WEEK 04

INSTRUCTOR: YANAN WU

TA: KHADIJA NISAR

SPRING 2025

2.1.1 REGRESSION

CAUSALITY & REGRESSION

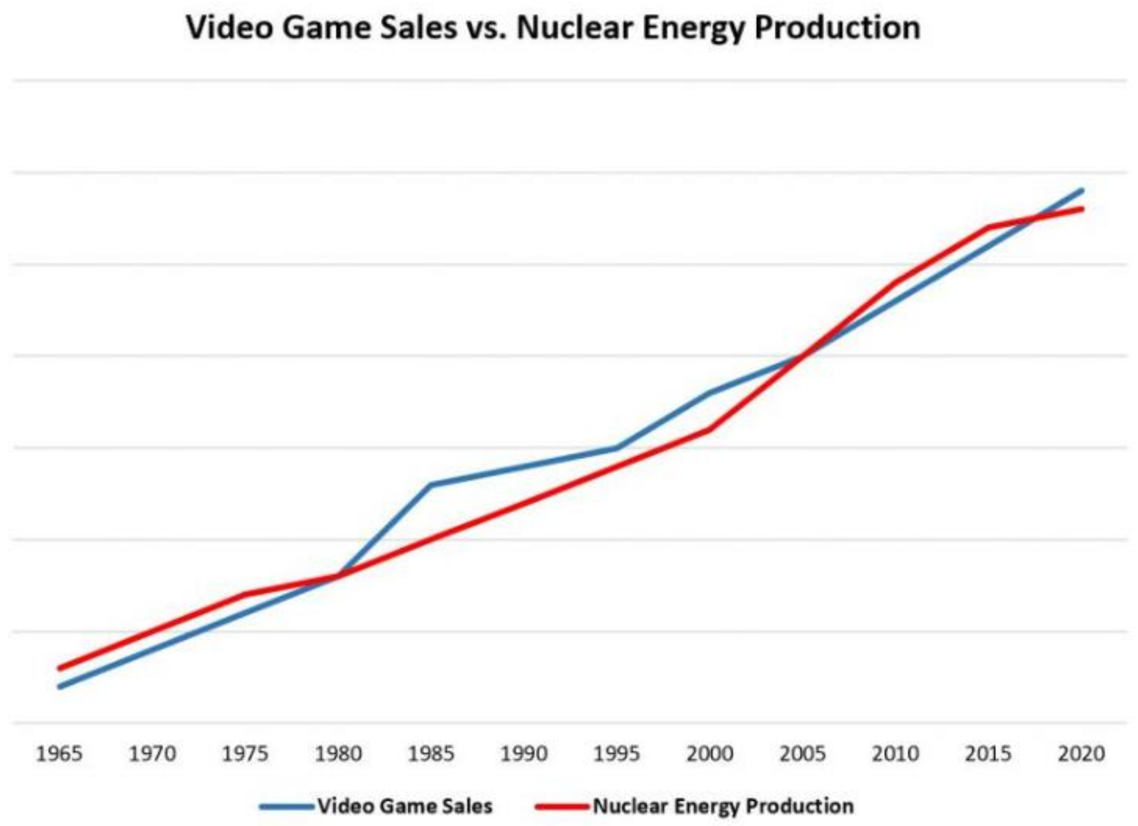
- Causality: Relationship between cause and effect, where one event (the cause) directly influences another event (the effect).

  Example: the relationship between rainfall and flooding

- Co-variation: Two variables change together. If two variables tend to increase or decrease in a related manner, they are said to **covary (not causality)**

SPURIOUS RELATIONSHIP

The covariation between X and Y can be influenced by their joint relationship to another variable Z (or a set of variables)



Examples of spurious relationship

2.1.2 BIVARIATE REGRESSION

MODEL BASED ON POPULATION AND SAMPLE

- For the i-th observation, the **population model** is $y_i = \beta_0 + \beta_1 * x_{i1} + \varepsilon_i$
  1) The β_0, β_1 are constant across all observations
  2) ε_i (the error also called disturbance) is directly associated to the i-th observation

Can we directly observe the population parameter $\beta_0, \beta_1, \varepsilon_i$ from sample? ⍰

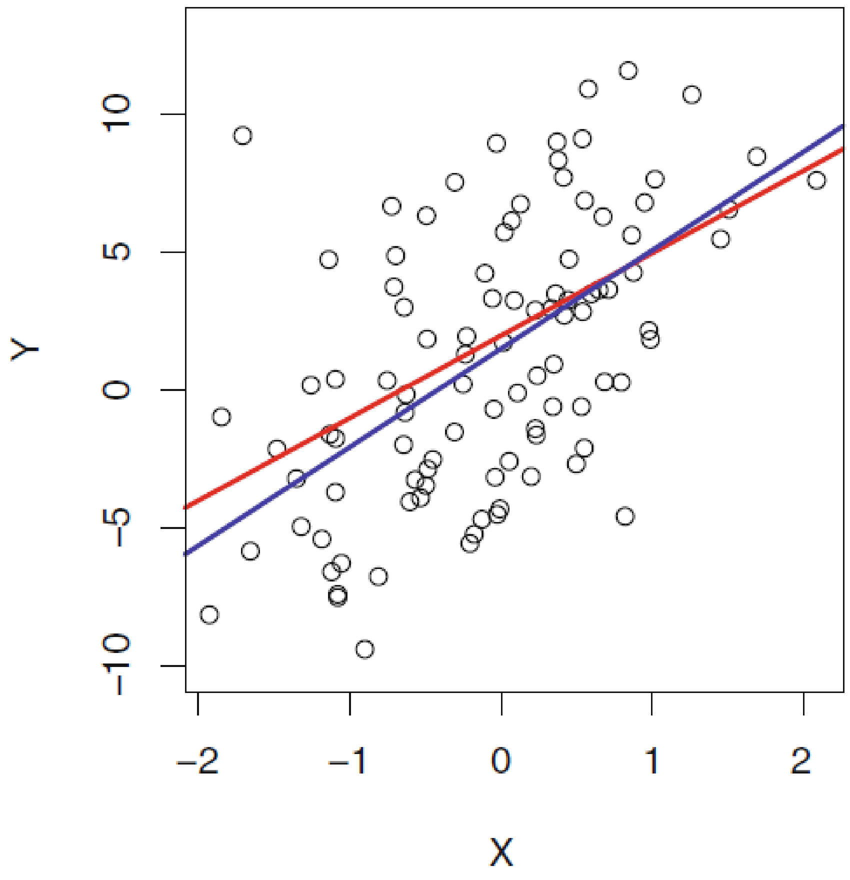
- For the i-th observation, the **estimated model based on sample** is:

$\hat{y}_i = b_0 + b_1 * x_{i1} + e_i$ with the residual $e_i = y_i - \hat{y}_i$

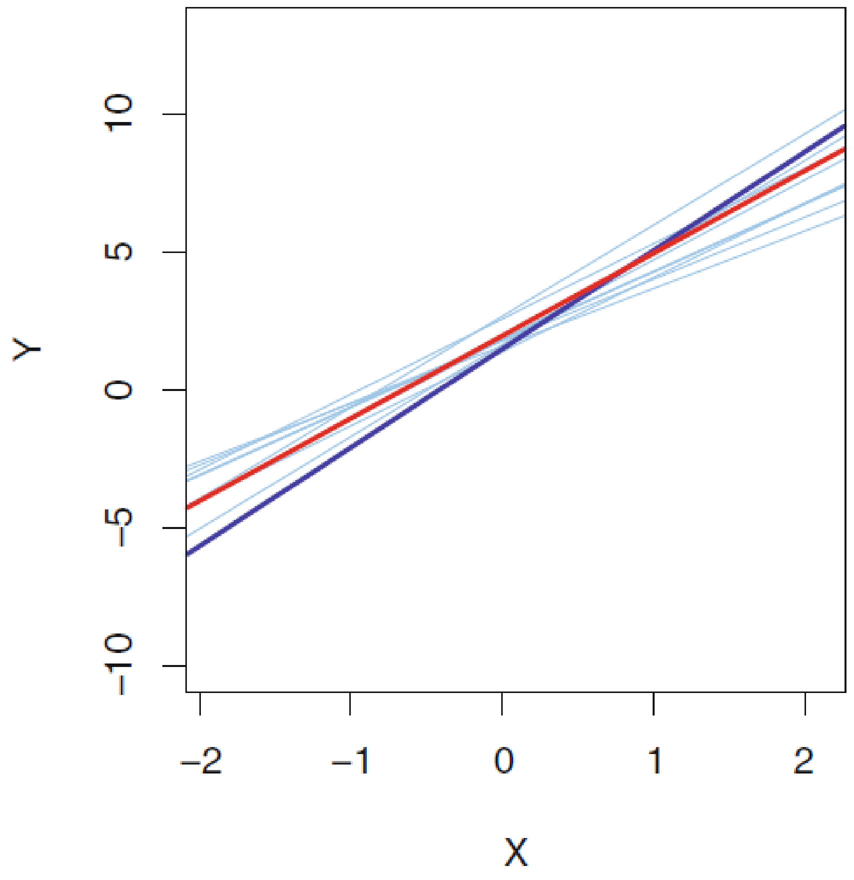
If you were analyzing a dataset on housing prices, what could x_{i1} and $\hat{y}_i$ represent in a regression model? ⍰

POPULATION REGRESSION LINE VS SAMPLE REGRESSION LINE

Red line: population regression line
Dark blue: sample regression line



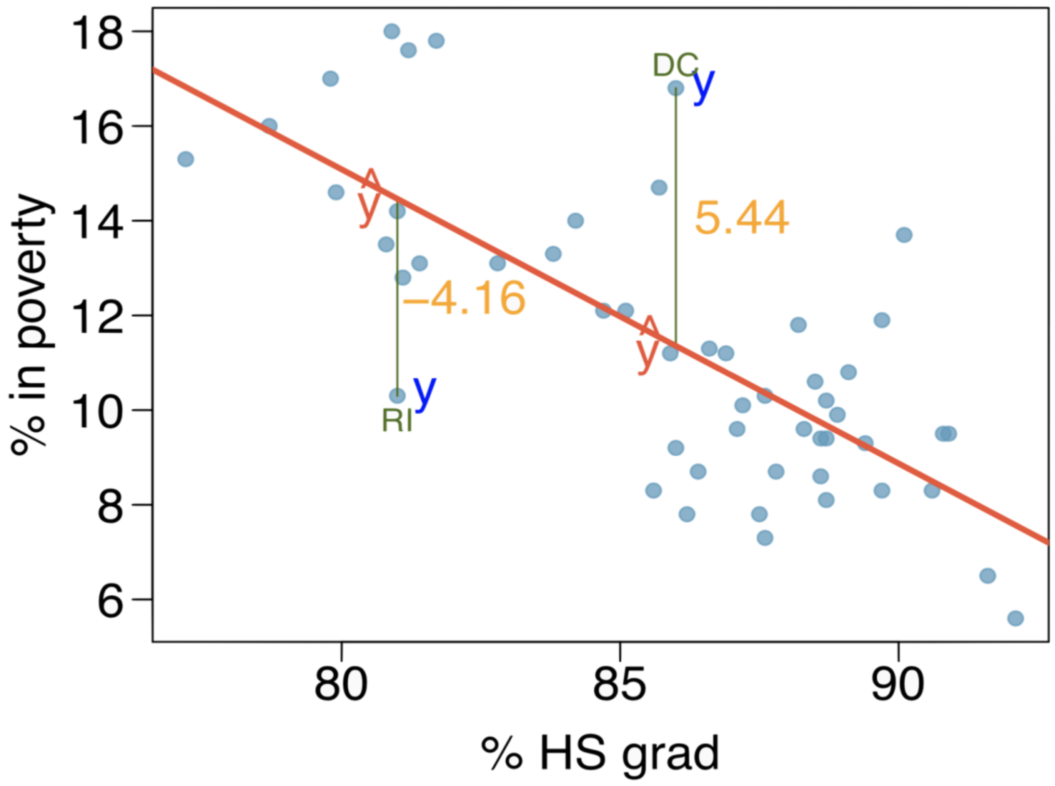
Light blue: sample regression line based on different samples



2.1.3 ORDINARY LEAST SQUARES ESTIMATION

Ordinary Least Squares

1. A straight line can minimize the error (the difference between y_i and $\widehat{y_i}$)

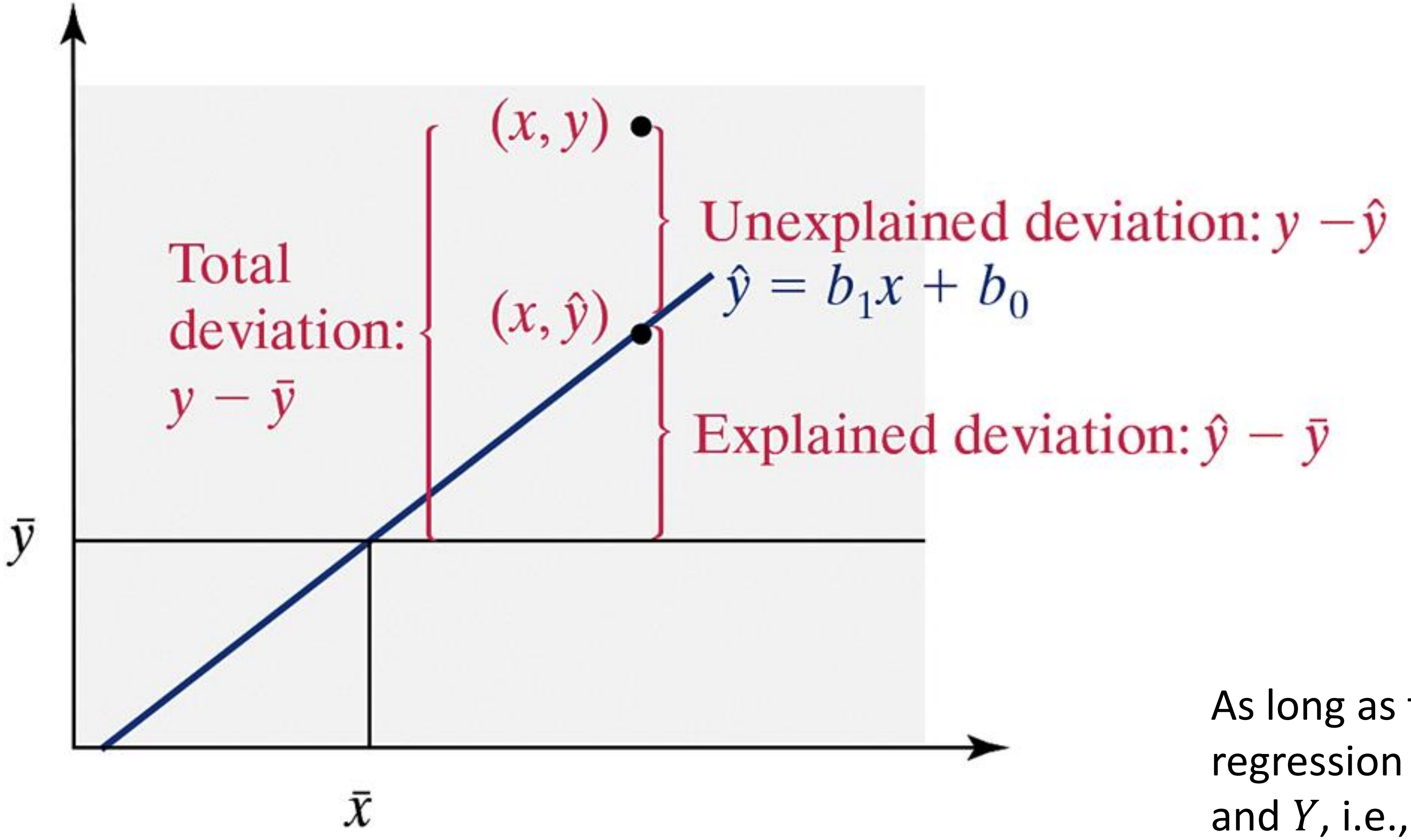


$$e_i = y_i - \widehat{y_i}$$

Do you want a smaller error or larger error? ?

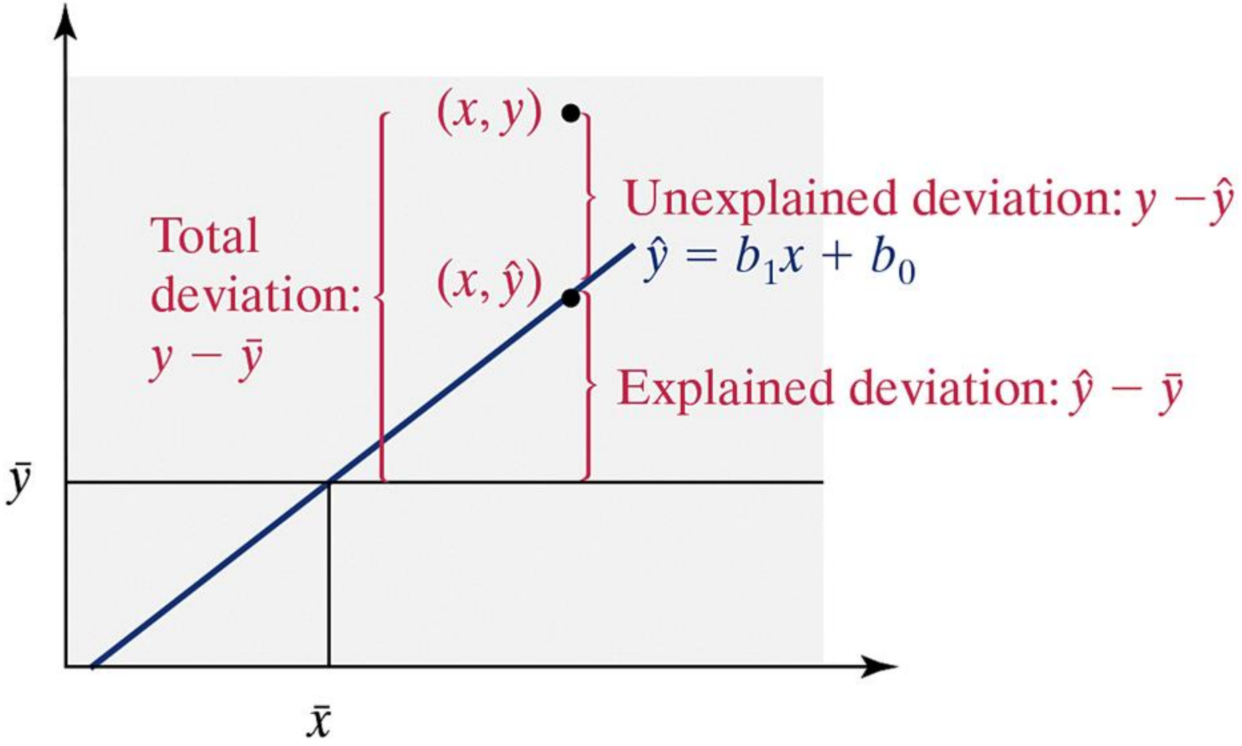
$$min \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

2. Variance Decomposition



As long as the linear model has an intercept, the regression line always goes through means of X and Y, i.e., the point $(\bar{\ },\bar{\ })$ will be on the regression line

2. TSS, RSS, ESS



$$TSS = \sum_{i=2}^{n} (y_i - \bar{y})^2$$

$$ESS = \sum_{i=2}^{n} (\hat{y}_i - \bar{y})^2$$

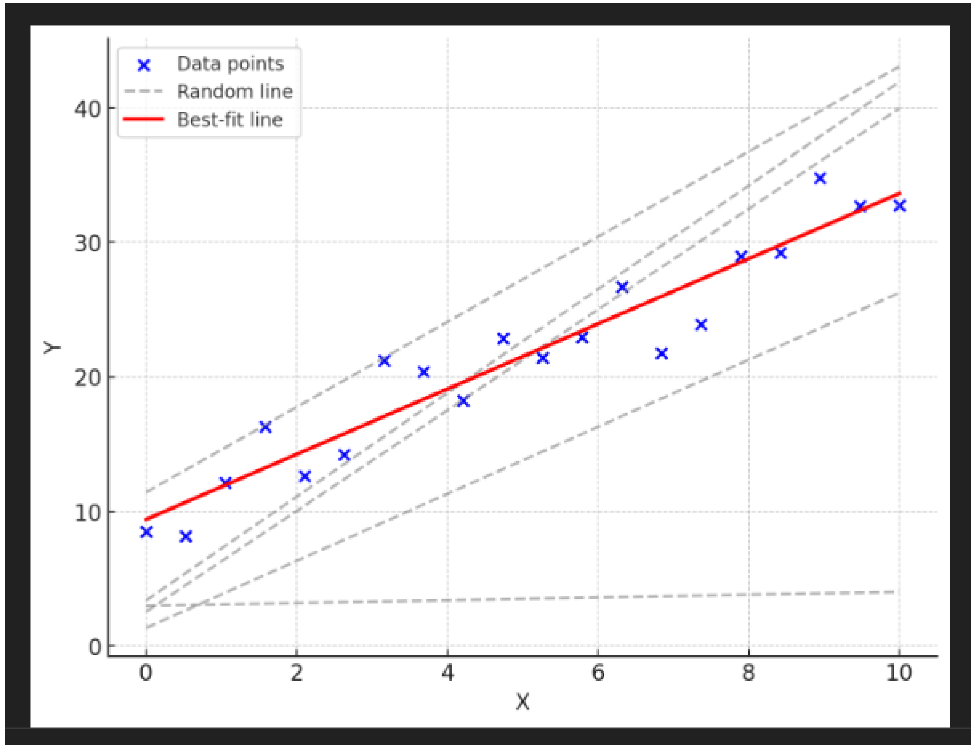
$$RSS = \sum_{i=2}^{n} (y_i - \hat{y}_i)^2$$

TSS = ESS + RSS

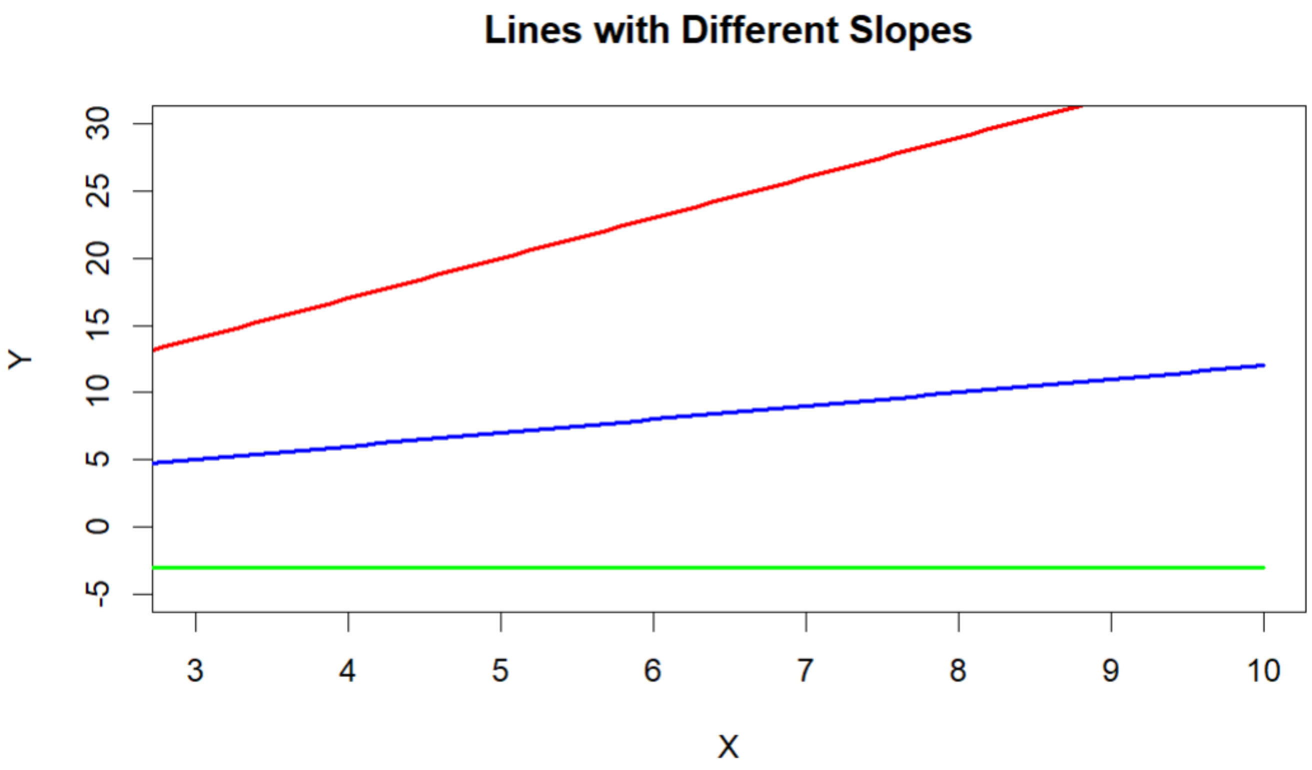
SLOPE AND INTERCEPT

The lest square approach use b_0 and b_1 to minimize the RSS

$RSS = e_1^2 + e_2^2 + \ ... + e_n^2$, where $e_i = y_i \ - \widehat{y_i} = \ y_i \ - (\boldsymbol{b_0} + \boldsymbol{b_1} * x_{i1})$



EXPLANATION ON b_0 AND b_1

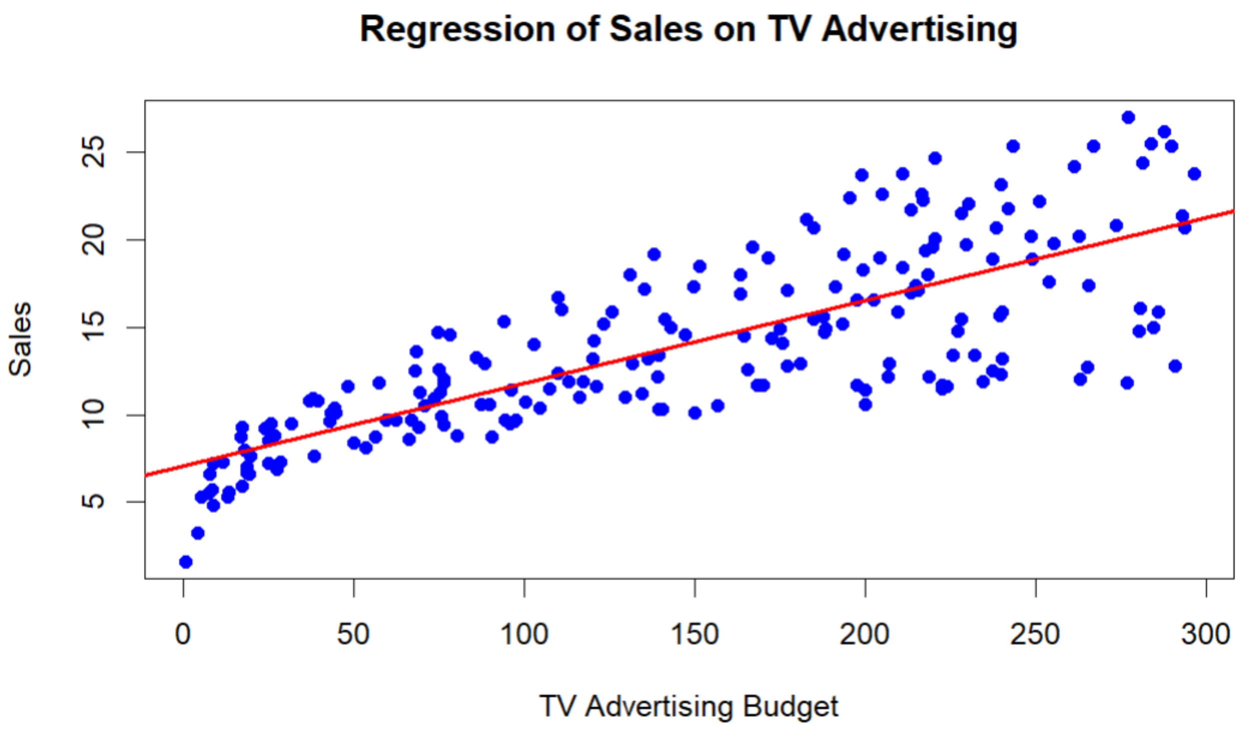


Which line has the highest slope?

Which line has the lowest intercept?

EXPLANATION ON b_0 & b_1

$$b_0 = 7.03, b_1 = 0.0475$$



x: The advertising budget on TV (unit: $)
y: The sales of the TA

If no money is spent on advertising ($x = 0$), what does the model predict for TV sales?

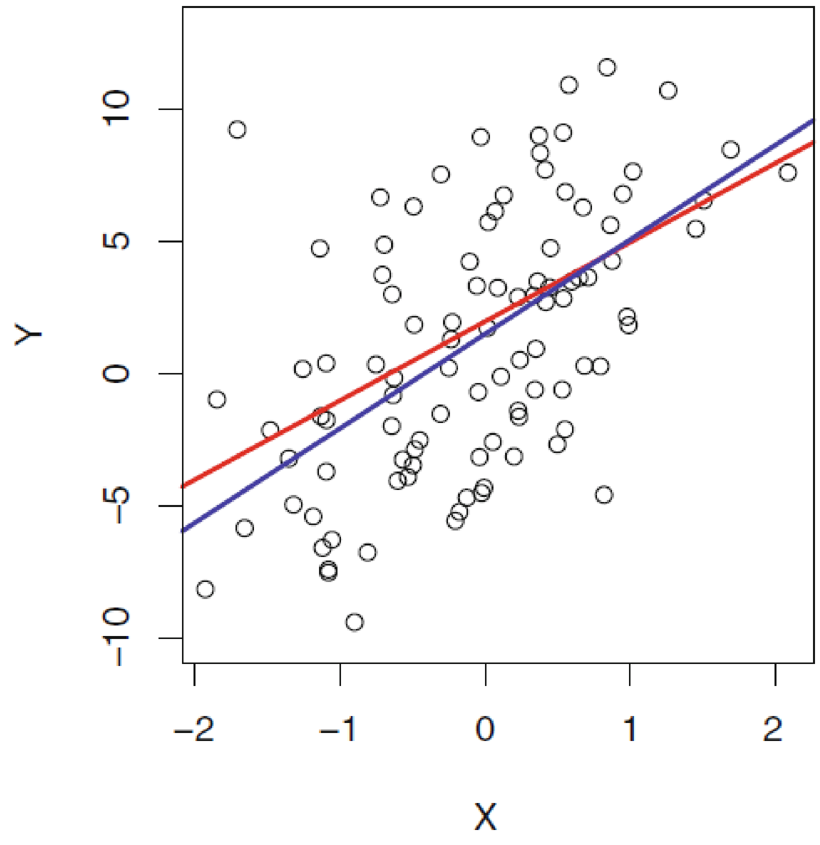
If the advertising budget increases by 1 dollar, how much does the model predict sales will increase?

If the advertising budget increases by $100, how much would we expect sales to increase?

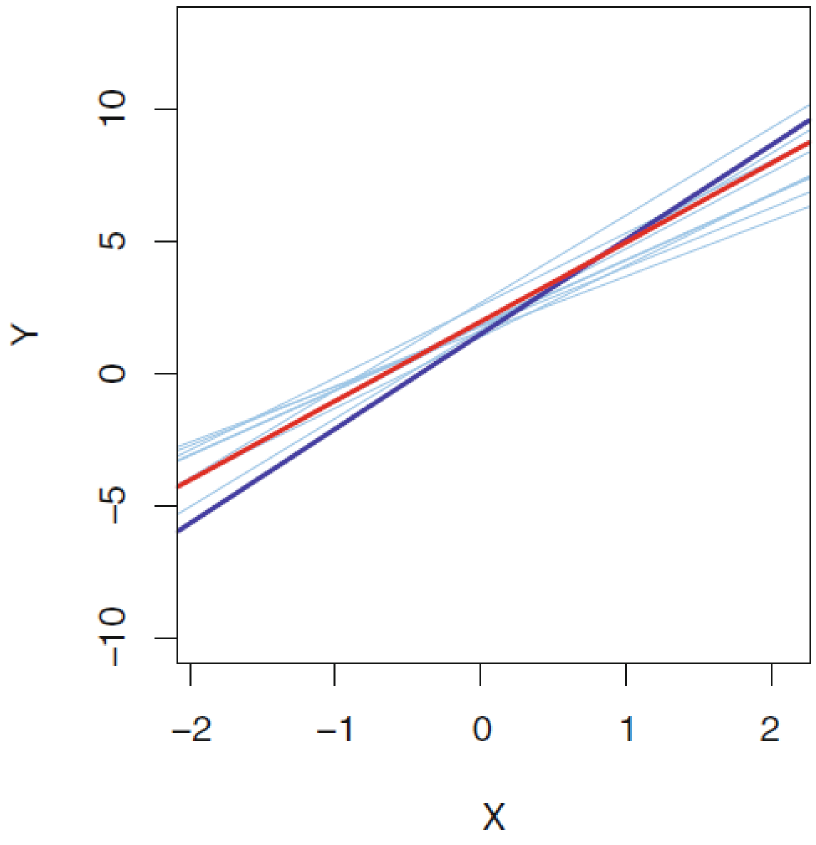
2.1.4 STANDARD ERROR OF b_0 & b_1

The estimated coefficients (b_0 and b_1) differ from sample to sample

Red line: population regression line
Dark blue: sample regression line



Light blue: sample regression line based on different samples



How close the b_0 and b_1 are to the true values β_0 and β_1 ?

STANDARD ERROR

- Standard error measure the uncertainty of the estimated parameter, b_0 and b_1

$$SE_{b_1} = \sqrt{Var(b_1)} = \frac{s_e}{\sqrt{TSS_X}} \quad \text{and } SE_{b_0} = \sqrt{Var(b_0)} = s_e * \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{TSS_X}},$$

With $s_e = \sqrt{\frac{RSS}{n-K}}$ and $TSS_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$

If the standard error of b_1 is large, what does it indicate about the reliability of our estimate?

What is the effect of a small residual sum of squares (RSS) on the standard error?

How does the total of sum of squares TSS_x affect the standard error?

Low uncertainty of any estimates is desirable properties.

2.1.5 CONFIDENCE INTERVAL

CONFIDENCE INTERVAL FOR b_0

A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter.

$$[\, b_0 + 2 * SE(b_0), \mathrm{b}_0 - 2 * SE(b_0)]$$

CONFIDENCE INTERVAL FOR b_1

A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter.

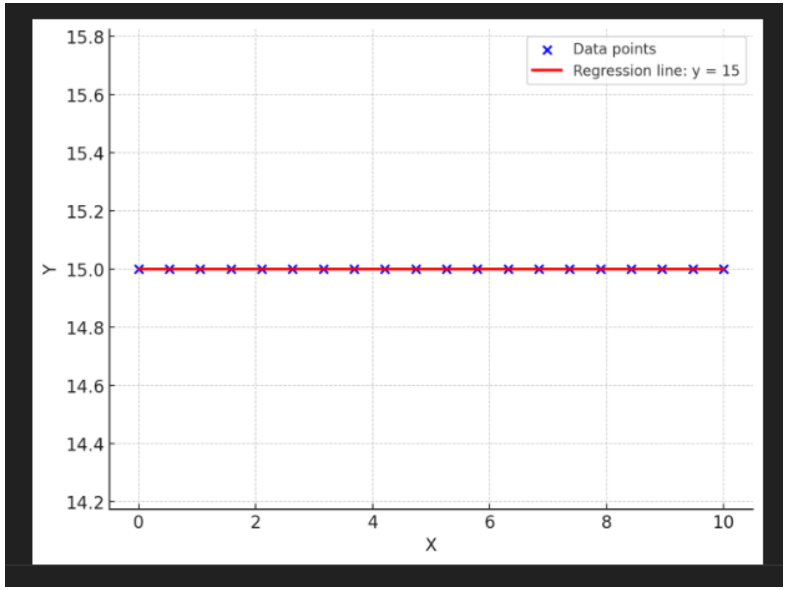
$$[b_1 + \ 2 * SE(b_1), b_1 - 2 * SE(b_1)]$$

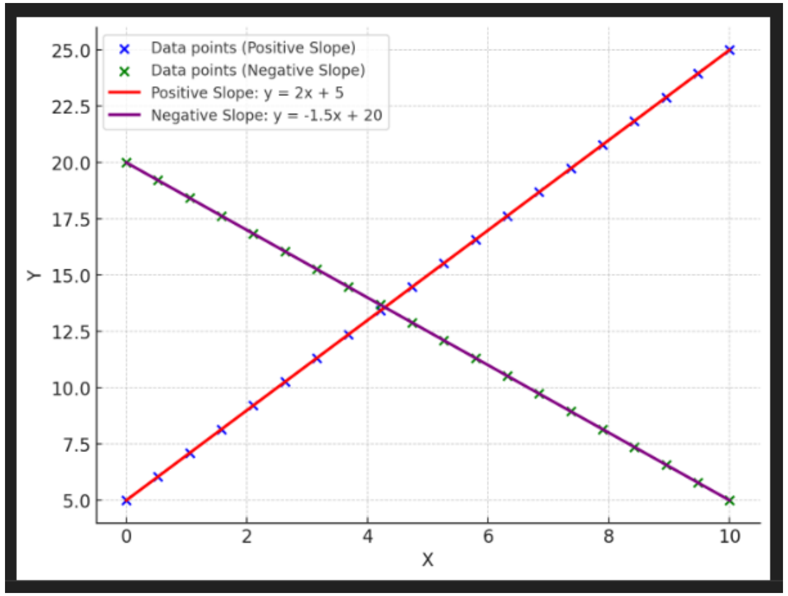
2.1.6 *HYPOTHESIS TEST*

HYPOTHESIS TEST FOR b_1

If the X variable does not explain any variation in Y, then there is no relationship between X and Y

$$H_0: b_1 = 0 \text{ (There is no relationship between } X \text{ and } Y\text{)}$$
$$H_1: b_1 \neq 0 \text{ (There is some relationship between } X \text{ and } Y\text{)}$$





t-statistics

$$t = \frac{b_1 - 0}{SE_{b_1}}$$

p-value: A small p-value indicates that it is unlikely to observe such as substantial association between the X and Y.

Reject the null hypothesis: if the p-value is small enough. Typical p-value cutoffs for rejecting the null hypothesis are 5 or 1%.

2.1.7 R^2 AND $R^2_{adjusted}$

RESIDUAL STANDARD ERROR

$$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

The RSE is considered a measure of the *lack of fit* of the model

Roughly speaking, it is the average amount that the $\hat{y}_i$ will deviate from the y_i.

For data with the same scale, Does a smaller RSE indicate a better or worse model fit?

Assessing the Accuracy of the Model

The goodness of fit

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

The adjusted goodness of fit

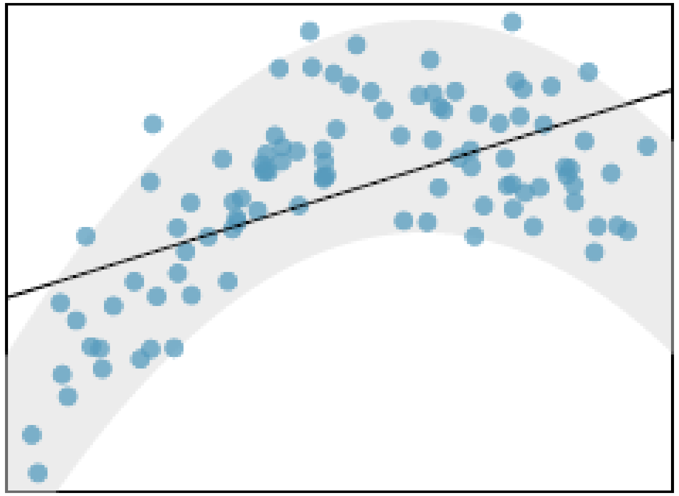
More variables are considered into the regression equation, the better the fit of the model will be

$$R^2_{adj} = 1 - \frac{RSS/(n-K)}{TSS/(n-1)}$$

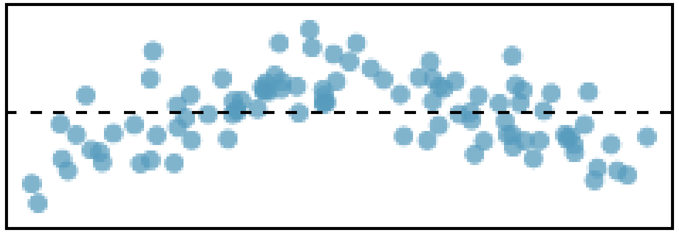
2.1.8 KEY ASSUMPTION ON REGRESSION ANALYSIS

1. Linearity: The relationship between the independent variable and dependent variable is **linear**, if there is a nonlinear trend, an advanced regression method should be applied

Non-linearity of the Data



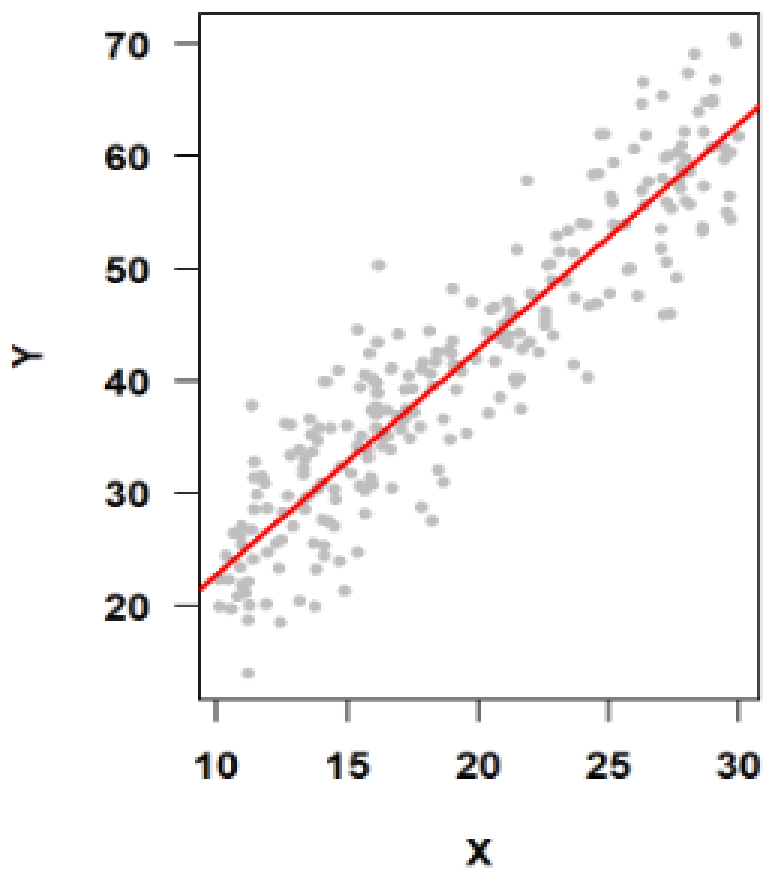
Liner regression line



Residual plot is a useful graphical tool for identifying non-linearity.

2. The error at any level of x_i share an **identical distribution**, with $mean = 0$ and constant variance

constant variance of Error Terms



non-constant variance of Error Terms

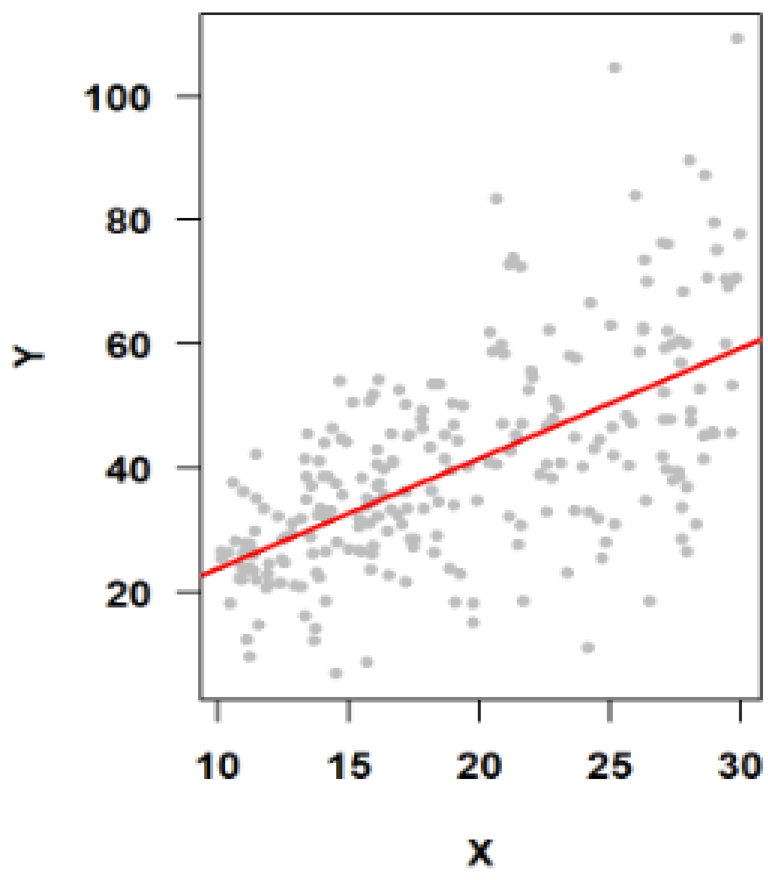
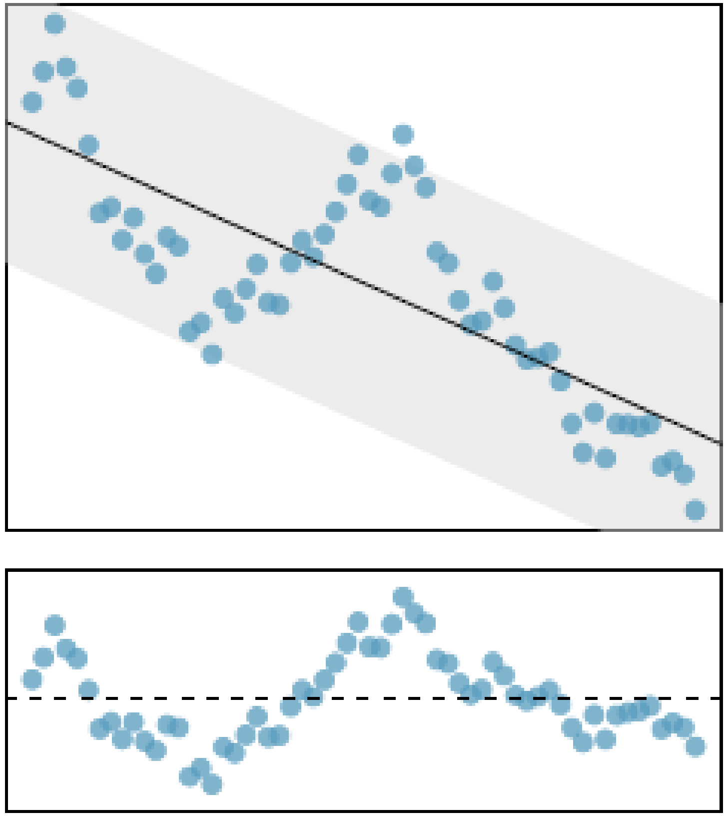


IMAGE SOURCE: HTTPS://WWW.BOOKDOWN.ORG/RWNAHHAS/RMPH/MLR-CONSTANT-VARIANCE.HTML

3. Error are assumed to be **independent** (uncorrelated) among each other

Example of correlated Error



4. i.i.d Normality of Error

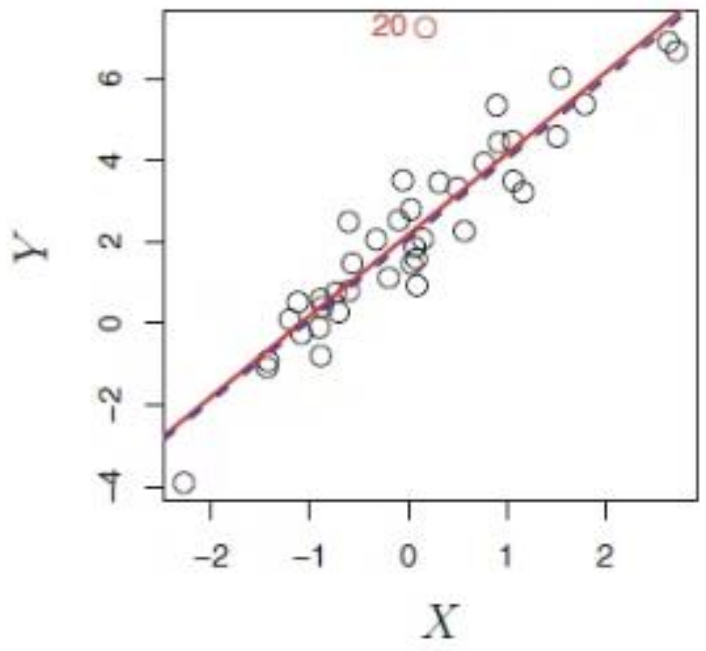
This assumption states that the **disturbances (errors) in a regression model are**:

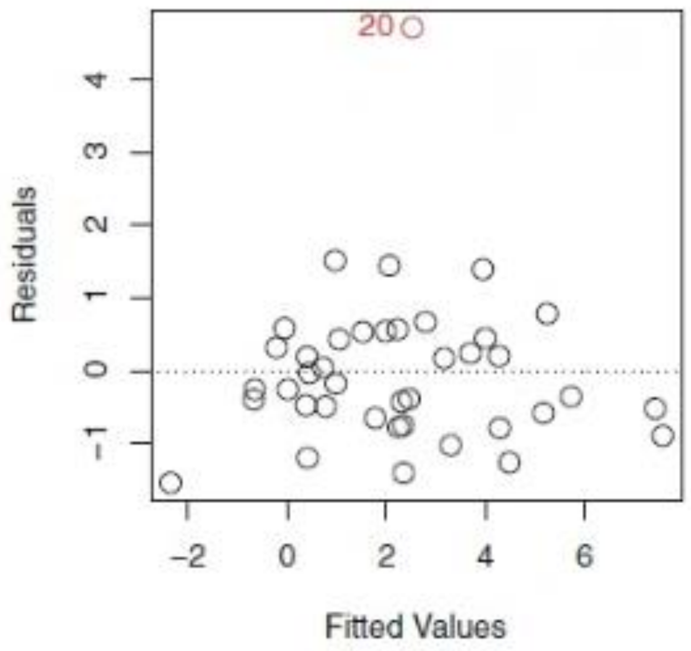
1) Independently and identically distributed (i.i.d)

2) Normally distributed (i.e.,$\varepsilon_i \sim N(0, \sigma^2)$)

This assumption is **important** because it allows for **valid hypothesis testing and confidence intervals**, even when the sample size is **very small**.

4. OUTLIER





If we drop outlier, 20, the RSE decrease from 1.09 to 0.77.
R^2 increase from 0.805 to 0.892.

If we believe the outlier is due to an error in data collection, we can simply remove the observation.

However, care should be taken ,since an outlier may indicate a deficiency in the model, such as missing x variables

WEEK 03

CODE DEMO SESSION

Instructor: Yanan Wu
TA: Khadija Nisar

Spring 2025

CODE

- week04_Demo