STAT 408
Applied Regression Analysis

Miles Xi

Department of Mathematics and Statistics

Loyola University Chicago

Fall 2022

Simple Linear Regression

Motivation

- Often in data analysis, we want to find a relationship between two variables
  - In NCbirth dataset, what is the relationship between baby's birth weight and mother's smoking habit?
  - In brain cancer study, what is the relationship between patients' survival time and tumor volume?

- So far, we have covariance and correlation coefficient to describe the relationship between two variables

- We can also visualize this relationship

Motivation

- Pima diabetes dataset
  - A study on 768 adult female Pima Indians living near Phoenix
  - 9 variables were recorded
    - Pregnant: number of times pregnant
    - Glucose: plasma glucose concentration
    - Diastolic: diastolic blood pressure
    - Triceps: triceps skin fold thickness
    - Insulin: 2-hour serum insulin
    - Bmi: body mass index (weight/height)
    - Diabetes: diabetes pedigree function (diabetes likelihood due to family history)
    - Age
    - Test: if the patient showed signs of diabetes

Motivation

- What is the relationship between insulin and glucose (both quantitative)?

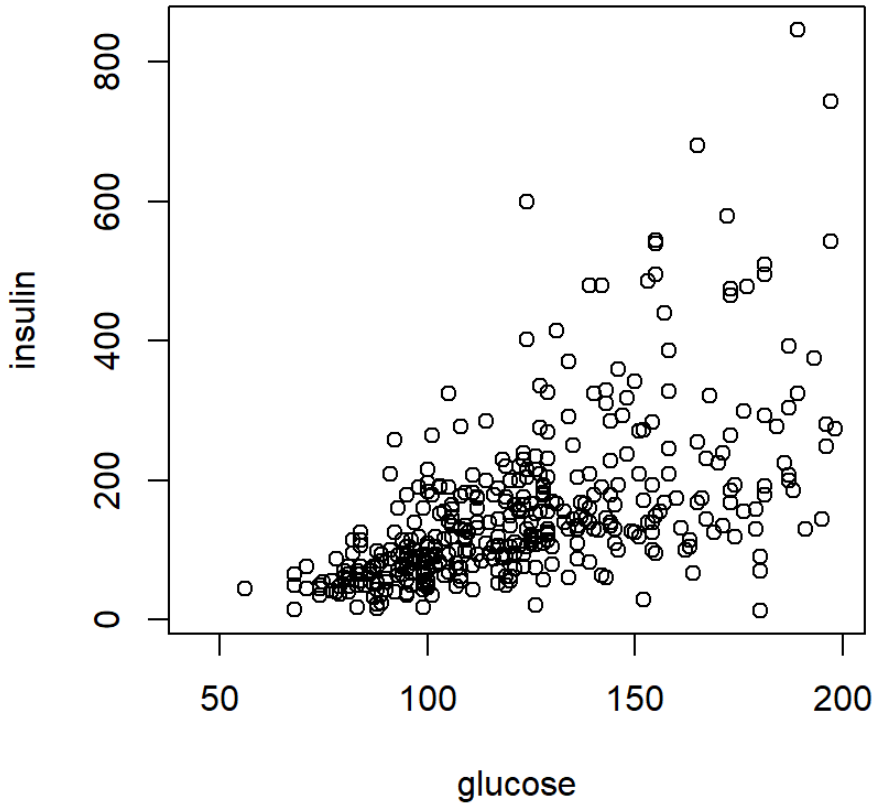
```
# set working directory
setwd("C:/Users/mxi1/OneDrive - Loyola University Chicago/Loyola/STAT 408 Fall 2022")

# read data
pima <- read.csv('pima.csv')

# correlation coefficient without missing values
cor(pima$insulin, pima$glucose, use='complete.obs')

# visualization
plot(insulin~glucose, data = pima)
```

Motivation

- The correlation is about 0.58, which implies a moderate positive relationship between insulin and glucose

- However, what will insulin changes if glucose increases by one unit?

- The simply correlation cannot answer this question



Motivation

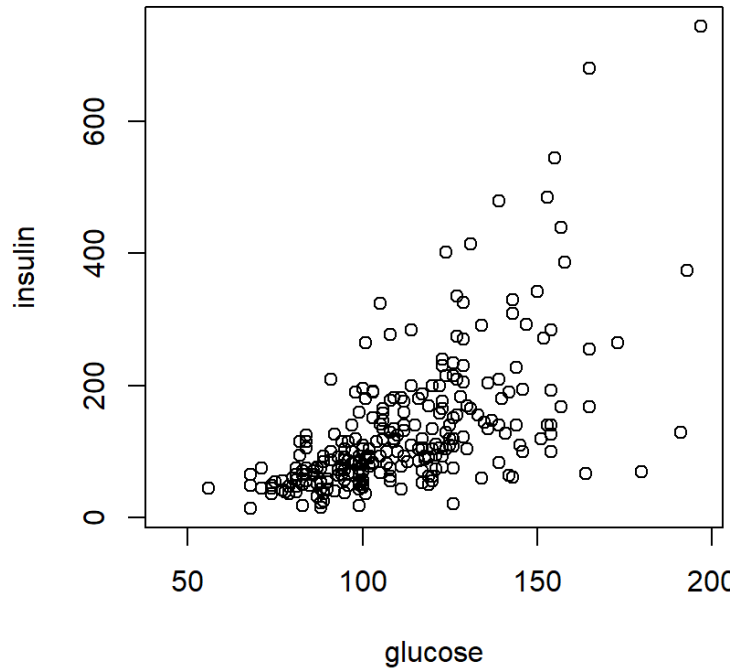
- Another issue is that a simple correlation may hide the difference of insulin~glucose relation between healthy people and diabetes patients

```
# correlation in healthy group
pima.neg <- pima[pima$test=='negative',]
cor(pima.neg$insulin, pima.neg$glucose, use='complete.obs')
plot(insulin~glucose, data = pima.neg)

# correlation in diabetes group
pima.pos <- pima[pima$test=='positive',]
cor(pima.pos$insulin, pima.pos$glucose, use='complete.obs')
plot(insulin~glucose, data = pima.neg)
```

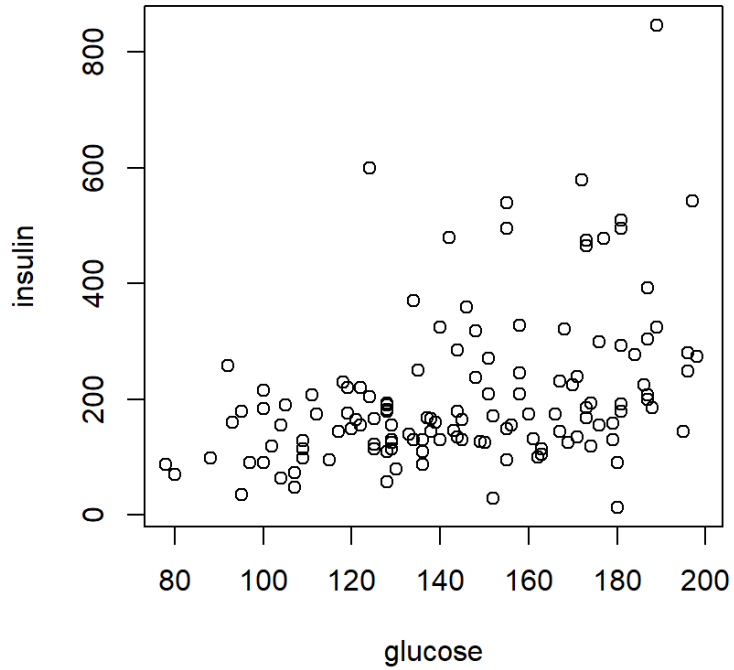
Motivation

- The correlation in healthy group is about 0.61

- The correlation in diabetes group is about 0.39



healthy group

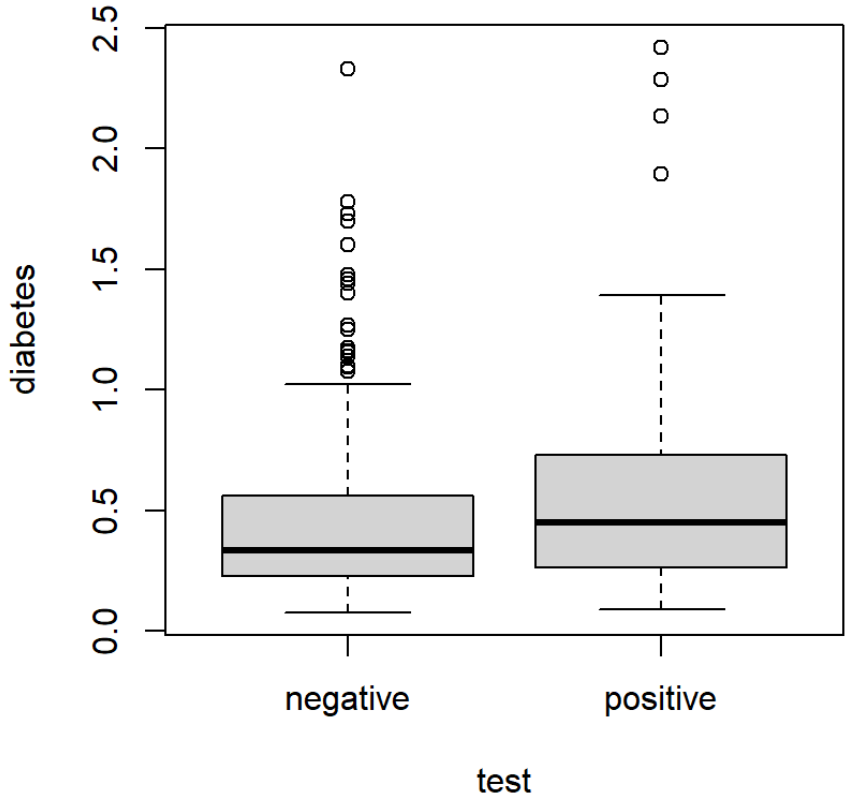


diabetes group

Motivation

- Another question: is diabetes inherited? How likely it is inherited?
  - In other words, how diabetes variable compares between healthy group and diabetes patients?

```
# compare diabetes pedigree function
plot(diabetes ~ test, pima)
```



Motivation

- Is this difference in diabetes pedigree function significant?

- We can use a two-sample t test

```
t.test(pima.neg$diabetes, pima.pos$diabetes)
```

- The p-value is 0.0000061

- But we still don’t know the degree of inheritance
  - Exact change of pedigree function with or without diabetes

Regression Analysis

- Regression analysis is to solve those issues
  - Explicitly assume a function relationship between two variables
  - Use data to estimate this function
  - With estimated function, the relationship between two variables is completely quantified

- There are many function forms we can assume, but we will start with a linear function between two variables
  - This is called linear model or linear regression
  - Most classical, easy to estimate, widely used

Terminology and Notation

- In previous examples, there are two variables X and Y
  - Y is our interest or outcome: insulin and diabetes pedigree function
  - In regression analysis, Y is called response variable or dependent variable

  - X is the variable that “causes” or “predicts” Y: glucose and diabetes test
  - In regression analysis, X is called predictor or independent variable

Simple Linear Model

- A linear model assumes the function that describes the relation between X and Y is

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

- Since this function only contains one predictor, it is called simple linear model or simple linear regression

- If there are more than one predictor in the linear function, it is called multiple linear model or multiple linear regression

- We will first focus on simple linear regression

Simple Linear Model

- What are the meanings of components in simple linear model?

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

- X and Y are predictor and response variables
  - Their values are given by the data

- $\beta_0,\ \beta_1$ are model parameters or coefficients
  - Model parameters define the relationship between X and Y
  - Before we collect data and do model estimation, $\beta_0,\ \beta_1$ are unknown

- ϵ is called error and is the part of Y cannot be linearly explained by X

Simple Linear Model

- Question
  - Why do we believe X and Y follow a linear model?
  - Is this linear relation true?
  - Why do we add error ϵ to the model?
  - How do we interpret parameters β_0 and β_1?

Model Estimation

- Geometrically, linear model is a straight line in two-dimensional space

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

- β_0 is the intercept
- β_1 is the slope

- How do we estimate parameters β_0 and β_1?

Model Estimation

- Remember, the ultimate goal of linear model is to fit the data as much as possible

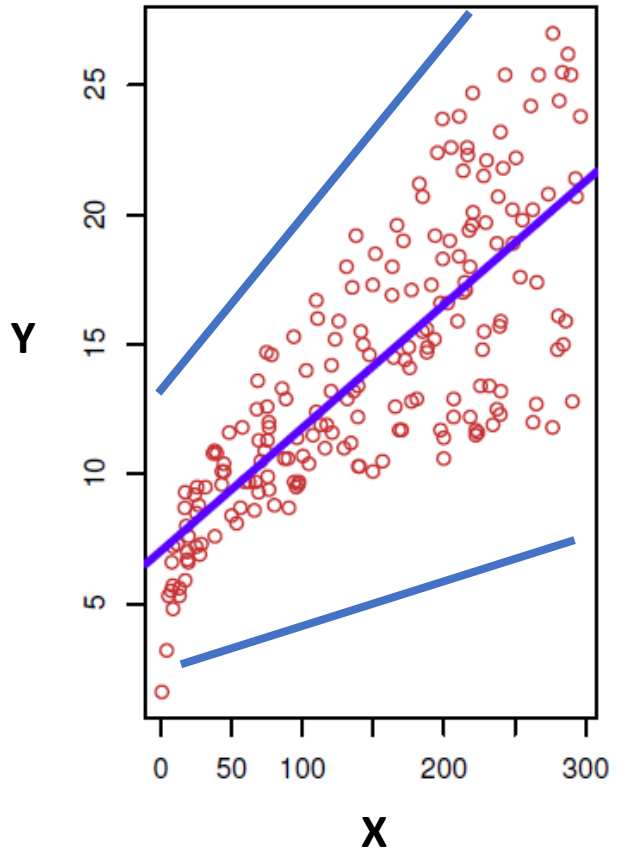
- In simple linear model, each data point is one dot in a X-Y scatter plot

- Suppose the data we have is (x_1, y_1), (x_2, y_2), …, (x_n, y_n), that is, n points in the scatter plot
  - What kind of linear model would fit the data best?

Model Estimation

- The best fitted model should be the straight line with the smallest total distance to all the data points

- Mathematically, how can we achieve this goal for simple linear model?
  - This is an optimization problem



Model Estimation

- Suppose the estimated parameters in simple linear model are $\hat{\beta}_0$ and $\hat{\beta}_1$, then the fitted response $\hat{y}$ is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- For any data point (x_i, y_i), the "distance" between fitted response and true response is

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

- We call this "distance" residual
  - There are n residuals
  - e_i is the ith residual where i = 1, 2, …, n

Model Estimation

- The best fitted model should minimize the sum of all residuals
- Define the residual sum of square RSS as

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^2$$

$$= e_1^2 + e_2^2 + \cdots + e_n^2$$

$$= (y_1 - \beta_0 - \beta_1 x_1)^2 + (y_2 - \beta_0 - \beta_1 x_2)^2 + \cdots + (y_n - \beta_0 - \beta_1 x_n)^2$$

- The estimation of model parameter $\beta_0,\ \beta_1$ is an optimization problem

$$Minimize_{\hat{\beta}_0, \hat{\beta}_1}\ RSS(\beta_0, \beta_1)$$

Model Estimation

- Recall calculus knowledge, we need to take derivative with respect to β_0 and β_1 and let the derivatives be zero

$$\begin{cases} \dfrac{\partial RSS(\beta_0, \beta_1)}{\partial \beta_0} = 0 \\ \dfrac{\partial RSS(\beta_0, \beta_1)}{\partial \beta_1} = 0 \end{cases}$$

- Solving this system of functions will give us the estimated parameters $\hat{\beta}_0, \hat{\beta}_1$

Model Estimation

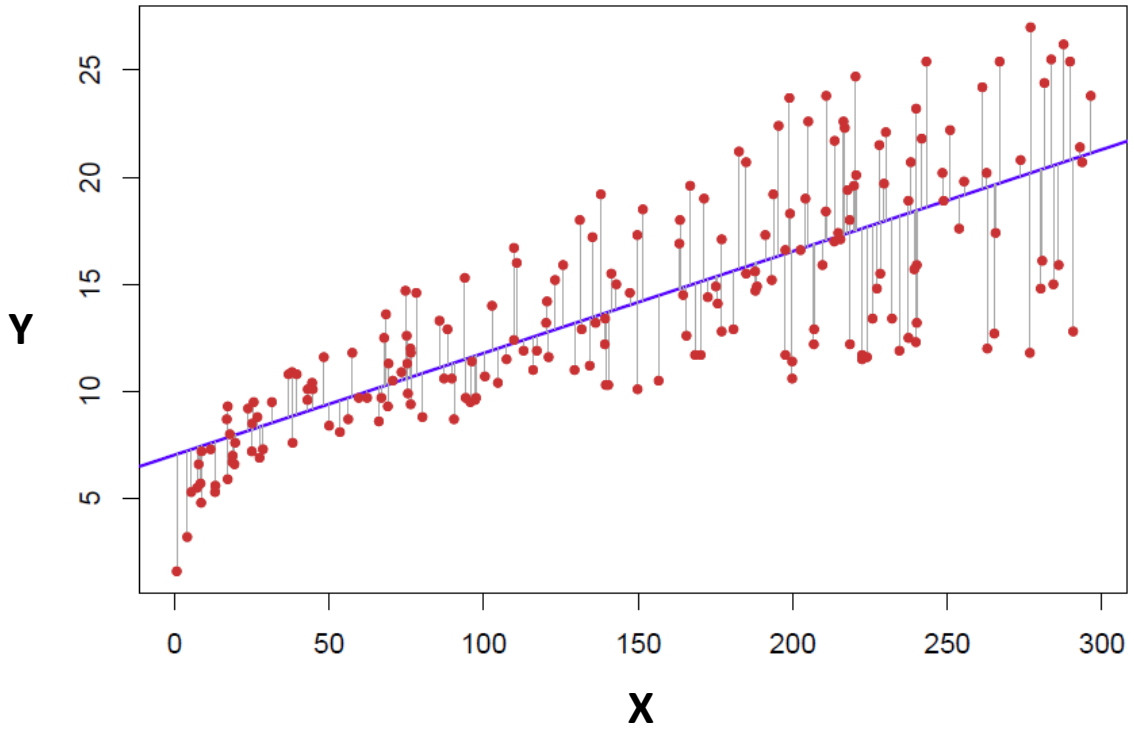
- After tedious algebra, the estimation for simple linear model is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

  where $\bar{x}$ and $\bar{y}$ are sample means

- We call $\hat{\beta}_0$ and $\hat{\beta}_1$ the least square estimation
- The fitted straight line is called regression line

Model Estimation



A fitted regression line (linear model) using least square estimation

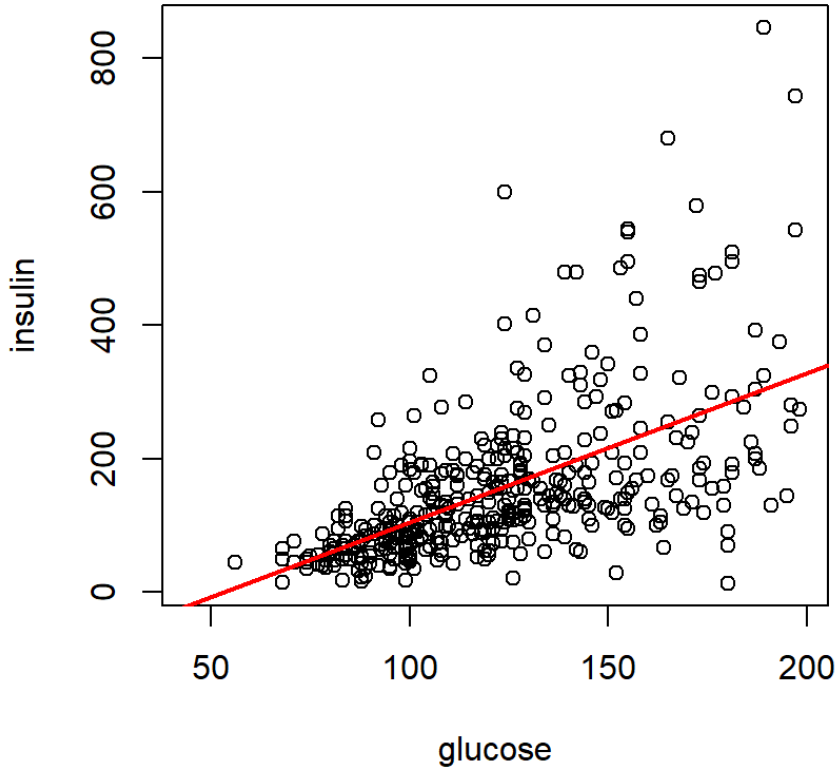
Model Estimation

- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ indicates the linear model fitted by least square estimation must go through the average data point $(\bar{x}, \bar{y})$

- Question
  - Why don't we minimize $\sum_{i=1}^{n} |e_i|$?

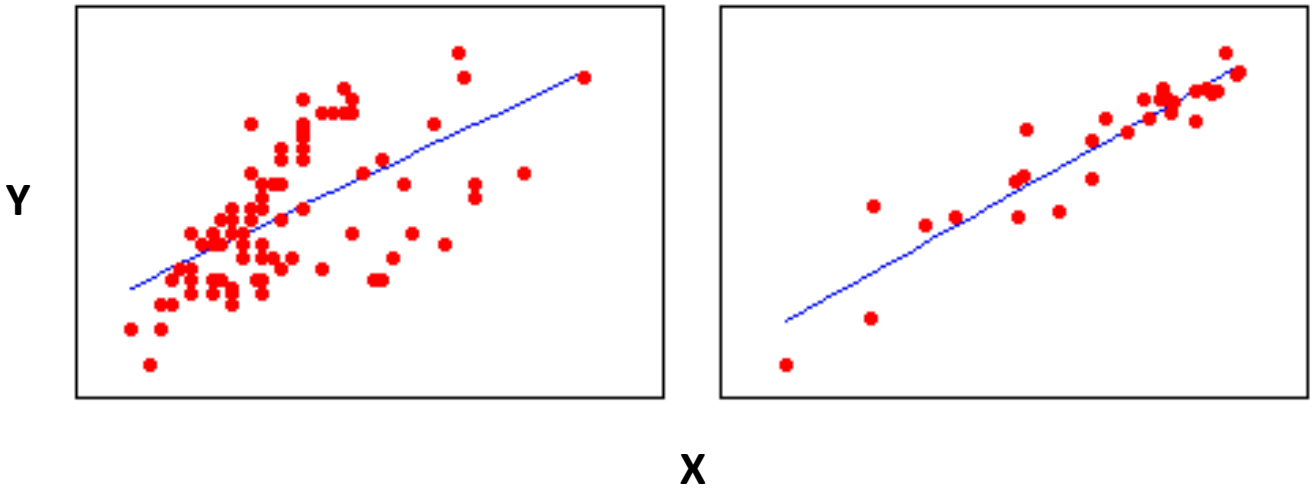
Model Estimation in R

```
# simple linear regression with insulin as response and glucose as predictor
lm.model <- lm(insulin~glucose, data = pima)

# show lease square estimation
lm.model

# plot the fitted regression line in the scatter plot
plot(insulin~glucose, data = pima)
abline(lm.model,col='red', lwd=2)
```



Goodness of Fit

- Even with least square estimation, the linear doesn't fit different data same
- Between the following two linear models, which one fits the data better?



- Can we quantify the "quality" of linear regression model?
  - How about RSS?

Goodness of Fit

- We use the fraction of variance explained by the model to assess the goodness of fit

- Define total sum of squares (TSS)

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- TSS is the total information contained in the dataset

Goodness of Fit

- Recall that the residual sum of squares is the information not explained by the model

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

- Then the information explained by the model is $TSS - RSS$

- The fraction of variance explained by the model is

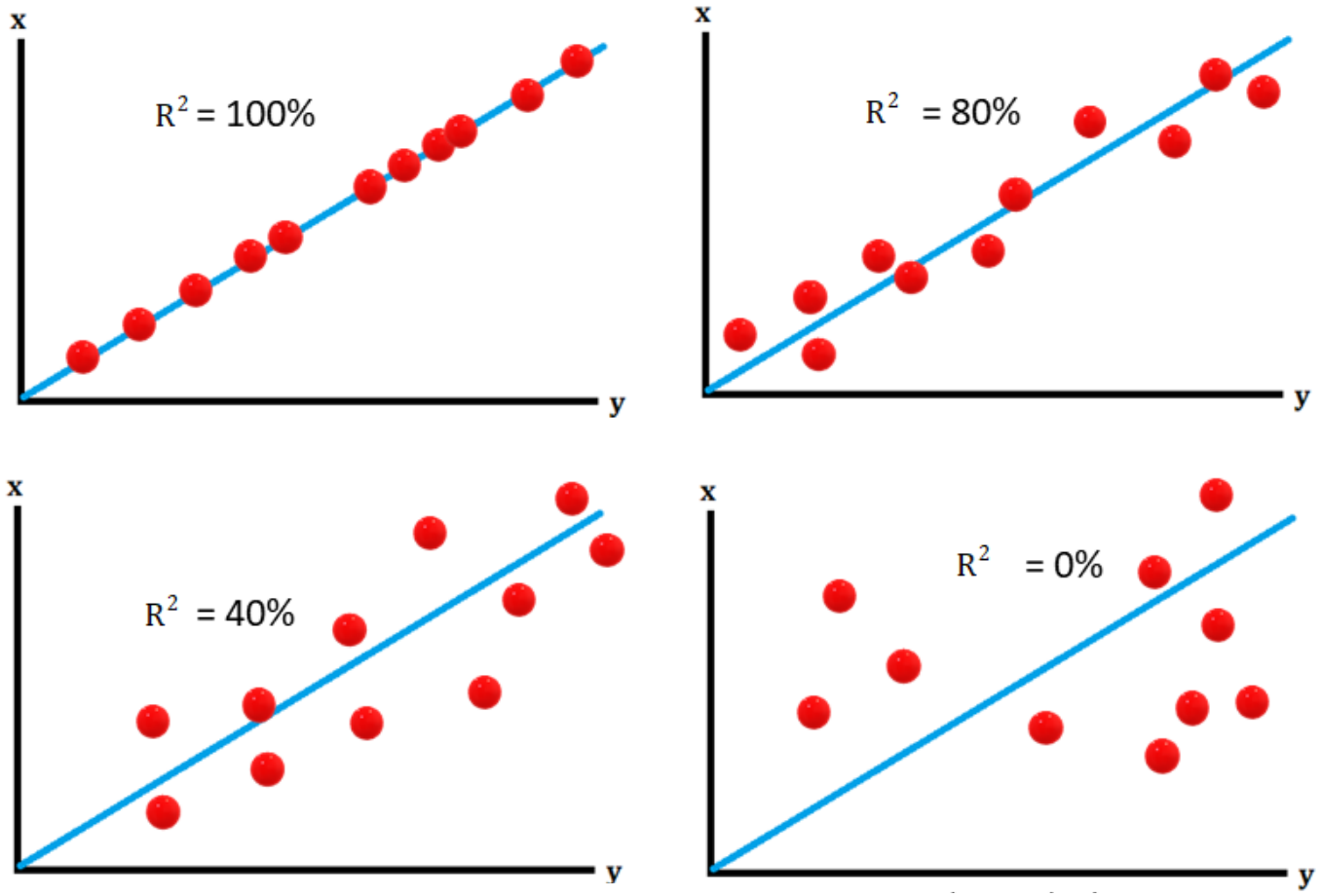
$$R^2 = \frac{TSS - RSS}{TSS}$$

Goodness of Fit

- We use R^2 to assess the goodness of fit for linear model
  1. $0 \leq \mathrm{R}^2 \leq 1$: values closer to one indicates better fits
  2. Better fitting -> smaller RSS -> larger R^2
  3. $\mathrm{R}^2 = 1$: perfect fitting
  4. $\mathrm{R}^2 = 0$: RSS = TSS (e.g., using $\bar{\mathrm{y}}$ for all fitted values, not model at all)

Goodness of Fit

- Comparison of linear models with different R^2



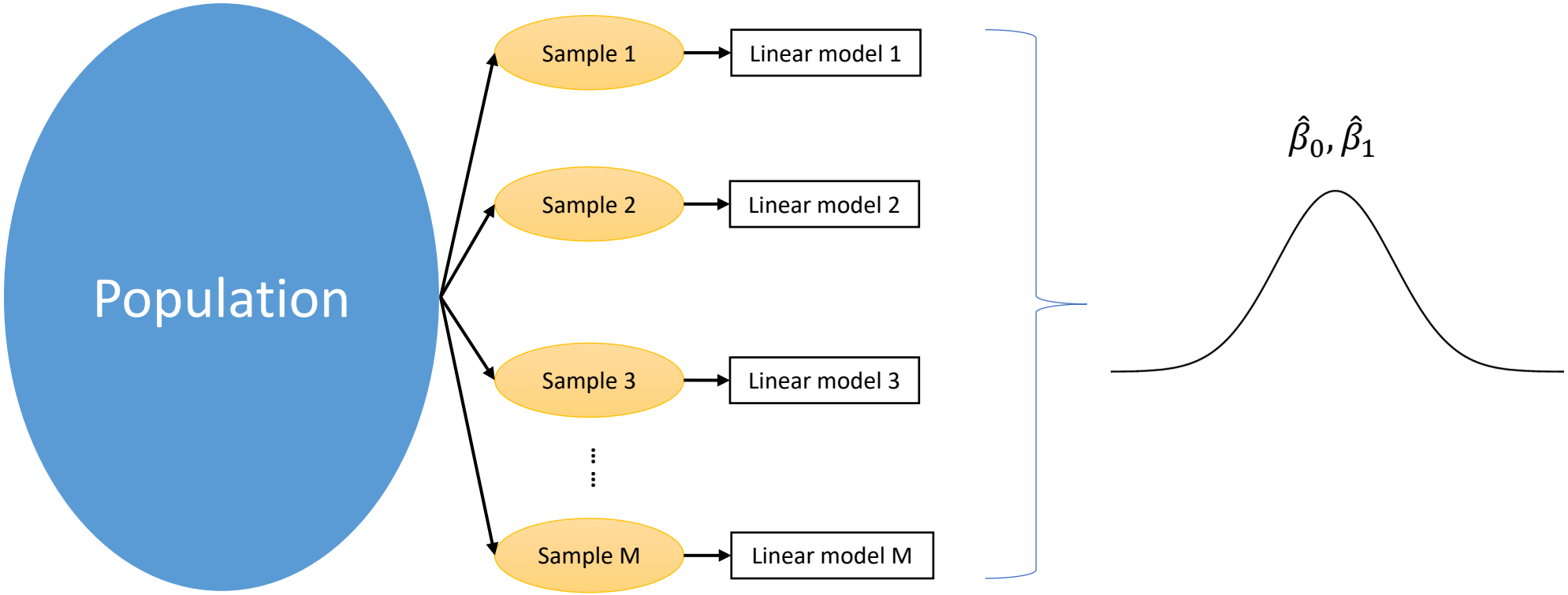
Inference in Simple Linear Model

- The data we used to fit a linear model is one sample from the population
  - Pima is a sample of size 768 from the Pima Indian population

- Imagine if we collect data from the population to obtain another sample, then the linear model fitted on that sample would be different

- Therefore, the estimated parameters $\hat{\beta}_0, \hat{\beta}_1$ in linear model are random variables

Inference in Simple Linear Model



Inference in Simple Linear Model

- In simple linear model, we care most the parameter β_1, because it is the "effect" of X on Y

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

- If the true value of β_1 is zero, then there is no effect

- If the estimated $\hat{\beta}_1 \neq 0$, can we say the true β_1 is nonzero?

Inference in Simple Linear Model

- The answer is no, because $\hat{\beta}_1$ is just one value from its distribution

- We need to conduct a hypothesis test to test the true value of β_1

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

- If we reject H_0, X has significant impact on Y
- If we fail to reject H_0, X is not associated with Y

Inference in Simple Linear Model

- The process of hypothesis test is same as usual
    1. Assume H_0 is true (under H_0)
    2. Find a test statistic (a function of data) and its distribution
    3. Calculate the value of test statistic
    4. Find the probability of observing such a test statistic (p-value)

Inference in Simple Linear Model

- If we assume the error term ϵ follows a normal distribution and different ϵ_i's are independent

$$\epsilon_i \sim N\left(0, \sigma^2\right)$$

$$\epsilon_i \perp \epsilon_j \text{ for } i \neq j$$

- Then under H_0, the t statistic follows a t distribution with n-2 degree of freedom

$$t = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$$

where

$se\left(\hat{\beta}_1\right) = \sqrt{\frac{Var(e)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$ is the standard error of $\hat{\beta}_1$

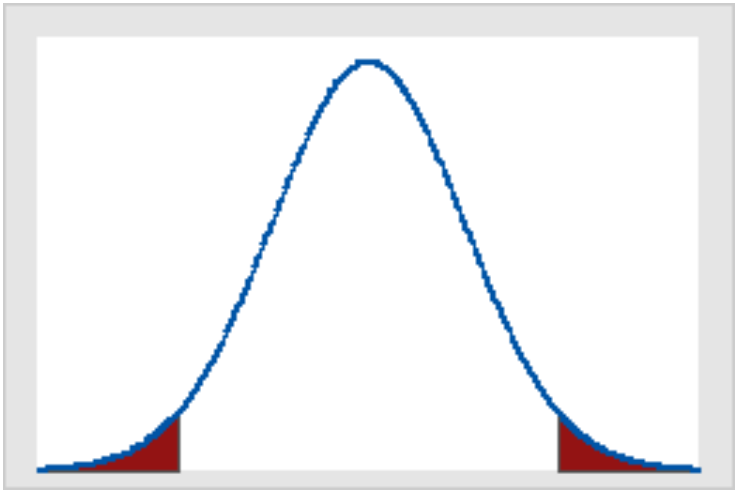
Inference in Simple Linear Model

- Note that the test for β_1 is a <u>two-sided</u> t test

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

- The p-value is the probability of t statistic being too "extreme" (against H_0)



Two Final Questions

1. To conduct the previous t test, we assume

$$\epsilon_i \sim N\left(0, \sigma^2\right)$$

$$\epsilon_i \perp \epsilon_j \text{ for } i \neq j$$

is it a valid assumption?

2. In statistical terms, what are β_0, β_1, $\hat{\beta}_0, \hat{\beta}_1$, and ϵ? (random or non-random)

The Answer

- Recall central limit theorem

  Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with mean μ and variance σ^2. If sample size n is reasonably large (>30), then the sample mean $\bar{X}$ is approximately normally distributed with mean μ and variance $\frac{\sigma^2}{n}$

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \sim N(\mu, \frac{\sigma^2}{n}) \text{ and } n\bar{X} = N(n\mu, \sigma^2)$$

1. Error is the sum of different types of sources (measurement, sampling, human error)
2. Based on central limit theorem, its sum can be approximated by a normal distribution, which is $\epsilon_i \sim N\left(n\mu, \sigma^2\right)$
3. The $n\mu$ part can be absorbed into the intercept β_0

The Answer

1. All model parameters $\beta_0,\ \beta_1$ are fixed (ground truth)
2. All estimated parameters $\hat{\beta}_0, \hat{\beta}_1$ are random variables
3. The error term ϵ is a random variable
4. The dataset (X, Y) used to fit the model is a sample from the population