Lecture 16 Math 390.03-02 4/6/16

n draws from bivariate data $\langle X, Y \rangle$

Y: response / outcome / dep var.

X: covariate / regressor / indep. var

too easily confused with dep/indep in prob.

$\langle x_1, y_1 \rangle, \langle x_2, y_2 \rangle, \ldots, \langle x_n, y_n \rangle$

$X \xrightarrow[\text{through}]{\text{effects change}} Y$

through a mechanism f and error ε

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

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

all possible functions of X is too big of a space to work with

and $\varepsilon \neq h(X)$

not a function of X

Forget everything about r.v.'s and prob for a moment...

$$\left.\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 \\ y_2 &= \beta_0 + \beta_1 x_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_n \end{aligned}\right\}$$

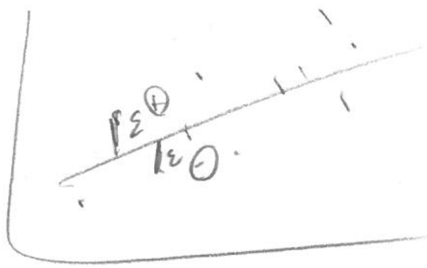
overdetermined system

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

loss function

Laplace's idea... non-unique

$$\hat{\beta}_0,\hat{\beta}_1 = \arg\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \ell(\varepsilon_i)$$

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

$L_2: \ell_2(\varepsilon) = \varepsilon^2$ sqd err / gaussian loss

$L_{0\text{-}1}: \ell_{0\text{-}1}(\varepsilon) = \begin{cases} 0 & \text{if } |\varepsilon| \le c \\ 1 & \text{if } |\varepsilon| > c \end{cases}$

lets go with L_2 loss

error is diff between lin model and actual y

$-\beta_0 - \beta_1 x_i$

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

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

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

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

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

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

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

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

$$\cancel{\frac{\partial}{\partial\beta_0}[f] = -2\bar{y}n - 2\sum y_i x_i + 2n\beta_0 + 2\beta_1\bar{x}n \overset{set}{=} 0 \Rightarrow b_0 = \bar{y}n + \sum y_i x_i - b_1\bar{x}n}$$

$$\cancel{\frac{\partial}{\partial\beta_1}[f] = 2\beta_1\sum x_i^2 - 2\sum y_i x_i + 2\beta_0\bar{x}n \overset{set}{=} 0 \quad b_1 = \sum y_i x_i - b_0\bar{x}n}$$

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

I mean (n-1)sxy = here

$= \sum y_i x_i - n\bar{x}\bar{y} \cancel{+ n\bar{y}\bar{x} + n\bar{y}\bar{x}}$

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

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

$b_1 := \frac{\partial}{\partial \beta_1}(\ ) \overset{set}{=} 0$

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

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

$b_0 := \frac{\partial}{\partial \beta_0}[\ ] \overset{set}{=} 0$

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

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

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

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

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

"L.S. estimates"

Gauss's method of L.S.

line of least fit

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

best est of β_0 " β_1

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

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

$Y = \beta_0 + \beta_1 X + \varepsilon$ where $E[\varepsilon] = 0$

$E(Y) = E_X\left[E_Y[Y|X]\right] = E_X[\beta_0 + \beta_1 X + 0] = \beta_0 + \beta_1 E[X]$

What if $\varepsilon_1 \ldots \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2)$ (Core OLS Assumption)

ord least sqs

Condition on X
so keep things simple (and other reasons - PhD in stats req'd)

1. Normality of errors
2. Indep of errors
3. Homoskedasticity all var $= \sigma^2$ independent of X

$$Y | X \sim N(\beta_0 + \beta_1 X, \sigma^2)$$

Why is normality reasonable? CLT

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

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

when $X_2, \ldots, X_p$ are unobserved and ind.

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

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

$$\alpha + \beta_2 X_2 + \ldots + \beta_p X_p \overset{d}{\approx} N(\beta_0, \sigma^2)$$

let $\varepsilon = \sum_{i=2}^{p} \beta_i X_i - \beta_0$ Noise is the results 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 ?

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

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

$$\ell(\beta_0, \beta_1 | x, y) = \ln(\mathcal{L}) - \frac{1}{2\sigma^2}$$

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

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

$$\ell'(\beta_0 | \beta_1, x, y) = -\frac{1}{2\sigma^2}\left(2n\beta_0 - 2\bar{y}n + 2\beta_1 \bar{x} n\right) = 0$$

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

the MLE's = the L.S. estimates !

L_2 loss i.e. $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ seems to be "natural" once again...

let B_0 be the estimator (r.v) for b_0 $\quad B_0 \sim ?$

let B_1 ,, ,, ,, ,, ,, ,, ,, b_1 $\quad B_1 \sim ?$ normal,)

$E(B_1) = \ldots \beta_1 \quad E(B_0) = \ldots \beta_0$ (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 of the Bern. model

Does it matter if $P(Y=0|x)$? No 0/1 arbitrary ... labels can be flipped. Usually you know what you want to predict

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

model

lin'l models $\in \mathbb{R}$

prob $\in [0,1]$

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

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

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

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

Recall $Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Bern}(\theta)$

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

$\Rightarrow$

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