

Lesson 16 Math 390.03-02 4/6/16

n draws from bivariate dist (X, Y)
 $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$

$X \xrightarrow[\text{through}]{\text{directly change}} Y$
 n mechanisms
 f and error ε

$$Y = f(X) + \varepsilon$$

We limit $f \in \mathcal{F}_{\text{lin}} := \{ \beta_0 + \beta_1 x \text{ s.t. } \beta_0, \beta_1 \in \mathbb{R} \} \subset \mathcal{F}$

and $\varepsilon \neq h(X)$
 not a function of X

too easily
 control w/it
 dep/indep is
 prob.

all possible
 functions
 of X is
 too big if n
 spec to work
 with

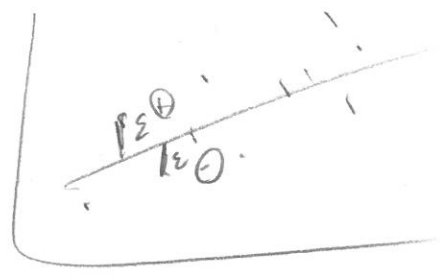
Forget everything about r.v.'s and prob for
 n moments...

$$\left. \begin{aligned} Y_1 &= \beta_0 + \beta_1 X_1 \\ Y_2 &= \beta_0 + \beta_1 X_2 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 X_n \end{aligned} \right\}$$

underdetermined
 system
 need ε to
 make these
 into eq's.

Gauss (1795)
 Legendre (1805)

let's try to figure out β_0 and β_1



let's try to minimize the error terms

$\hat{\beta}_0, \hat{\beta}_1 = \arg \min_{\beta_0, \beta_1}$

$\sum_{i=1}^n \ell(\epsilon_i)$ loss function

Laplace's idea... non-unique

$L_1: \ell_1(\epsilon) = |\epsilon|$ absolute loss

$L_2: \ell_2(\epsilon) = \epsilon^2$ sqd error / gaussian loss

$L_{0.1}: \ell_{0.1}(\epsilon) = \begin{cases} 0 & \text{if } |\epsilon| \leq c \\ 1 & \text{if } |\epsilon| > c \end{cases}$

lets go with L_2 loss

$$\hat{\beta}_0, \hat{\beta}_1 = \arg \min_{\beta_0, \beta_1} \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

error is diff between lin model and actual y
 $y_i^2 - 2y_i\bar{y} + \bar{y}^2$

$$(n-1)S_y^2 = \sum (y_i - \bar{y})^2$$

$$= \sum y_i^2 - 2\bar{y} \sum y_i + \sum \bar{y}^2$$

$$= \sum y_i^2 - 2\bar{y}^2 + n\bar{y}^2$$

$$= \sum y_i^2 - n\bar{y}^2$$

$$= \sum y_i^2 - 2y_i\beta_0 - 2y_i x_i \beta_1 + \beta_0^2 + 2\beta_0\beta_1 x_i + \beta_1^2 x_i^2$$

$$= \sum y_i^2 + \beta_1^2 \sum x_i^2 + n\beta_0^2 - 2\bar{y}n\beta_0 - 2\beta_1 \sum x_i y_i + 2\beta_0\beta_1 \bar{x}n$$

$$\frac{\partial}{\partial \beta_0} = -2\bar{y}n - 2\sum x_i y_i + 2n\beta_0 + 2\beta_1 \bar{x}n = 0 \Rightarrow \beta_0 = \bar{y} + \frac{\sum x_i y_i}{n} - \beta_1 \bar{x}$$

$$\frac{\partial}{\partial \beta_1} = 2\beta_1 \sum x_i^2 - 2\sum x_i y_i + 2\beta_0 \bar{x}n = 0 \Rightarrow \beta_1 = \frac{\sum x_i y_i}{\sum x_i^2} - \beta_0 \bar{x}$$

$$r := \frac{s_{xy}}{s_x s_y} \Rightarrow r s_x s_y = s_{xy} := \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - x_i \bar{y} - y_i \bar{x} + \bar{y} \bar{x}$$

I mean $(n-1)s_{xy}$ = here $= \sum y_i x_i - n \bar{x} \bar{y} + n \bar{y} \bar{x} + n \bar{y} \bar{x}$

$$\sum \epsilon_i^2 = s_y^2 + n \bar{y}^2 + \beta_1^2 s_x^2 + \beta_1^2 n \bar{x}^2 + n \beta_0^2 - 2 \bar{y} n \beta_0 - 2 \beta_1 s_{xy} + 2 \beta_1 n \bar{y} \bar{x} + 2 \beta_0 \beta_1 n \bar{x}$$

This should be $(n-1)s^2_y$ instead of just s^2_y

$$b_1 := \frac{\partial}{\partial \beta_1} \left(\sum \epsilon_i^2 \right) \stackrel{!}{=} 0$$

$$\Rightarrow 2 \beta_1 (s_y^2 + n \bar{x}^2) - 2 s_{xy} - 2 n \bar{y} \bar{x} + 2 \beta_0 n \bar{x} = 0$$

$$\Rightarrow b_1 = \frac{s_{xy} + n \bar{y} \bar{x} - \beta_0 n \bar{x}}{s_y^2 + n \bar{x}^2} = \frac{s_{xy} + n \bar{y} \bar{x} - (\bar{y} - b_1 \bar{x}) n \bar{x}}{s_y^2 + n \bar{x}^2}$$

$$b_0 := \frac{\partial}{\partial \beta_0} \left(\sum \epsilon_i^2 \right) \stackrel{!}{=} 0$$

$$\Rightarrow 2 n \beta_0 - 2 \bar{y} n + 2 \beta_1 n \bar{x} = 0$$

$$b_0 = \frac{\bar{y} n - \beta_1 n \bar{x}}{n} = \bar{y} - b_1 \bar{x}$$

$$\Rightarrow s_{xy} + n \bar{y} \bar{x} - n \bar{x} \bar{y} + b_1 n \bar{x}^2 = b_1 s_y^2 + b_1 n \bar{x}^2$$

$$\Rightarrow b_1 = \frac{s_{xy}}{s_y^2} = \frac{r s_x s_y}{s_y^2} = r \frac{s_x}{s_y}$$

The $(n-1)$'s cancel out between numerator and denominator here

"LS. estimator"

Gauss method of L.S.

line of least fit

$$\hat{y}_i = b_0 + b_1 x_i$$

best est of β_0

" β_1

$$b_0 = \beta_0, b_1 = \beta_1 \quad \text{No...}$$

The true formula is

$$b_1 = r s_y / s_x \text{ or}$$

$$b_1 = s_{xy} / s_x^2$$

I made a mistake somewhere.

Now... r.v. $\langle X, Y \rangle$ joint density ~~st.~~

$$Y = \beta_0 + \beta_1 X + \epsilon \quad \text{where } E[\epsilon] = 0$$

$$E[Y] = E_X[E[Y|X]] = E_X[\beta_0 + \beta_1 X + 0] = \beta_0 + \beta_1 E[X]$$

What if $\varepsilon_1, \dots, \varepsilon_n \stackrel{iid}{\sim} N(0, \sigma^2)$ (DLS Assumption)

Condition on X

to keep things simple (but other versions - PhD n stuff)

(Ord least sqs)

- ① Normality of errors
 - ② Independence of errors
 - ③ Homoskedasticity all var = σ^2
- Independence of X ~~is not~~

$$Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2) \quad \text{reg'd}$$

Why is normality reasonable? CLT

In class X_2, \dots, X_p I called Z_1, \dots, Z_p (I like the Z's better)

Consider $Y = \alpha + \beta_1 X + \beta_2 X_2 + \dots + \beta_p X_p$ (no ε here)

when X_2, \dots, X_p are unobserved but i.i.d.

$$\Rightarrow E[Y|X] = \beta_1 X + \underbrace{\left(\alpha + \sum_{i=2}^p \beta_i E[X_i] \right)}_{\beta_0}$$

$$\text{Var}[Y|X] = \underbrace{\sum_{i=2}^p \beta_i^2 \text{Var}[X_i]}_{\sigma^2}$$

$$\alpha + \beta_2 X_2 + \dots + \beta_p X_p \stackrel{d}{\sim} N(\beta_0, \sigma^2)$$

let $\varepsilon = \sum_{i=2}^p \beta_i X_i - \beta_0$ Noise is the result of not seeing a bunch of stuff
(philosophical point)

Back to $y|x \sim N(\beta_0 + \beta_1 x, \sigma^2)$

What is $\hat{\beta}_0^{MLE}$, $\hat{\beta}_1^{MLE}$?

$$L(\beta_1 | x, y) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - (\beta_0 + \beta_1 x_i))^2}$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2} \sum (y_i - \beta_0 - \beta_1 x_i)^2}$$

$$l(\beta_1 | x, y) = \ln(\downarrow) = -\frac{1}{2\sigma^2} \downarrow$$

$$l'(\beta_1 | x, y) = -\frac{1}{\sigma^2} (2\beta_1 (s_x^2 - n\bar{x}^2) - 2s_{xy} - 2n\bar{x}\bar{y} + 2\beta_0 \bar{x}) = 0$$

Same thing with the (n-1) terms when I did this derivation in class

$$l'(\beta_0 | x, y) = -\frac{1}{\sigma^2} (2n\beta_0 - 2\sum y_i + 2\beta_1 \sum x_i) = 0$$

$$\hat{\beta}_0^{MLE} = b_0 \text{ and } \hat{\beta}_1^{MLE} = b_1$$

the MLE's = the LS estimates!

L2 loss is $l(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ seems to be "natural" once again...

let β_0 be the estimator (r.v) for b_0 $\beta_0 \sim ?$
let β_1 " " " " " " b_1 $\beta_1 \sim ?$ (binned)

$$E(\beta_1) = \dots \beta_1 \quad E(\beta_0) = \dots \beta_0 \quad (\text{unbiased})$$

We will see soon why this is...

This works if $Y \in \mathbb{R}$... What about $Y \in \{0,1\}$ i.e. binary?

$P(Y=1|x) \in (0,1)$ by def of the param. ^{space} Does it make if $P(Y=0|x)$?

$$Y \sim \text{Bern}(P(Y=1|x))$$

model

No 0/1 arbitrary ... labels
Can be flipped. Usually
you know who you want
to predict

linear models $\in \mathbb{R}$

prob $\in [0,1]$

What to do? Introduce "link function", the most
famous being:

$$l := \text{logit}(P(Y=1|x)) := \ln\left(\frac{P(Y=1|x)}{1-P(Y=1|x)}\right) = \beta_0 + \beta_1 x$$

$$P(Y=1|x) = \text{logit}^{-1}(l) = \frac{e^l}{1+e^l} \Rightarrow 1-P(Y=1|x) = 1 - \frac{e^l}{1+e^l} = \frac{1}{1+e^l}$$

The denominator should be $1+e^l$... carry that through everywhere...

Recall $Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} \text{Bern}(\theta)$

$$\Rightarrow L(\theta; y) = \prod_{i=1}^n \theta^{y_i} (1-\theta)^{1-y_i} = \theta^{\sum y_i} (1-\theta)^{n - \sum y_i} \propto \left(\frac{\theta}{1-\theta}\right)^{\sum y_i}$$

$$L(\beta_0, \beta_1; y, x) = \prod_{i=1}^n \left(\frac{e^{a_i}}{1+e^{a_i}}\right)^{y_i} \left(\frac{1}{1+e^{a_i}}\right)^{1-y_i} = \prod_{i=1}^n (1+e^{a_i})^{-1} \prod_{i=1}^n e^{a_i y_i} = \prod_{i=1}^n (1+e^{(\beta_0 + \beta_1 x_i)})^{-1} e^{\sum_{i=1}^n y_i (\beta_0 + \beta_1 x_i)}$$