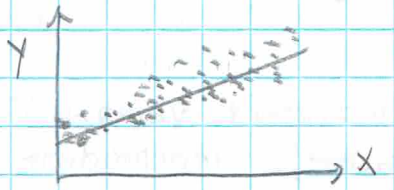
Lecture 16

Suppose we have these ordered pairs $\langle x_1, y_1 \rangle, \langle x_2, y_2 \rangle, \ldots, \langle x_n, y_n \rangle$. X affects Y through some linear function $f \in F_0$ & through some noise. Why do we need the noise? Well our data are points which look like the following:



We see $\nexists$ a solution to the system

$$Y_1 = \beta_0 + \beta_1 X_1$$
$$Y_2 = \beta_0 + \beta_1 X_2$$
$$\cdots$$
$$Y_n = \beta_0 + \beta_1 X_n$$

This is an overdetermined system! To resolve this we need ε_i's.

Our goal is to figure out β_0 & β_1. To do so, we 1st must establish a loss function. Here are a few loss functions

① $L_1 : \ell_1(\varepsilon_i) = |\varepsilon_i|$

② $L_2 : \ell_2(\varepsilon_i) = \varepsilon_i^2$

③ $L_3 : \ell_3(\varepsilon_i) = \begin{cases} 0 & \text{if } |\varepsilon_i| \leq c \\ 1 & \text{if } |\varepsilon_i| > c \end{cases}$

We will use the L_2 loss function aka "quadratic loss."

Given our n 2-dimensional data points our TOTAL ERROR is $\sum_{i=1}^{n} \varepsilon_i^2$. Now $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \Rightarrow \varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$

We want to minimize the error. So we have

$$\hat{\beta}_0, \hat{\beta}_1 := \underset{\beta_0, \beta_1 \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i) \right)^2$$

Before we minimize this, we define the following

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

② $(n-1) S_{xy} := \sum (y_i - \bar{y})(x_i - \bar{x})$

③ $r = \dfrac{S_{xy}}{S_x S_y}$

To minimize, we solve $\nabla \left(\sum \varepsilon_i^2 \right) = 0$ where

$$\nabla := \begin{bmatrix} \frac{\partial}{\partial \beta_0} \\ \frac{\partial}{\partial \beta_1} \end{bmatrix}$$

Solving yields $\hat{\beta}_1 = r\frac{S_x}{S_y}$ & $\hat{\beta}_0 = \bar{y} - r\frac{S_x}{S_y}\bar{x}$

As a matter of notation, we call $\hat{\beta}_0$ & $\hat{\beta}_1$ b_0 & b_1 respectively. The reason for doing so is because the lower case letters are estimators of their counterparts. I.e. b_0 is the estimator of β_0 & b_1 is the estimator of β_1.

b_0 & b_1 are called "the least squares estimates"

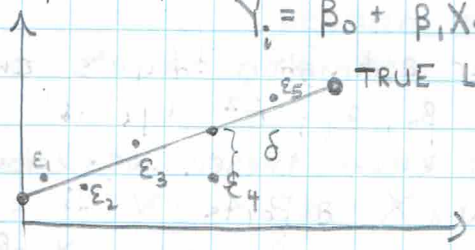
Q: Let X, Y be a bivariate distribution. (These are r.v.'s)
For $Y = f(X) + \varepsilon = \beta_0 + \beta_1 X + \varepsilon$,
does $b_0 = \beta_0$ & does $b_1 = \beta_1$? NO. Why?

$\beta_0 = \bar{Y} - R\frac{S_x}{S_y}\bar{X}$ & $\beta_1 = R\frac{S_x}{S_y}$ but

$b_0 = \bar{y} - r\frac{S_x}{S_y}\bar{x}$ & $b_1 = r\frac{S_x}{S_y}$.

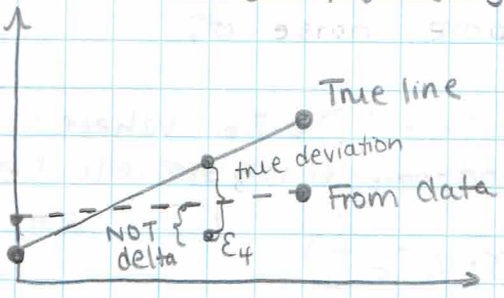
β_0 depends on the r.v. $\bar{Y}$ but b_0 depends on $\bar{y}$ from the samples $\langle x_1, y_1\rangle, \ldots, \langle x_n, y_n\rangle$. Similar reason for b_1. So b_0 depends on the data & often we cannot determine the EXACT distribution of the r.v. from some data so of course $b_0 \neq \beta_0$, $b_1 \neq \beta_1$. Instead b_0 & b_1 are good approximates.

As a quick picture, this is what is going on. In reality, $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ looks like this



$\langle x_1, y_1\rangle, \ldots, \langle x_n, y_n\rangle$ & we get

BUT WE DO NOT SEE THIS. Instead we see the line from the data, i.e. the line $b_0 + b_1 x_i + \varepsilon_i = y_i$ obtained from the data.



This answers HW 8 q2i
The realization of the r.v. ε_i, call it ε_i, does NOT always equal e_i.

Now $Y = f(X) + \varepsilon = \beta_0 + \beta_1 X + \varepsilon$

Let us look at the r.v. ε.

$E(\varepsilon) = 0$ b/c if $E(\varepsilon) \neq 0$, then $\exists$ a systematic deviation from Y to $f(X)$ so the error isn't random like it should be.

Note $E[Y] \neq E[\beta_0 + \beta_1 X + \varepsilon]$ b/c ε is a r.v.

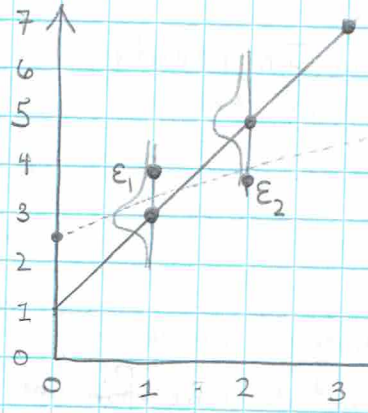
$$E_\varepsilon [E_X [\beta_0 + \beta_1 X + \varepsilon \mid \varepsilon]] = E_\varepsilon [\beta_0 + \beta_1 E[X] + \varepsilon] = \beta_0 + \beta_1 E[X]$$

OLS Assumptions

1) Conditional on X
2) Linearity (f)
3) Normality of errors
4) Independence of errors
5) Homoskedasity (i.e. same variance)

$\Big\}\ \varepsilon_1, \dots, \varepsilon_n \overset{i.i.d}{\sim} N(0, \sigma^2)$

All these 5 conditions can be written as the following $Y_i \mid X_i = x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ for $i \in \{1, 2, \dots, n\}$ all independent.



Suppose $Y = 1 + 2X$. Then you draw from $N(\beta_0 + \beta_1 x_i, \sigma^2)$

Our estimation targets are $\beta_0, \beta_1, \sigma^2$ b/c if we know these, we know how X affects Y.

Remember, X affects Y through some linear function $f(X) = 1 + 2X$ & then some noise σ^2.

Suppose $Y = \alpha + \beta_1 X + \gamma_1 z_1 + \gamma_2 z_2 + \dots + \gamma_p z_p$ where $z_1, \dots, z_p$ are independent, UNOBSERVED r.v.'s

Then

$$E[Y \mid X] = \beta_1 X + \underbrace{\left(\alpha + \sum_{i=1}^{p} \gamma_i E[z_i]\right)}_{:= \beta_0}$$

$$\text{Var}(Y \mid X) = \sum_{i=1}^{p} \gamma_i^2 \text{var}(z_i) := \sigma^2$$

Normality by C.L.T. since we are adding a bunch of independent r.v.'s

Now remember β_0 & β_1 are r.v.'s. (recall that $Y = \beta_0 + \beta_1 X + \varepsilon$). So what are the MLE's of β_0 & β_1? Sticking to our notation the MLE for β_0 is denoted by $\hat{\beta}_0^{MLE}$ & the MLE for β_1 is denoted by $\hat{\beta}_1^{MLE}$. So

$\hat{\beta}_0^{MLE} = ?$

$\hat{\beta}_1^{MLE} = ?$

Recall the MLE is a POINT which maximizes the likelihood function. It turns out $\hat{\beta}_0^{MLE} = b_0$ & $\hat{\beta}_0^{MLE} = b_1$.

Recall how we got b_0 & b_1. b_0 & b_1 were the LEAST SQUARES ESTIMATE. It turns out they are also the MLE's for β_0 & β_1.

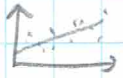
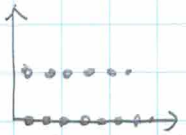
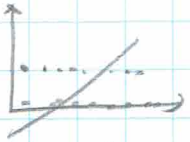
Again, β_0 & β_1 are r.v.'S. We said

$$\beta_0 = \bar{Y} - R\frac{s_x}{s_y}\bar{X}$$

&

$$\beta_1 = R\frac{s_x}{s_y}.$$

A natural question to ask is what is the distribution of B_0 & B_1? We will answer this later...

Let us again consider the following:  & we see $Y = f(X) + \varepsilon = \beta_0 + \beta_1 X + \varepsilon$. We can then impose a linear line  In that case, $\text{support}(Y) \subset \mathbb{R}$. What happens if $\text{supp}(Y) = \{0, 1\}$? Then the graph looks like  so when we fit a line, we get  which is a bad approximation. We need to somehow estimate Y. It might be better to look at $P(Y=1|X)$ Well, $Y = 1|X \sim \text{Bern}(?)$

How do we figure out a model for $Y = 1|X$?

Recall that if $p \in (0,1)$ then $\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) \in (-\infty, \infty)$
so logit transforms a unit interval into the entire real line

Then $\text{logit}\left(p(Y=1|X)\right) = \beta_0 + \beta_1 X$ (no epsilon)

$$\Rightarrow \quad p(Y=1|X) = \text{logit}^{-1}(\beta_0 + \beta_1 X)$$

$$p(Y=1|X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

Fact: if $\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$, then $\text{logit}^{-1}(p) = \frac{e^p}{e^p+1}$

So

$$Y=1|X \sim \text{Bern}\left(\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}\right)$$

Now the likelihood for a Bernoulli(p) is $\mathcal{L}(x;\theta) = \prod_{i=1} p(x;\theta)$

$$= \theta^{x_1}(1-\theta)^{1-x_1} \cdot \ldots \cdot \theta^{x_n}(1-\theta)^{1-x_n}$$
$$= \theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}$$

For us, $Y=1|X \sim \text{Bern}\left(\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}\right)$ so

$$\mathcal{L}(\beta_0, \beta_1; Y, X) = \prod \left(\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}\right)^{y_i} \left(\frac{1}{1+e^{\beta_0+\beta_1 x}}\right)^{1-y_i}$$

$$= \prod_{i=1}^{n} \left(\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}\right)^{\Sigma y_i} \left(\frac{1}{1+e^{\beta_0+\beta_1 x}}\right)^{n-\Sigma y_i}$$