Lec 7 Math 390 2/21/19 $x_{raw} \in \{Red, Green\}$ (1

$x \in \{0, 1\}$ binary encoding

$g(x) = \begin{cases} \bar{y}_r & \text{if } x = red \\ \bar{y}_g & \text{if } x = green \end{cases}$ this is the OLS soln!

$$\Rightarrow \hat{y} = \underbrace{\bar{y}_{red}}_{b_0} + \underbrace{(\bar{y}_{green} - \bar{y}_{red})}_{b_1} x$$

Can we prove this? let $\sum x_i := n_g$ # green $\Rightarrow n - n_g = n_r$ # red

$$\bar{x} = \frac{\sum x_i}{n} = \frac{n_g}{n} = p \text{, prop. of green}$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{\sum_{i:green} y_i + \sum_{i:red} y_i}{n} = \frac{\sum_{i:green} y_i}{n} \cdot \frac{n_g}{n_g} + \frac{\sum_{i:red} y_i}{n} \cdot \frac{n_r}{n_r}$$

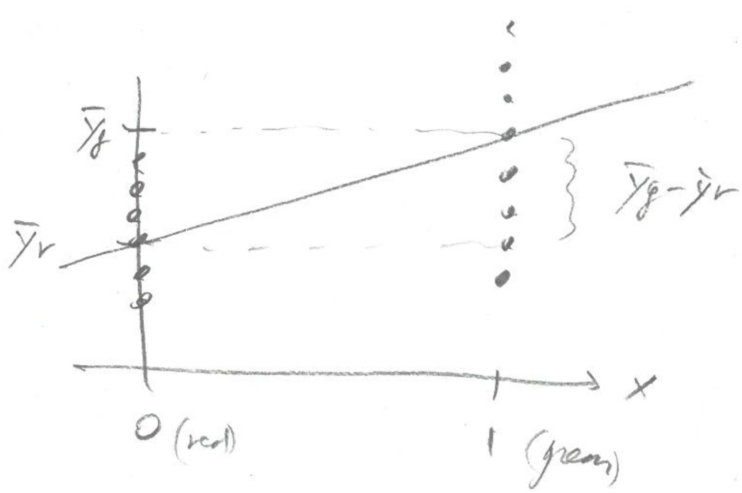
$$= \bar{y}_{green} \cdot p + \bar{y}_{red}(1-p)$$

$$b_1 = \frac{\overbrace{\sum x_i y_i}^{n_g \bar{y}_g} - n \overbrace{\bar{x}}^{p} \bar{y}}{\underbrace{\sum x_i^2}_{n_g} - n \underbrace{\bar{x}^2}_{p^2}} = \frac{n_g \bar{y}_g - n p \cdot \bar{y}}{n_g - n p^2} \cdot \frac{\frac{1}{n}}{\frac{1}{n}} = \frac{p \bar{y}_g - p \bar{y}}{p - p^2} = \frac{\bar{y}_g - \bar{y}}{1-p}$$

$$= \frac{\bar{y}_g - (\bar{y}_g p + \bar{y}_r (1-p))}{1-p} = \frac{(1-p)\bar{y}_g - (1-p)\bar{y}_r}{1-p} = \bar{y}_g - \bar{y}_r$$

$$b_0 = \bar{y} - b_1 \overbrace{\bar{x}}^{p} = p\bar{y}_g + (1-p)\bar{y}_r - (\bar{y}_g - \bar{y}_r)p = (1-p)\bar{y}_r + p\bar{y}_r = \bar{y}_r \checkmark$$

Visual



Note: x_{raw} has $L = 2$ but only one intercept is necessary. Why? Intercept takes place of the red category. This is called a "reference level". Red is the reference level. b_1 is effect of green rel. to red. Not "overall effect of red".

What if $x_{raw} = \{\text{red, green, blue}\}$

We can create two dummies $x_1 = \mathbb{1}_{x_{raw} = \text{green}}$, $x_2 = \mathbb{1}_{x_{raw} = \text{blue}}$,

If $x_1 = x_2 = 0 \Rightarrow x_{raw} = \text{red}$. Red is the reference category.

$\mathcal{H} = \{w_0 + w_1 x_1 + w_2 x_2 : w_0, w_1, w_2 \in \mathbb{R}\}$ $p = 2$! Even though one variable.

The answer will be the same

$b_0 = \bar{y}_r$

$b_1 = \bar{y}_g - \bar{y}_r$ Can $x_1 = 1$ & $x_2 = 2$ at the same time? No...

$b_2 = \bar{y}_b - \bar{y}_r$ So you never add $b_0 + b_1 + b_2$ all together.

This can be genralized to any categorical variable.

Does it matter if variable is nominal or ordinal?

No... except if you force the relationship in y to be monotonic too....

if $x_{raw} = \{\text{low, medium, high}\}$

set reference category

$\hat{y}(\text{low}) = \bar{y}_L$

$\hat{y}(\text{medium}) = \bar{y}_M$ if I force $\hat{y}(\text{low}) \le \hat{y}(\text{medium}) \le \hat{y}(\text{high})$,

$\hat{y}(\text{high}) = \bar{y}_H$ this OLS alg. does ~~not~~ always work.

Hard problem!

Consider two r.v.'s X, Y. They are dep. if $\exists x_1, x_2$ s.t. $P(Y|X=x_1) \neq P(Y|X=x_2)$

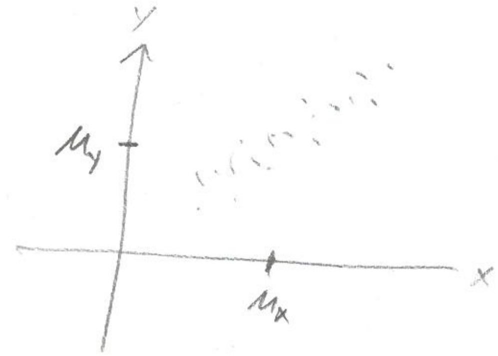
AKA "association"

Review: ρ, r, σ_{xy}, s_{xy}

correlation

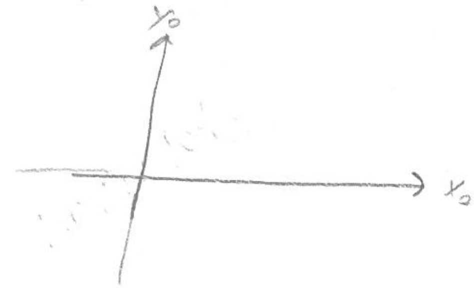
$\rho = \mathrm{Corr}[X,Y] := \frac{\sigma_{xy}}{\sigma_x \sigma_y}$ estimated by $r :=$

$\sigma_{xy} := E[(X-\mu_X)(Y-\mu_Y)]$ estimated by s_{xy}

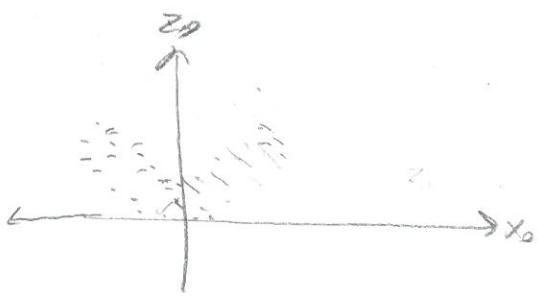


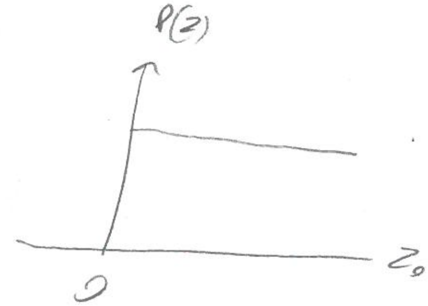
$\Longrightarrow$

let $X_0 := X - \mu_X$, $Y_0 = Y - \mu_Y$



let $Z_0 = X_0 Y_0$



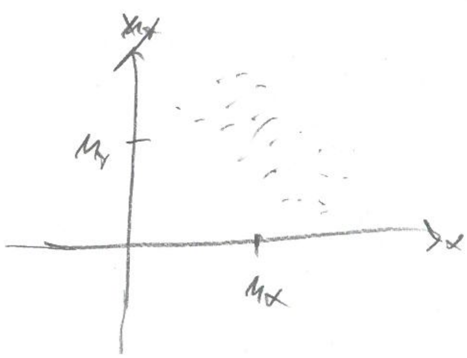


$\sigma_{XY} := E[(X-\mu_X)(Y-\mu_Y)] > 0$

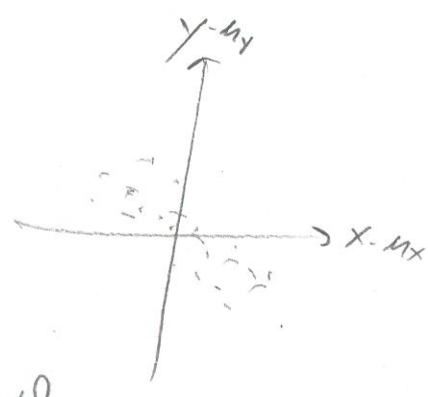
$\parallel$

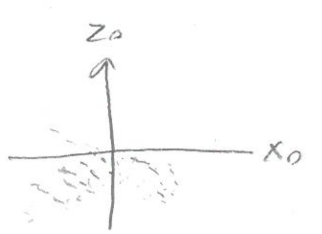
$E[Z] > 0$

if lin. ass & + slope

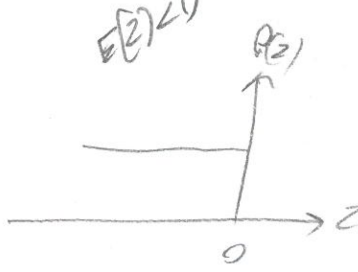


$\Longrightarrow$





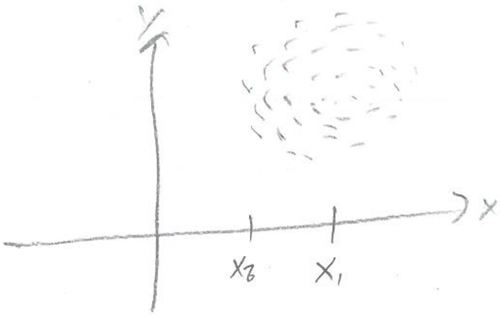
$\Rightarrow$



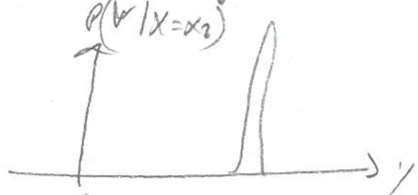
$\sigma_{XY} < 0$

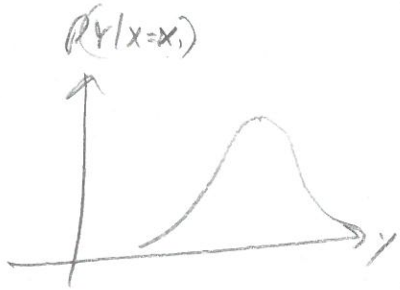
if lin. ass & - slope

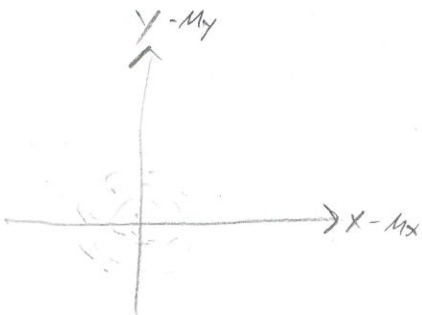
What if

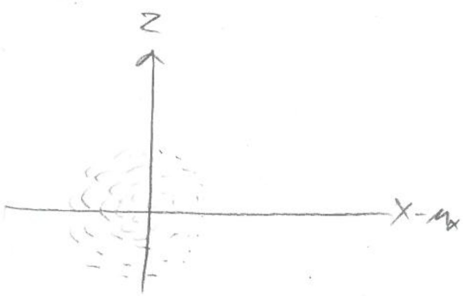


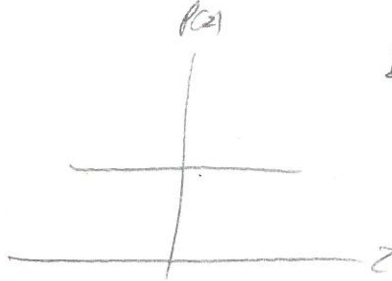
Clearly dep. (& assoc.)











$E[Z] \approx 0 \Rightarrow \sigma_{XY} \approx 0$

assoc. $\not\Rightarrow$ correlation (linear assoc)

better name for "corr."

corr $\in$ {association types}

$\rho \in [-1,1]$, $r \in [-1,1]$

For $p=1$ $R^2 = r^2$ (this's where it gets its name from)

on HW.

midterm 1 ↑

Midterm 2 ↓

$\mathcal{Y} = \mathbb{R}$

OLS with $p = 2$ $\Rightarrow$ $\mathcal{H} = \{ w_0 + w_1 x_1 + w_2 x_2 : w_0, w_1, w_2 \in \mathbb{R} \}$

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - w_0 - w_1 x_1 - w_2 x_2)^2 = \ldots \text{ long...}$$

then $\underset{w_0}{\text{argmin}}\{SSE\}$, $\underset{w_1}{\text{argmin}}\{SSE\}$, $\underset{w_2}{\text{argmin}}\{SSE\}$ lots of derivatives

Easier way?

$$\mathbb{D} = \langle X, y \rangle, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix} \in \mathbb{R}^{n \times \overset{p+1}{3}}$$

$$\vec{\hat{y}} = X\vec{w} = \begin{bmatrix} w_0 + w_1 x_{11} + w_2 x_{12} \\ w_0 + w_1 x_{21} + w_2 x_{22} \\ \vdots \\ w_0 + w_1 x_{n1} + w_2 x_{n2} \end{bmatrix}$$

$$\vec{e} = \vec{y} - \vec{\hat{y}} \qquad (a+b)^T = a^T + b^T$$

Recall $SSE = \sum e_i^2 = \vec{e}^T \vec{e} = (\vec{y} - \vec{\hat{y}})^T (\vec{y} - \vec{\hat{y}}) = (\vec{y}^T - \vec{\hat{y}}^T)(\vec{y} - \vec{\hat{y}})$

$$= \vec{y}^T \vec{y} - \vec{\hat{y}}^T \vec{y} - \vec{y}^T \vec{\hat{y}} + \vec{\hat{y}}^T \vec{\hat{y}} \qquad (a^T b)^T = b^T a$$

$$= \vec{y}^T \vec{y} - 2\vec{\hat{y}}^T \vec{y} + \vec{\hat{y}}^T \vec{\hat{y}}$$

$$= \vec{y}^T \vec{y} - 2 \vec{w}^T X^T \vec{y} + \vec{w}^T X^T X \vec{w}$$

Now we need to do argmin which means we need $\frac{\partial}{\partial w_0}[SSE] \ldots$

Imagine taking deriv. of whole vector

$$\underbrace{\frac{\partial SSE}{\partial \vec{w}}}_{\text{scalar}} := \begin{bmatrix} \frac{\partial}{\partial w_0}[SSE] \\ \frac{\partial}{\partial w_1}[SSE] \\ \frac{\partial}{\partial w_2}[SSE] \end{bmatrix}$$ "the vector derivative"

Let's get some rules for this..

let $\vec{x}$ have length n

$$\frac{\partial}{\partial \vec{x}}[a] \quad \text{where } a \text{ is a constant wrt } \vec{x}$$

$$= \begin{bmatrix} \frac{\partial}{\partial x_1}[a] \\ \vdots \\ \frac{\partial}{\partial x_n}[a] \end{bmatrix} = \vec{0}_n$$

let $\vec{a} \in \mathbb{R}^n$ const wrt $\vec{x}$

$$\frac{\partial}{\partial \vec{x}}[\vec{a}^T\vec{x}] = \begin{bmatrix} \frac{\partial}{\partial x_1}[a_1x_1 + a_2x_2 + \dots + a_nx_n] \\ \frac{\partial}{\partial x_2}[\quad] \\ \vdots \\ \frac{\partial}{\partial x_n}[\quad] \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \vec{a}$$

$\vec{a}^T\vec{x} = \vec{x}^T\vec{a}$

NOTE: $\neq \vec{a}^T$ different dimension

let b constant wrt to x

$$\frac{\partial}{\partial \vec{x}}[a f(\vec{x}) + b g(\vec{x})] = \begin{bmatrix} \frac{\partial}{\partial x_1}[a f(\vec{x}) + b g(\vec{x})] \\ \vdots \\ \frac{\partial}{\partial x_n}[a f(\vec{x}) + b g(\vec{x})] \end{bmatrix} = \begin{bmatrix} a \frac{\partial}{\partial x_1}[f(\vec{x})] + b\frac{\partial}{\partial x_1}[g(\vec{x})] \\ \\ \\ \end{bmatrix} = a\frac{\partial}{\partial \vec{x}}[f(\vec{x})] + b\frac{\partial}{\partial \vec{x}}[g(\vec{x})]$$

quadratic form

let $A \in \mathbb{R}^{n \times n}$ constant wrt $\vec{x}$

$$\frac{\partial}{\partial \vec{x}}[\vec{x}^T A \vec{x}]$$ & symmetric

$$A\vec{x} = \begin{bmatrix} \leftarrow \vec{a}_1. \rightarrow \\ \leftarrow \vec{a}_2. \rightarrow \\ \leftarrow \vec{a}_n \rightarrow \end{bmatrix}\begin{bmatrix} \uparrow \\ \vec{x} \\ \downarrow \end{bmatrix} = \begin{bmatrix} \vec{a}_1. \vec{x} \\ \vec{a}_2. \vec{x} \\ \vdots \\ \vec{a}_n \vec{x} \end{bmatrix}$$

$$A\vec{x} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n \\ \vdots \\ a_{n1}x_1 + a_{12}x_2 + \dots + a_{nn}x_n \end{bmatrix} =$$

$$X^T(A\vec{x}) = [x_1\ x_2 \dots x_n]\begin{bmatrix}\vec{a}_1\cdot\vec{x}\\ \vec{a}_2\cdot\vec{x}\\ \vdots \\ \vec{a}_n\cdot\vec{x}\end{bmatrix} = x_1\vec{a}_1\cdot\vec{x} + x_2\vec{a}_2\cdot\vec{x} + \dots + x_n\vec{a}_n\vec{x}$$

$$x_1(a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n) + x_2(a_{21}x_1 + a_{22}x_2 + \dots) + \dots +$$

same due to symmetry

$$\frac{\partial}{\partial \vec{x}}[\ ] = \frac{\partial}{\partial x_1}[\ ] = 2a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n + a_{21}x_2 + \dots + a_{n1}x_n$$

$$= 2(a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n) = 2\vec{a}_1\cdot\vec{x}$$

$$\frac{\partial}{\partial x_2}[\ ] = a_{12}x_1 + (a_{21}x_1 + 2a_{22}x_2 + a_{23}x_3 + \dots + a_{2n}x_n) + a_{32}x_3 + \dots + a_{n2}x_2$$

$$= 2(a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n)$$

$$= 2\vec{a}_2\cdot\vec{x}$$

$$\frac{\partial}{\partial \vec{x}}[x^TAx] = 2\begin{bmatrix}\vec{a}_1\cdot\vec{x}\\ \vec{a}_2\cdot\vec{x}\\ \vdots \\ \vec{a}_n\cdot\vec{x}\end{bmatrix} = 2A\vec{x}$$

Math lesson over...

$$\frac{\partial}{\partial \vec{w}}[SSE] = \frac{\partial}{\partial \vec{w}}\left[\vec{y}^T\vec{y} - 2\vec{w}^TX^Ty + \vec{w}^TX^TX\vec{w}\right]$$

symmetric?

$$(X^TX)^T = X^T(X^T)^T = X^TX \checkmark$$

$$= \frac{\partial}{\partial \vec{w}}[\underbrace{\vec{y}^T\vec{y}}_{\text{const}}] - 2\frac{\partial}{\partial \vec{w}}[\vec{w}^T(X^T\vec{y})] + \frac{\partial}{\partial \vec{w}}[\vec{w}^T\underbrace{(X^TX)}_{3\times3}\vec{w}]$$

$$= -2X^T\vec{y} + 2(X^TX)\vec{w} \overset{\text{set}}{=} 0$$

$$\Rightarrow (X^TX)\vec{w} = X^T\vec{y}$$

OLS estimate

$$(X^TX)^{-1}(X^TX)\vec{w} = (X^TX)^{-1}X^T\vec{y} \Rightarrow \boxed{\vec{b} = (X^TX)^{-1}X^T\vec{y}}$$

Proof valid for all p.

We assumed $X^\top X$ is invertible. When? only when

$\text{rank}[X] = p+1$ Note: $\text{rank}(X) \not> p+1$ can't be greater!

$\quad = \dim[\text{colsp}[X]] \not> p+1$ only $p+1$ col's!

We want to prove $\boldsymbol{X}^T\boldsymbol{X}$ is invertible only when $\text{rank}\left[\boldsymbol{X}\right] = p+1$ i.e. the design matrix is "full rank". This is equivalent to proving that

$$\text{rank}\left[\boldsymbol{X}\right] = p+1 \implies \text{rank}\left[\boldsymbol{X}^T\boldsymbol{X}\right] = p+1$$

Logically equivalent is the contrapositive:

$$\text{rank}\left[\boldsymbol{X}^T\boldsymbol{X}\right] \neq p+1 \implies \text{rank}\left[\boldsymbol{X}\right] \neq p+1$$

Not equal in this case is equivalent to less than because a matrix cannot have a rank that exceeds its number of columns since $\text{rank}\left[\boldsymbol{X}\right] := \dim \text{colsp}\left[\boldsymbol{X}\right]$ so the above is equivalent to:

$$\text{rank}\left[\boldsymbol{X}^T\boldsymbol{X}\right] < p+1 \implies \text{rank}\left[\boldsymbol{X}\right] < p+1$$

Beginning with the premise on the left hand side, a rank-deficient matrix has at least one non-trivial (i.e. non-zero) vector $\boldsymbol{v} \in \mathbb{R}^{p+1}$ that maps to the zero vector, i.e. there is at least one direction in the nullspace:

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{v} = \boldsymbol{0}_{p+1}$$

Saqib noticed we can multiply both sides on the right by $\boldsymbol{v}^\top$ to arrive at:

$$\boldsymbol{v}^\top\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{v} = \boldsymbol{v}^\top\boldsymbol{0}_{p+1} = 0$$

We can write this alternatively as

$$(\boldsymbol{X}\boldsymbol{v})^\top\boldsymbol{X}\boldsymbol{v} = 0 \implies \sum_{i=1}^{n}(\boldsymbol{X}\boldsymbol{v})_i^2 = 0$$

If all elements of $\boldsymbol{X}\boldsymbol{v}$ squared and summed yield zero, every single element must be zero and thus,

$$\boldsymbol{X}\boldsymbol{v} = \boldsymbol{0}_n$$

indicating that the vector $\boldsymbol{v}$ (which was assumed to be nontrivial above) is in the nullspace of $\boldsymbol{X}$ indicating that $\boldsymbol{X}$ is rank deficient and thus $\text{rank}\left[\boldsymbol{X}\right] < p+1$. ■