_1X_3} = \int_{X_2} x_2 \frac{f_{X_1X_2X_3}(x_1,x_2,x_3)}{f_{X_1X_3}(x_1,x_3)} \, dx_2.$$

joint marginal $f_{X_1X_3}(x_1,x_3) = \int_{X_2} f_{X_1X_2X_3}(x_1,x_2,x_3) \, dx_2 =$

$$= \int_0^1 (x_1+x_2)e^{-x_3} \, dx_2 = e^{-x_3} \left[x_1x_2 + \frac{x_2^2}{2} \right]_{x_2=0}^{1} = e^{-x_3}\left[x_1 + \frac{1}{2}\right]$$

So $$\mu_{X_2|X_1,X_3} = \int_0^1 x_2 \frac{(x_1+x_2)\cancel{e^{-x_3}}}{\cancel{e^{-x_3}}(x_1+\frac{1}{2})} dx_2$$

$$= \frac{1}{x_1+\frac{1}{2}} \int_0^1 (x_1x_2 + x_2^2)\, dx_2$$

$$= \frac{1}{x_1+\frac{1}{2}} \left[x_1 \cdot \frac{x_2^2}{2} + \frac{x_2^3}{3} \right]_{x_2=0}^1 = \frac{x_1 \cdot \frac{1}{2} + \frac{1}{3}}{x_1+\frac{1}{2}}$$

$$= \frac{x_1 + \frac{2}{3}}{2x_1+1}. \Leftarrow \not\Rightarrow \text{ indep of } X_3. \ \boxed{(0<x_1<1)}.$$

If $\mu_{X_2|X_1,X_3}$ is indep of $X_3 \not\Rightarrow X_2$ & X_3 indep.

<u>Thm:</u> Suppose $\mu_{Y|X} = \alpha + \beta x$ is linear.

Then we <u>know</u> α & β as a function of:

$\mathbb{E}(X) = \mu$, $\mathbb{E}(Y) = \nu$, $\mathrm{Var}(X) = \sigma^2$, $\mathrm{Var}(Y) = \tau^2$,

$\mathrm{Cov}(X,Y) = c$.

Pf:
$$\int_X \mu_{Y|X} f_X(x)\,dx = \int_X \int_Y y \frac{f_{X,Y}(x,y)}{f_X(x)}\,dy\, f_X(x)\,dx = \int_X (\alpha+\beta x) f_X(x)\,dx$$

$$\int_X\int_Y y \frac{f_{X,Y}(x,y)}{\cancel{f_X(x)}}\,dy \cdot \cancel{f_X(x)}\,dx = \int_X (\alpha+\beta x) f_X(x)\,dx.$$

$$\boxed{\nu = \mathbb{E}(Y) = \alpha+\beta\,\mathbb{E}(X) = \alpha+\beta\mu}.$$

$$\int_X x \int_Y y \frac{f_{X,Y}(x,y)}{\cancel{f_X(x)}}\,dy\, \cancel{f_X(x)}\,dx = \int_X x(\alpha+\beta x) f_X(x)\,dx$$

$$\mathbb{E}(X\cdot Y) = \boxed{c+\mu\nu = \alpha\mu+\beta\,\cancel{\mathbb{E}(X^2)}\,(\sigma^2+\mu^2)}$$

$$c = \mathrm{Cov}(X,Y) = \mathbb{E}\big((X-\mu)(Y-\nu)\big) = \mathbb{E}(XY)-\mu\nu$$

$$\sigma^2 = \mathrm{Var}(X) = \mathbb{E}\big((X-\mu)^2\big) = \mathbb{E}(X^2)-\mathbb{E}(X)^2 = \mathbb{E}(X^2)-\mu^2$$

$$\alpha + \beta\mu = \nu \Rightarrow -\mu\alpha - \mu^2\beta = -\mu\nu$$

$$\mu\alpha + (\sigma^2 + \mu^2)\beta = c + \mu\nu$$

$$\Rightarrow \sigma^2\beta = c \Rightarrow \boxed{\beta = \frac{c}{\sigma^2}}$$

$$\alpha = \nu - \mu\beta = \boxed{\nu - \frac{\mu c}{\sigma^2} = \alpha}$$

Thm: If $\mu_{Y|X}$ = linear, then,

$$\mu_{Y|X} = \nu - \frac{\mu c}{\sigma^2} + \frac{c}{\sigma^2}x.$$

$$\Leftrightarrow \mu_{Y|X} - \nu = \frac{c}{\sigma^2}(x - \mu).$$