Lecture 30

Simple Linear Regression

Idea: Given two RV's X and Y we assume a linear relationship.

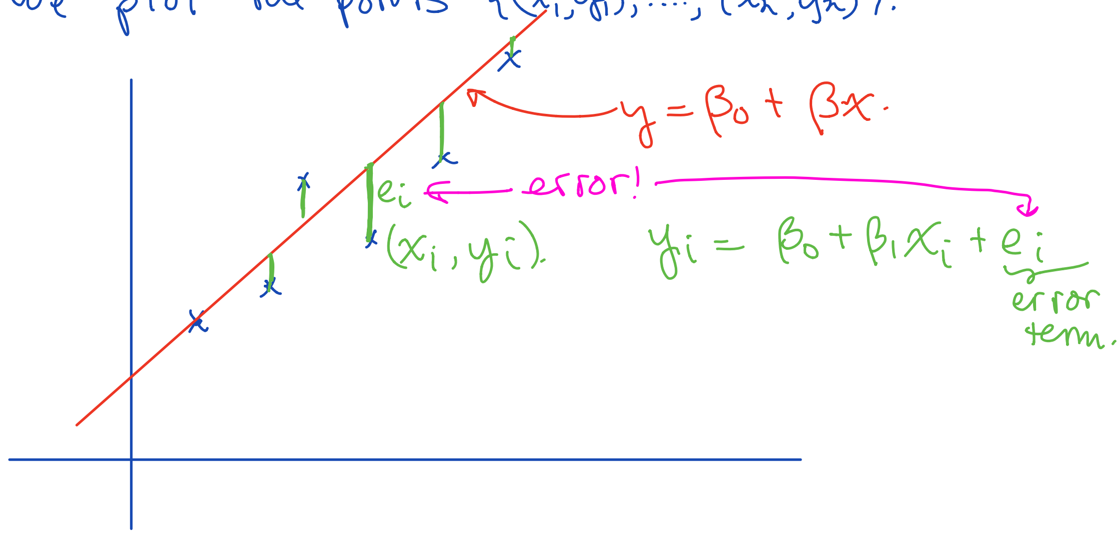
- We assume $Y = \beta_0 + \beta_1 X + \underbrace{\varepsilon}_{\text{error term: } E[\varepsilon] = 0}$.

- β_1 + β_0 are called the "regression coefficients".
- $E[Y] = E[\beta_0 + \beta_1 X + \varepsilon] = \beta_0 + \beta_1 E[X]$.

- suppose we have a set of observations $\{(x_1, y_1), \dots, (x_n, y_n)\}$

- want to determine β_0, β_1, the regression coefficients.

- we plot the points $\{(x_1, y_1), \dots, (x_n, y_n)\}$:



- write $e_i = y_i - \beta_1 x_i - \beta_0$

- $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0)^2$.

- Gauss: "least squares" best fit:

$$F(\beta_0, \beta_1) := \sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0)^2$$

is a function of β_0, β_1. $\leadsto$ choose (β_0, β_1) that minimize this function.

- this is an exercise in calculus: want $\frac{\partial F}{\partial \beta_0} = 0 = \frac{\partial F}{\partial \beta_1}$

- this gives two linear equations in two unknowns, so we can <u>always</u> solve!! :)

- the solutions are called $\hat{\beta}_0$ and $\hat{\beta}_1$, the "least squares estimates".

- <u>General solution</u> (exercise in linear algebra/calc).

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{S_{xy}}{S_{xx}}$$

$$\beta_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where $\bar{x}$, $\bar{y}$ are the sample averages.

- least squares regression lets us predict values of (x, y) not in our data.

Ex: Consider the data:

$$\{(1,2),\ (2,5),\ (3,3),\ (4,8),\ (5,7)\}.$$

- For computational reasons, it is useful to organize the data into a chart as follows:

i	x_i	y_i	$x_i y_i$	x_i^2
1	1	2	2	1
2	2	5	10	4
3	3	3	9	9
4	4	8	32	16
5	5	7	35	25
sum	15	25	88	55
mean	3	5		

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{5} x_i y_i - 5\bar{x}\bar{y}}{\sum_{i=1}^{5} x_i^2 - 5\bar{x}^2} = \frac{88 - 5(3)(5)}{55 - 5(3)^2}$$

$$= \frac{13}{10} = 1.3.$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 5 - (1.3)(3) = 1.1.$$

So in this case, the line of best fit is

$$y = (1.1) + (1.3)x.$$

So if someone asked what the value of y would be if $x=6$, then you could estimate

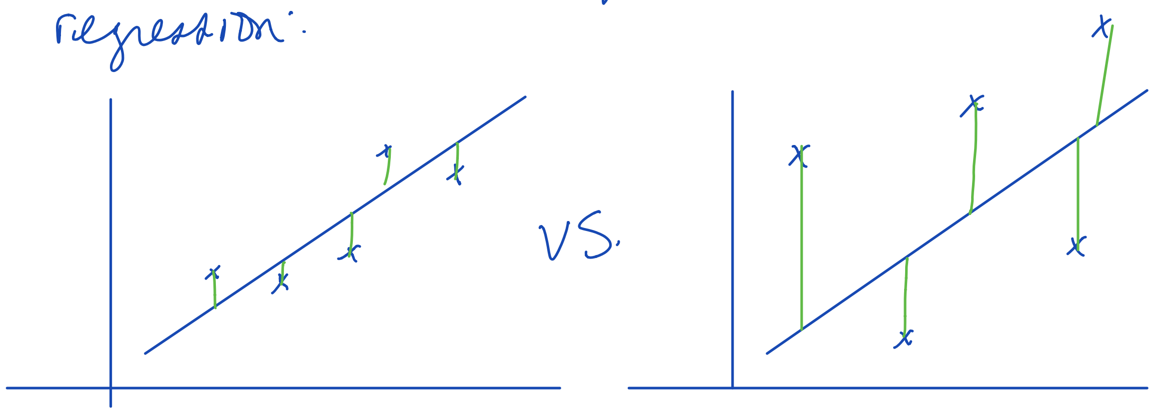
$$y = 1.1 + (1.3)6 = 8.9$$

<u>more on error</u>

- we assume we have a set of points $\{(x_1, y_1), \dots, (x_n, y_n)\}$.

- suppose $y = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is the least squares best fit.
- for each i, we have $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
  - e_i, the error, is then $e_i = y_i - \hat{y}_i$
  - we assume that the e_i's are realizations of a random variable ε with $E(\varepsilon)=0$.

- <u>issue:</u> two different samples can give the same regression:



VS.

small error
small $Var(\varepsilon)$

big error
big $Var(\varepsilon)$.

- So we want to estimate $Var(\varepsilon) = \sigma^2$.
- we use the unbiased estimator

$$\hat{\sigma}^2 = \frac{1}{n-2} \underbrace{SS_E}_{\text{sum of squares error.}}$$

$$SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- This is usually a pain to compute.
- Short cut:

$$SS_E = \underbrace{SS_T}_{\text{total sum of squares}} - \underbrace{SS_R}_{\text{Regression sum of squares.}}$$

$$SS_T = S_{yy} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

and

$$SS_R = \hat{\beta}_1 S_{xy} = \hat{\beta}_1 \left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \right).$$

Ex:

i	x_i	y_i	$x_i y_i$	x_i^2
1	1	2	2	1
2	2	5	10	4
3	3	3	9	9
4	4	8	32	16
5	5	7	35	25
sum	15	25	88	55
mean	3	5		

$$SS_T = (2^2 + 5^2 + 3^2 + 8^2 + 7^2) - 5(5)^2$$

$$= 151 - 5^3 = 26.$$

$$SS_R = \hat{\beta}_1 S_{xy} = (1.3)(88 - 5\cdot(3)(5))$$

$$= 16.9$$

so $SS_E = 26 - 16.9 = \boxed{9.1}$