STA302 - Lecture 1

* The statistical model for simple linear regression is:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Y = dependent variable, it's a r.v

X = independent variable

β_0, β_1 = parameters of model, they are unknown and we will use observed data to estimate them.

ε = random error/noise, variation in measures that we can't account for.

* linear in β's

Simple - only one X; i.e only one predictor.

* More formally

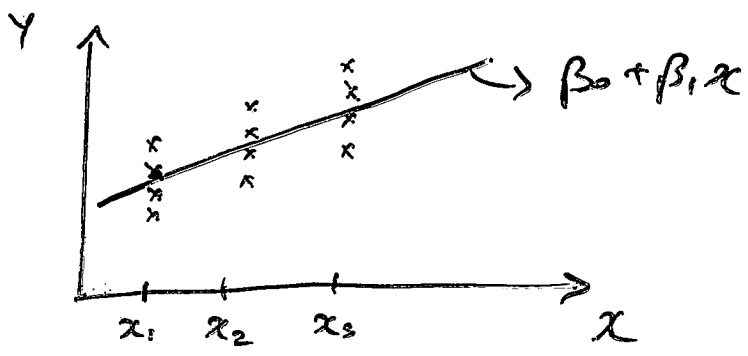
$$E(Y|X) = \beta_0 + \beta_1 X \qquad \text{if } X \text{ is r.v}$$

$$E(Y|X=x) = \beta_0 + \beta_1 x \qquad \text{if } X \text{ is fixed.}$$

* Can think about the different values of X as defining different subpopulations, one for each possible value of X. Each subpopulation consist of all individuals in the population having the same x value.

* Example:

Y = weight , X = height



(x_i, y_i)

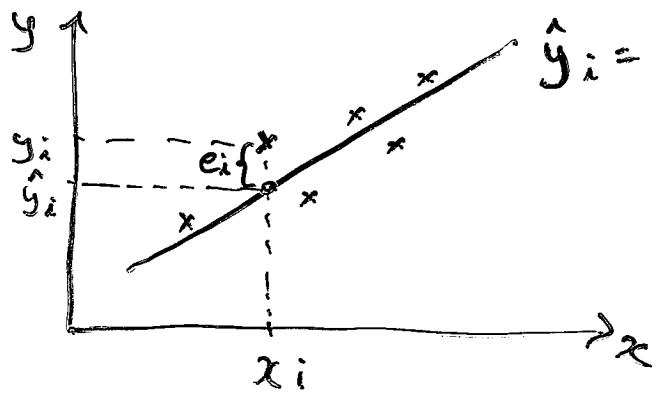
$i = 1, \ldots, 30$

SLR Model - Continuation

* Data (x_i, y_i) $i = 1, \dots, n$

* Model: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ $i = 1, \dots, n$

* Graphical representation of data is a simple scatterplot.



* Want to estimant β_0 and β_1 based on observed data. Estimated are denoted by b_0, b_1 or $\hat{\beta}_0$ and $\hat{\beta}_1$

* For each observed value x_i of X the fitted value of Y is $\hat{y}_i = b_0 + b_1 x_i$

* The deviations from the line in vertical direction are the errors in prediction of Y and are called "residuals". They are defined as:

$$e_i = y_i - \hat{y}_i$$

- e_i: residuals
- y_i: observed value of y for particular x_i value
- $\hat{y}_i$: predicted value of Y for that same x_i

* Note, in the SLR model ε_i's are random variables. The observed residuals are estimates of ε_i, i.e, $\hat{\varepsilon}_i = e_i$

* The estimated b_0 and b_1 are found by the Method of Least Square which is based on Minimizing the sum of squares of the residuals

Least-Square Estimates

* Why vertical distances?

- Want to predict Y from X and so we want $\hat{y}_i$ (the predicted value) to be as close as possible to y_i (observed value)
- If we minize the horisontal distances, we will get a different answer for b_0 and b_1.
- Regression is <u>not</u> symmetric - it matters which variable is dependent and which is independent.

* Why minimizing <u>squared</u> deviations?

Answer: Since $\sum_{i=1}^{n} e_i = 0$ (proof later...)

Derivation of LS estimates

* Residuals, vertical distances:

$$e_i = y_i - \hat{y}_i = y_i - \underbrace{(b_0 + b_1 x_i)}_{\hat{y}_i = \text{the fitted line.}}$$

* $$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

want to find b_0 and b_1 that minimized RSS.

* So, set $\frac{\partial RSS}{\partial b_0} = 0$ and $\frac{\partial RSS}{\partial b_1} = 0$

$$\Rightarrow \frac{\partial RSS}{\partial b_0} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)$$

$$\frac{\partial RSS}{\partial b_1} = -2 \sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i)$$

* So b_0 and b_1 must satisfy:

$$\text{①} \quad \sum_{i=1}^{n} y_i = n b_0 + b_1 \Sigma x_i$$

$$\text{②} \quad \sum_{i=1}^{n} x_i y_i = b_0 \Sigma x_i + b_1 \Sigma x_i^2$$

] "Normal Equations"

Review of Summation Rules

a = constant

* $\sum_{i=1}^{n} x_i = x_1 + x_2 + \dots + x_n$

* $\sum_{i=1}^{n} a x_i = a x_1 + a x_2 + \dots + a x_n = a \cdot \sum x_i$

---

* $\sum_{i=1}^{n} a = a + a + \dots + a = n \cdot a$

* $\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} (x_i - y_i)$

Derivation of LS Estimates - Cont's

* "Normal Equations"

① $\sum_{i=1}^{n} y_i = n b_0 + b_1 \Sigma x_i$

② $\sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2$

* Let $\bar{y} = \frac{1}{n}\Sigma y_i \Rightarrow \sum_{i=1}^{n} y_i = n\bar{y}$

$\bar{x} = \frac{1}{n}\Sigma x_i \Rightarrow \sum_{i=1}^{n} x_i = n\bar{x}$

* Using this we have:

① $b_0 = \bar{y} - b_1 \bar{x}$

② $\sum_{i=1}^{n} x_i y_i = b_0 \cdot n\bar{x} + b_1 \sum_{i=1}^{n} x_i^2$

$\Rightarrow \sum_{i=1}^{n} x_i y_i = (\bar{y} - b_1 \bar{x}) \cdot n\bar{x} + b_1 \sum_{i=1}^{n} x_i^2$

$$b_1 = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\Sigma x_i^2 - n\bar{x}^2} = \frac{S_{xy}}{S_{xx}}$$

S_{xy} = sample covariance

$S_{xx} = (n-1)S_x^2$ it's $(n-1)\times$ sample variance of x_i's

* Can show that:

$$b_1 = \frac{\Sigma(x_i - \bar{x})\cdot(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \quad (*)$$

* Exercises:

(1) Show that (*) is the same as the above results from normal equations.

(2) Show that b_0 and b_1 give minimum.

Properties of fitted line:

(1) Residuals: $e_i = y_i - \overbrace{(b_0 + b_1 x_i)}^{\hat{y}_i}$

$$= y_i - (\bar{y} - b_1\bar{x} + b_1 x_i)$$

$$= y_i - \bar{y} - b_1(x_i - \bar{x})$$

$$\Rightarrow \sum_{i=1}^{n} e_i = \sum_{i=1}^{n}\left[y_i - \bar{y} - b_1(x_i - \bar{x})\right]$$

$$= \sum_{i=1}^{n}(y_i - \bar{y}) - b_1\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

$\Sigma(y_i - \bar{y}) = 0$
$\Sigma(x_i - \bar{x}) = 0$ why?

Properties of fitted line Cont's

(1) $\sum e_i = 0$

(2) $\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} (b_0 + b_1 x_i)$

$= \sum_{i=1}^{n} (\bar{y} - b_1 \bar{x} + b_1 x_i)$

$= n\bar{y} - b_1 n\bar{x} + b_1 \underbrace{\sum x_i}_{n\bar{x}}$

$= n\bar{y} - b_1 \cancel{n}\bar{x} + b_1 \cancel{n}\bar{x}$

$= n\bar{y} = \sum_{i=1}^{n} y_i$

(3) $\sum_{i=1}^{n} e_i x_i = 0$

(4) $\sum_{i=1}^{n} e_i \cdot \hat{y}_i = 0$

] Exercises!

Statistical Assumptions for SLR

(1) We assume that the linear model is appropriate, i.e, that indeed

$$E(Y|X) = \beta_0 + \beta_1 X$$

* $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i = 1, \ldots, n$

Y_i: r.v; β_0, β_1: parameters; X_i: fixed; ε_i: random

* So the additional assumptions are:

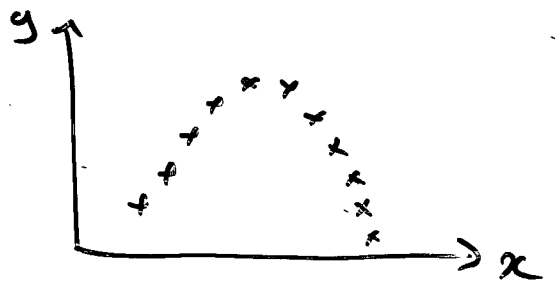
(2) $E(\varepsilon_i) = 0 \Rightarrow E(Y_i | X_i = x_i) = \beta_0 + \beta_1 x_i$

(3) $Var(\varepsilon_i) = \sigma^2 \Rightarrow Var(Y_i | X_i = x_i) = \sigma^2$

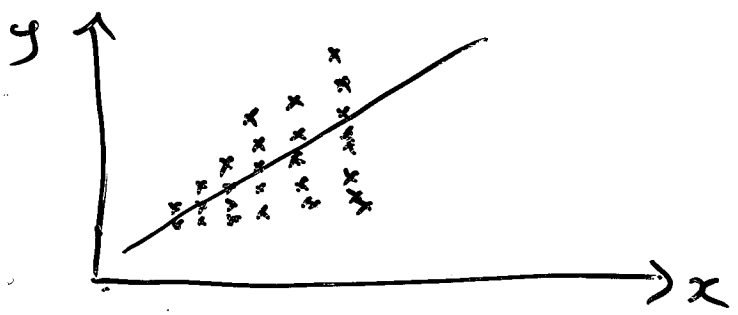
(4) ε_i's are uncorrelated $\Rightarrow$ Y_i's are uncorrelated

Possible Violation of Statistical Assumptions

* linear model is not appropriate, for example



* Variance of error, $[var(Y_i)]$ is increasing with x_i



* Can have influencial points that moves the line.