Introduction to Data Science

Unit 5(Part I) – Correlation and Simple Linear Regression

Preet Kanwal
Assistant Professor
Department of CSE
PESU Bangalore

1. Bivariate Data

2. Bivariate Analysis

3. Dependent(Y) and independent(X) variables

Bivariate Data

- The primary purpose of bivariate data is to compare the two sets of data or to find a relationship between the two variables. For example:
  - Weights and heights of college students.
- Bivariate data could also be two sets of items that are dependent on each other. For example:
  - Ice cream sales compared to the temperature that day.
  - Traffic accidents along with the weather on a particular day.
- Bivariate data has many practical uses in real life. For example:
  - It is pretty useful to be able to predict when a natural event might occur.

Bivariate Analysis

- Bivariate analysis means the analysis of bivariate data; used to find out if there is a relationship between two sets of values.
- It usually involves the variables X and Y, denoted as (X, Y).
- With bivariate analysis, there is a Y value for each X.
- The results from bivariate analysis are usually stored in a two-column data table.

Dependent(Y) and independent(X) variables

- The **dependent variables** represent the output or outcome whose variation is being studied. [most common symbol for the input is x]

- The **independent variables** represent inputs or causes, i.e. potential reasons for variation. [most common symbol for the output is y]

- Models test or explain the effects that the independent variables have on the dependent variables. [y = f(x)]

- It is possible to have multiple independent variables and/or multiple dependent variables.

Identify Dependent and independent variables

1) Relationship between caloric intake and weight.
2) Effect of temperature on pigmentation.
3) Effect of fertilizer on plant growth.

Solution : Dependent and independent variables

Independent Variable (X)	Dependent Variable (Y)
1. Caloric intake	Weight
2. Temperature	Pigmentation
3. Amount of fertilizer used	Growth in height or mass of the plant

Note

- Depending on the context, an independent variable is sometimes called a **"predictor variable", "regressor"**, "controlled variable", "manipulated variable", "explanatory variable", or "input variable."

- Depending on the context, a dependent variable is sometimes called a "**response variable**", **"regressand"**, "**predicted variable**", "measured variable", "explained variable", "experimental variable", "responding variable", "outcome variable", or "output variable"

Types of Bivariate Analysis

Types of Bivariate Analysis

1. Scatter plots : gives you a visual idea of the pattern that your variables follow.

2. Correlation Coefficients : This coefficient tells you if the variables are related. Refers to the extent to which two variables have a linear relationship with each other.

3. Regression Analysis : Regression analysis helps provide an equation for the curve or line that depicts pattern that your variables follow.

- It can also give you the correlation coefficient.
- It is a statistical process for estimating the relationships among variables.

Note:

The methods described here work only when there exists a linear relationships between bivariate data.

Correlation

Prof. Preet Kanwal

Correlation

- **Correlation** is any of a broad class of statistical relationships involving dependence; **commonly refers to the extent to which two variables have a linear relationship with each other (or related to each other).**

- Correlations are useful because they can indicate a **predictive relationship** that can be exploited in practice.

- There are several correlation coefficients that measure the degree of correlation.

Pearson's correlation coefficient

- Correlation coefficient detects only linear dependencies between two variables.

- The most familiar measure of dependence between two quantities is the Pearson product-moment correlation coefficient, or "Pearson's correlation coefficient", commonly called simply "**the correlation coefficient".**

- If the variables are independent, Pearson's correlation coefficient is 0.

Note

- If X is correlated with Y, there could be five explanations:
  - X causes Y
  - Y causes X
  - X causes Y and Y causes X
  - Some third variable Z causes X and Y
  - The correlation is a coincidence; there is no causal relationship between X and Y.

Population Pearson correlation coefficient – ρ

For a population, Pearson's correlation coefficient when applied to a population is commonly represented by the Greek letter ρ (rho) and may be referred to as the population correlation coefficient or the population Pearson correlation coefficient.

Sample Pearson correlation coefficient – r

Pearson's correlation coefficient when applied to a sample is commonly represented by the letter r and may be referred to as the sample correlation coefficient or the sample Pearson correlation coefficient.

Sample Pearson correlation coefficient - r

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

Expanding s_x and s_y ,

$$r = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}}$$

Computing Formulas

1. $$\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$$

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

3. $$\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - n\,\overline{x}\,\overline{y}$$

Points to Ponder

- The absolute values of both the sample and population Pearson correlation coefficients are less than or equal to 1.
- Correlations equal to 1 or −1 correspond to data points lying exactly on a line.
- The Pearson correlation coefficient is symmetric:

$$corr(X,Y) = corr(Y,X)$$

- Correlation coefficient is invariant to separate changes in location and scale in the two variables.

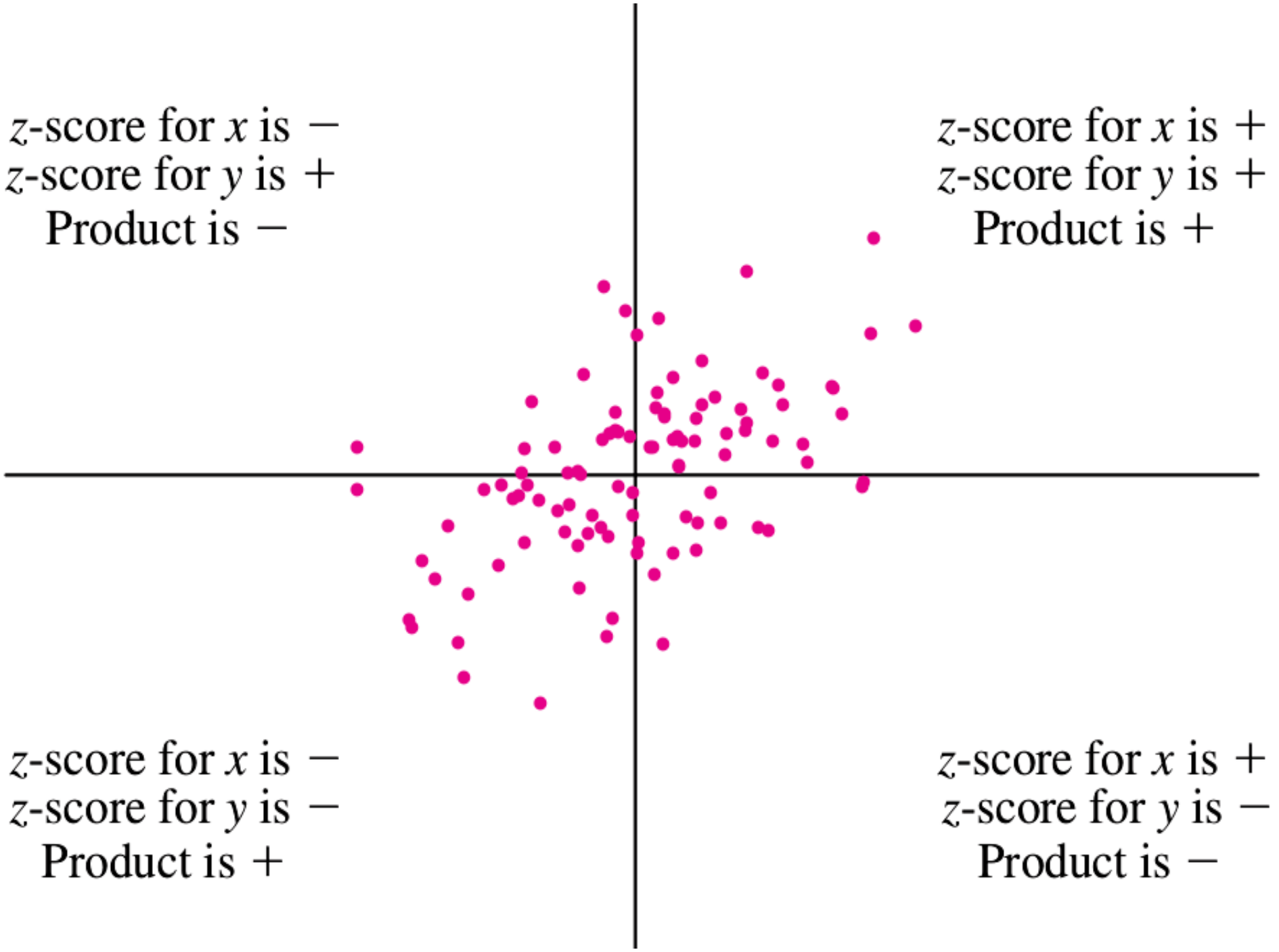
The correlation coefficient remains unchanged under each of the following operations:

- Multiplying each value of a variable by a positive constant.
- Adding a constant to each value of a variable.

Working of Correlation Coefficient

- The correlation coefficient is positive if X_i and Y_i tend to be simultaneously greater than, or simultaneously less than, their respective means.

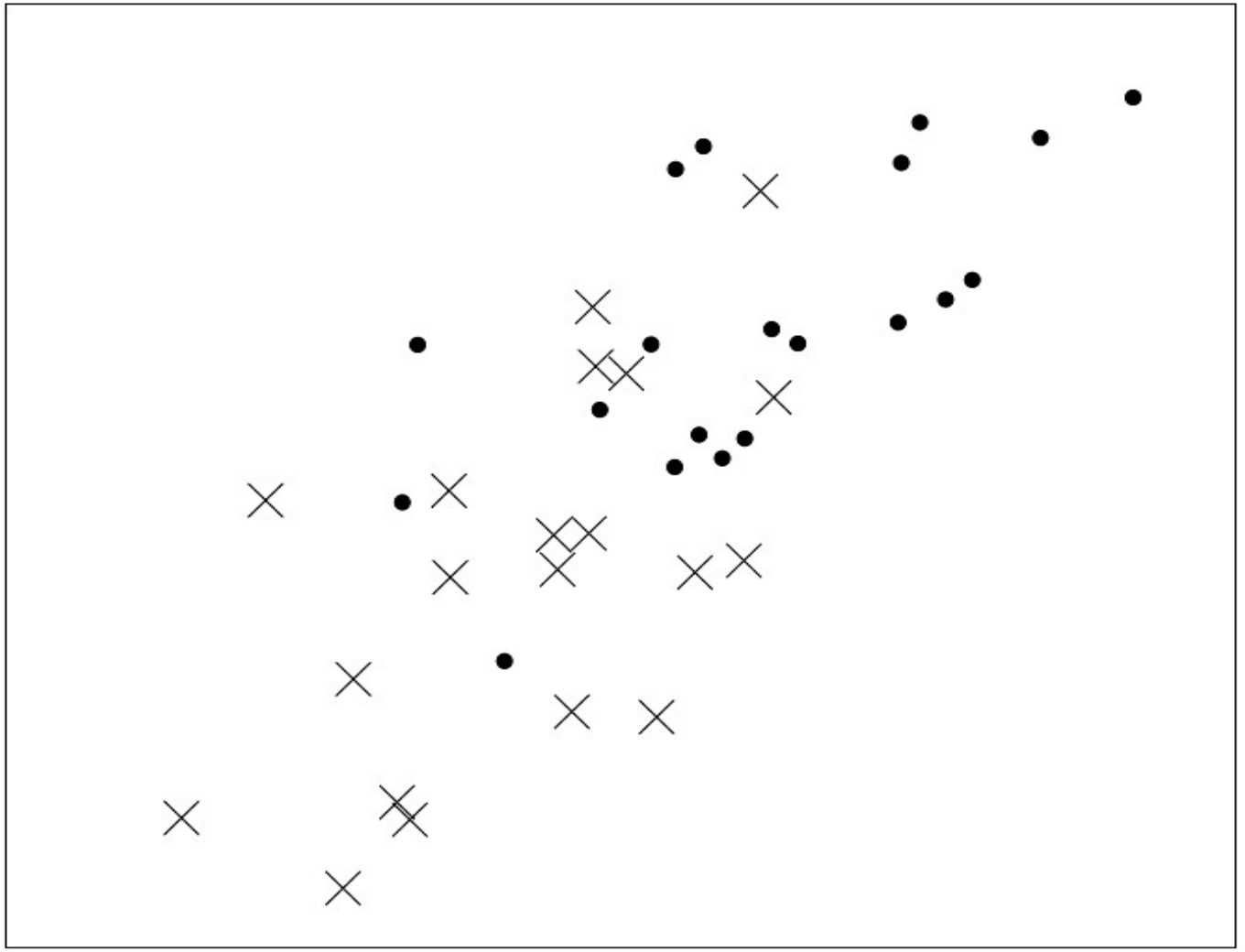
- The correlation coefficient is negative if X_i and Y_i tend to lie on opposite sides of their respective means.


z-score for x is −
z-score for y is +
Product is −
z-score for x is +
z-score for y is +
Product is +
z-score for x is −
z-score for y is −
Product is +
z-score for x is +
z-score for y is −
Product is −

Problem 1

- An investigator collected data on heights and weights of college students.
- The correlation between height and weight for men was about 0.6, and for women it was about the same.
- If men and women are taken together, will the correlation between height and weight be more than 0.6, less than 0.6, or about equal to 0.6?

Problem 1 : Solution



- (dot) represents height and weight of men

x (cross) represents height and weight of women

Problem 1 : Solution

- The heights and weights for the men (dots) are on the whole greater than those for the women (xs).
- Therefore the scatterplot for the men is shifted up and to the right.
- The overall plot exhibits a higher correlation than either plot separately.
- The correlation between heights and weights for men and women taken together will be more than 0.6.

Correlation does not imply causation

Correlation does not imply causation

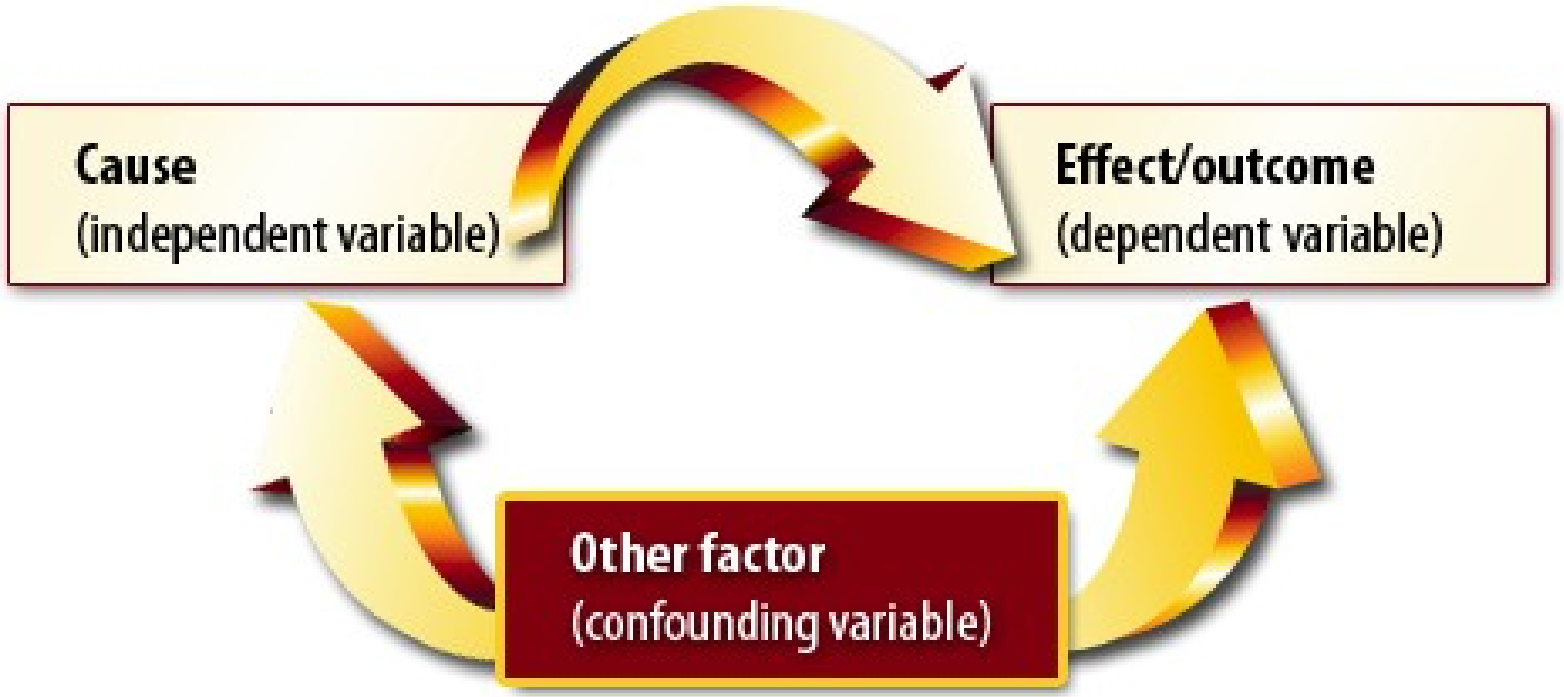
- Correlation cannot be used to infer a **causal relationship** between the variables.

- In other words, a correlation can be taken as evidence for a possible causal relationship, but cannot indicate what the causal relationship, if any, might be.

Example : Relationship between reading ability and shoe size. (Strong Positive Correlation).

This does not mean large shoes cause good reading skills.

The part age plays in this example is known as a "confounding variable" or "confounding factor," and is something that is not being controlled for in the experiment.



Confounding

- Something that interferes with or obscures your research.
- A confounding variable is an “extra” variable that you didn’t account for.
- They can ruin an experiment and give you useless results. They can suggest there is correlation when in fact there isn’t.
- Confounding variables can cause two major problems:
  - Increase variance
  - Introduce bias.
- A confounding variable are like extra independent variables that are having a hidden effect on your dependent variables.
- A confounding variable can be what the actual cause of a correlation is, hence any studies must take these into account and find ways of dealing with them.

MURDER AND ICE CREAM

It is known that throughout the year, murder rates and ice cream sales are highly positively correlated. There are three possible explanations for this correlation:

Possibility #1: Murders cause people to purchase ice cream.

Possibility #2: Purchasing ice cream causes people to murder or get murdered.

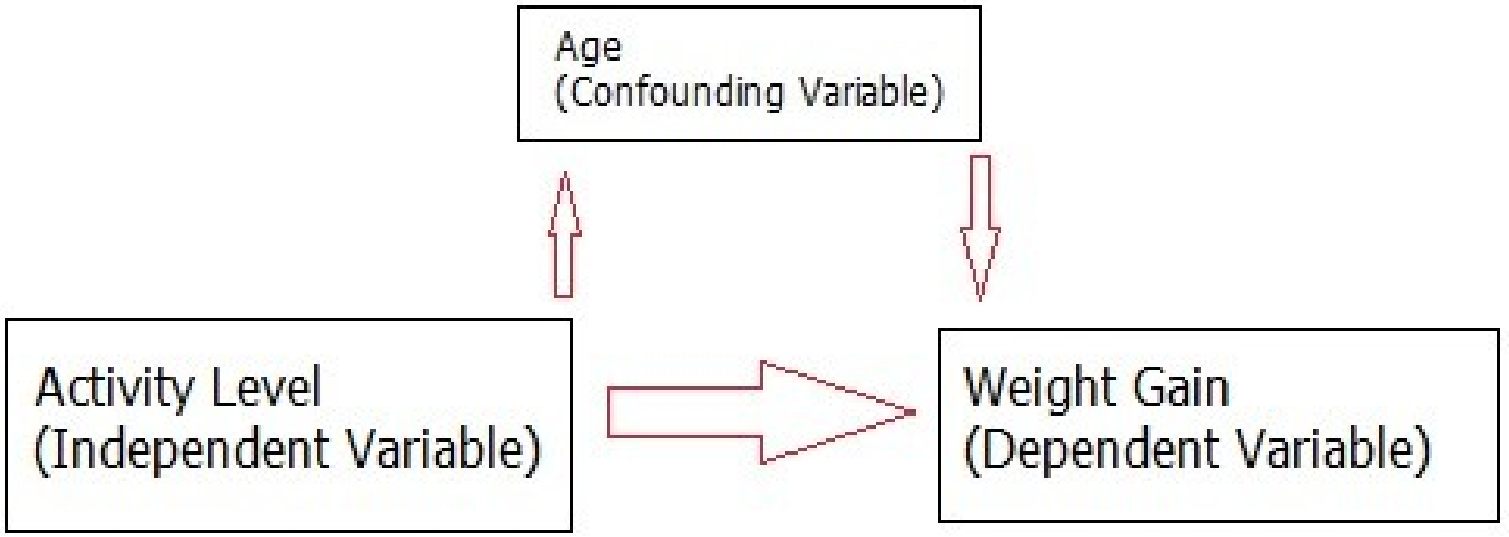
Possibility #3: There is a third variable—a confounding variable—which causes the increase in BOTH ice cream sales AND murder rates. For instance, the weather.

Relationship between the force you apply to a ball and the distance the ball travels.

- Naturally, you predict that the more force you apply, the further the ball will travel.
- After you run your experiment, you observe that the ball travels further in Condition 2 than it does in Condition 1.
- In other words, you find that the less force you apply, the further the ball travels.
- No, there's a clear confounding variable in this experimental design: the angle of the slope.

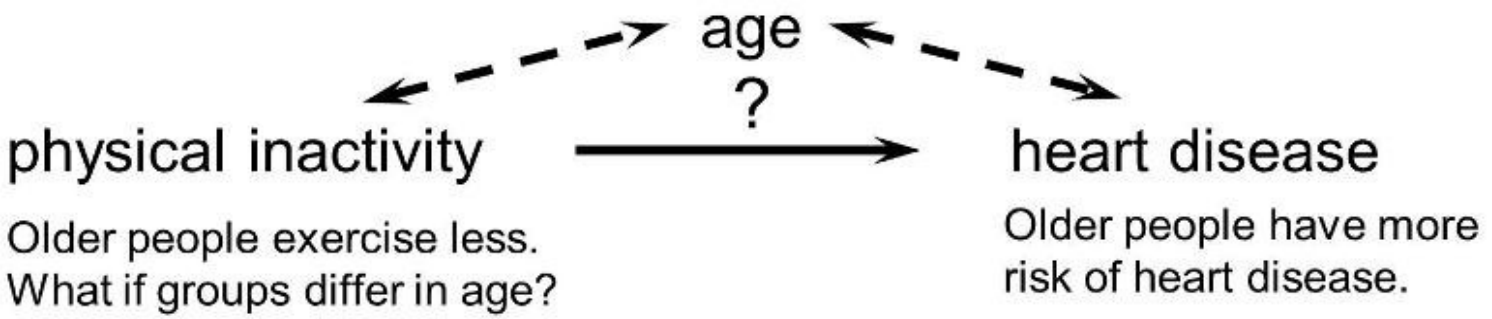
Example : Relationship between Physical activity and weight gain.

- One confounding variable is **how much people eat.** It's also possible that men eat more than women; this could also make **sex** a confounding variable.
- Nothing was mentioned about starting weight, occupation or age either.
- A poor study design like this could lead to bias.
- For example, if all of the women in the study were middle-aged, and all of the men were aged 16, age would have a direct effect on weight gain.
- That makes **age** a confounding variable.



Example : Strength of association between physical inactivity and heart disease.

- Age is a confounding factor because it is associated with the exposure (meaning that older people are more likely to be inactive), and it is also associated with the outcome (because older people are at greater risk of developing heart disease).



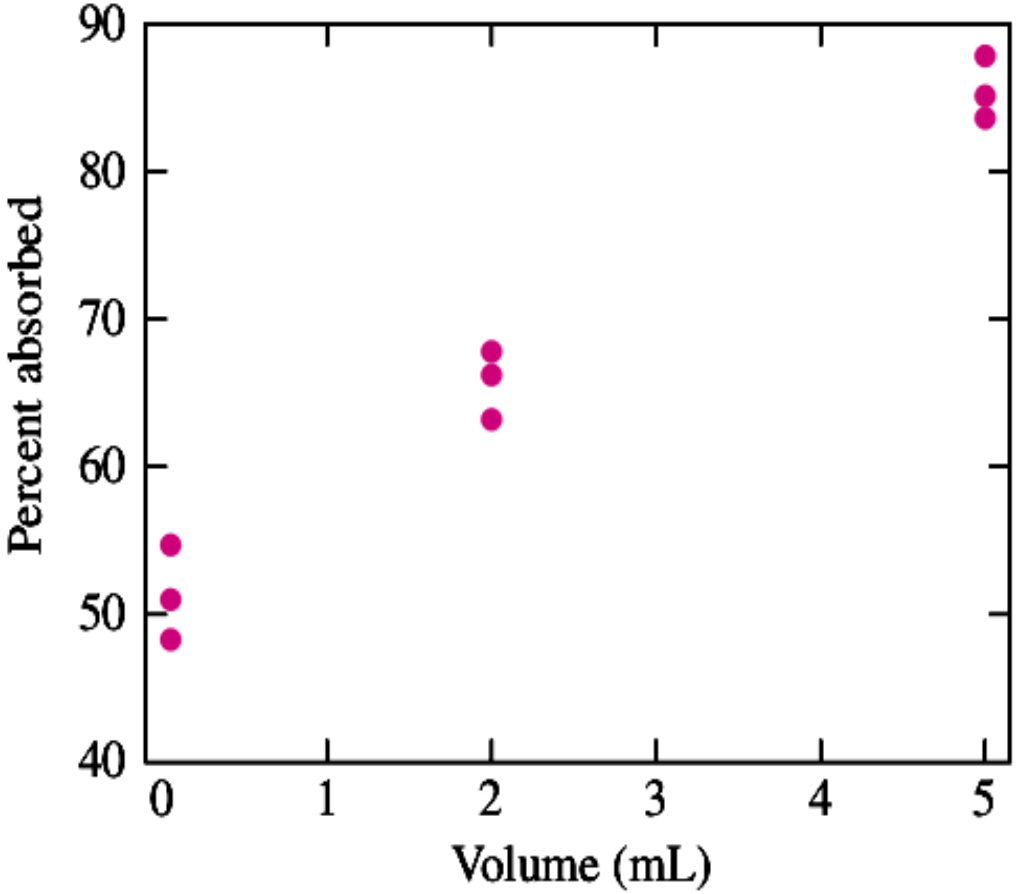
- Or, if the age distribution is similar in the exposure groups being compared, then age will not cause confounding.

Problem

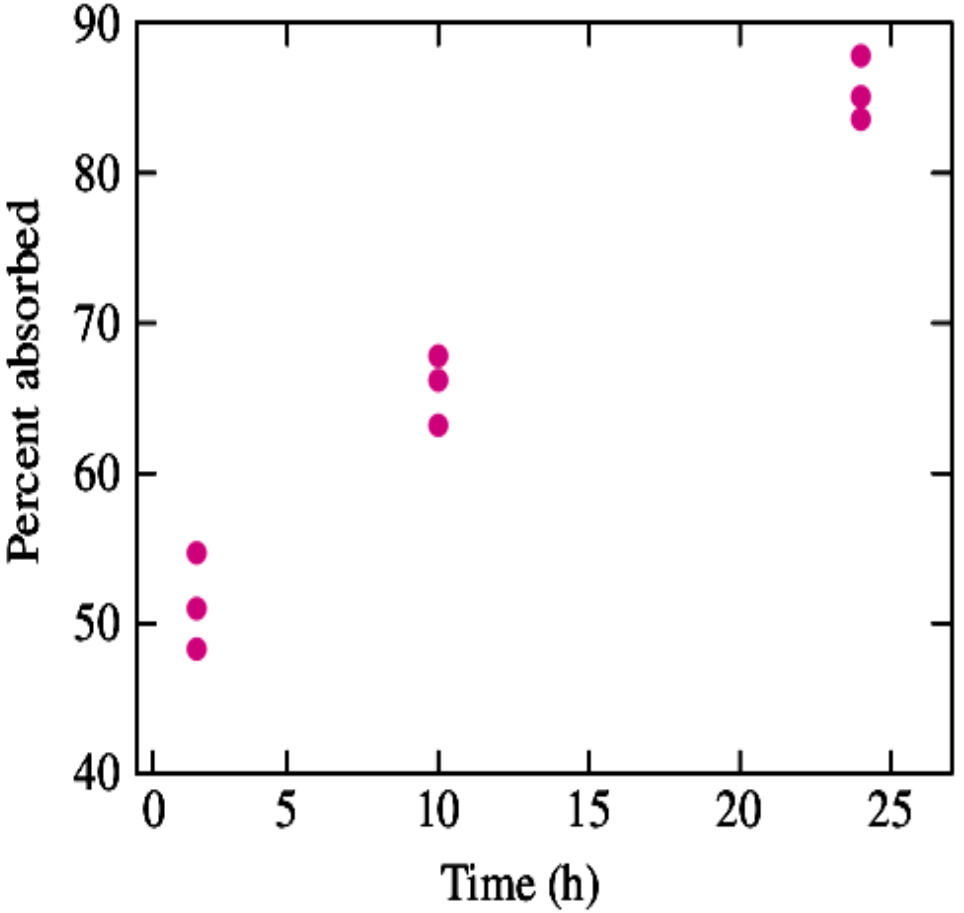
- An environmental scientist is studying the rate of absorption of a certain chemical into skin.
- She places differing volumes of the chemical on different pieces of skin and allows the skin to remain in contact with the chemical for varying lengths of time.
- She then measures the volume of chemical absorbed into each piece of skin.
- She obtains the results shown in the following table.

Volume (mL)	Time (h)	Percent Absorbed
0.05	2	48.3
0.05	2	51.0
0.05	2	54.7
2.00	10	63.2
2.00	10	67.8
2.00	10	66.2
5.00	24	83.6
5.00	24	85.1
5.00	24	87.8

The scientist plots the percent absorbed against both volume and time. She wants to determine whether it is the time or the volume that is having an effect.



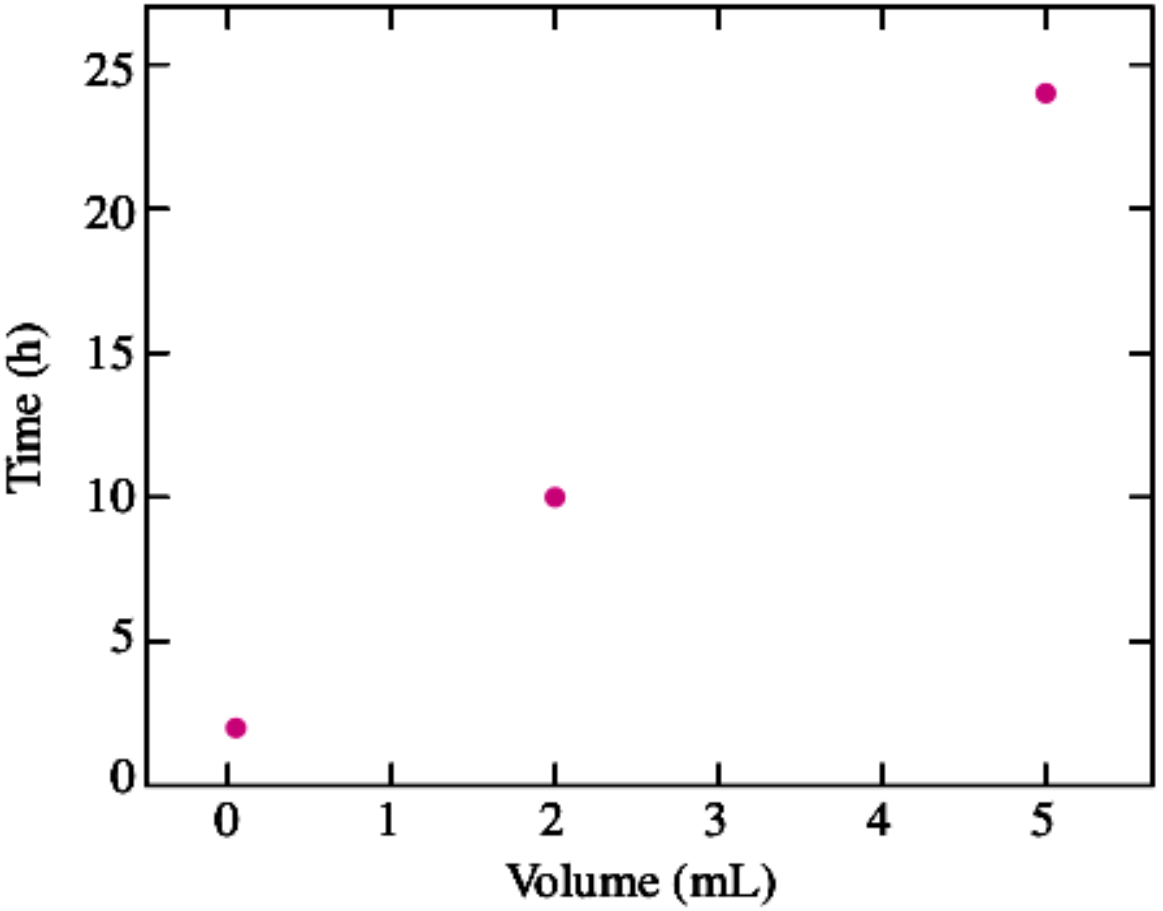
r = 0.988



r = 0.987

- She concludes that increasing the volume of the chemical causes the percentage absorbed to increase.
- She concludes that increasing the time that the skin is in contact with the chemical causes the percentage absorbed to increase as well.

Are these conclusions justified?



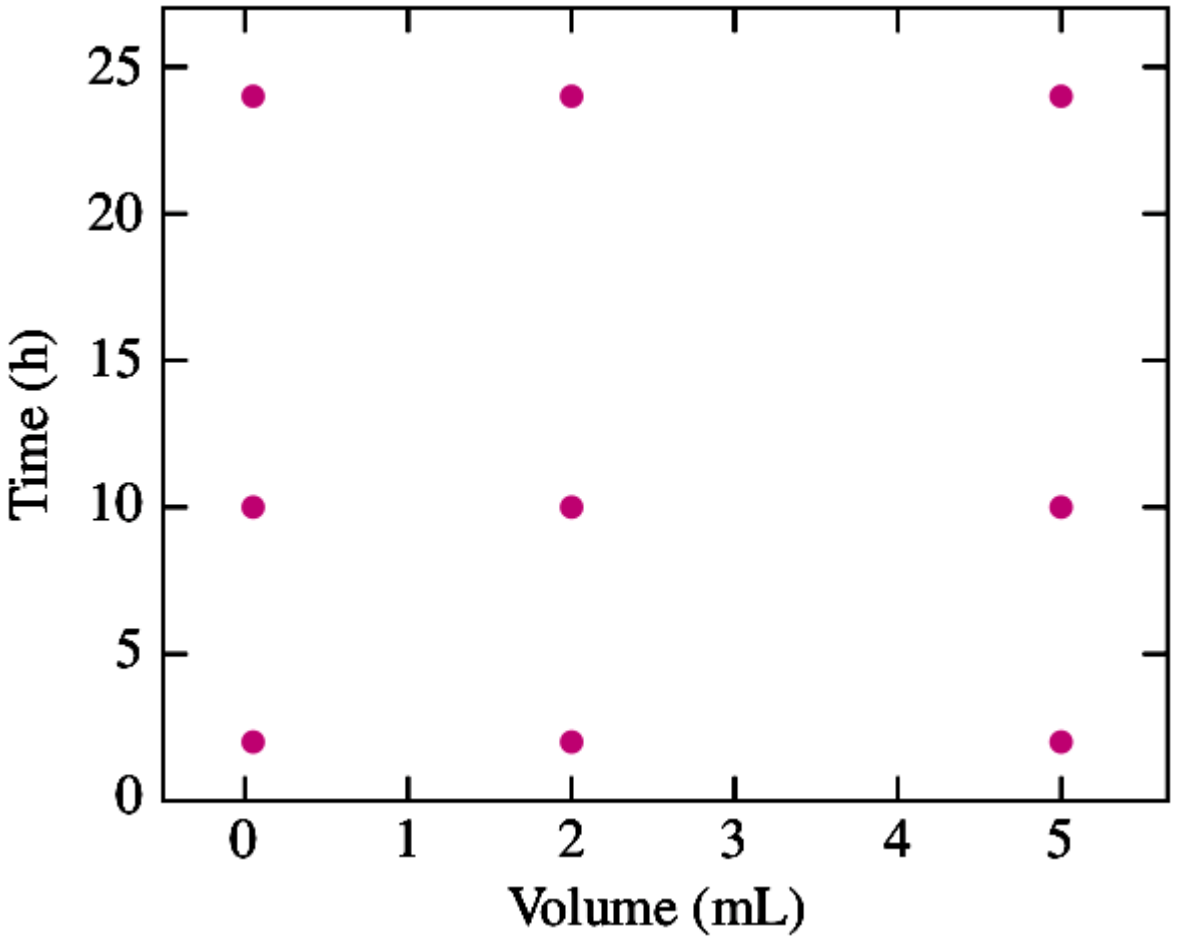
r = 0.999

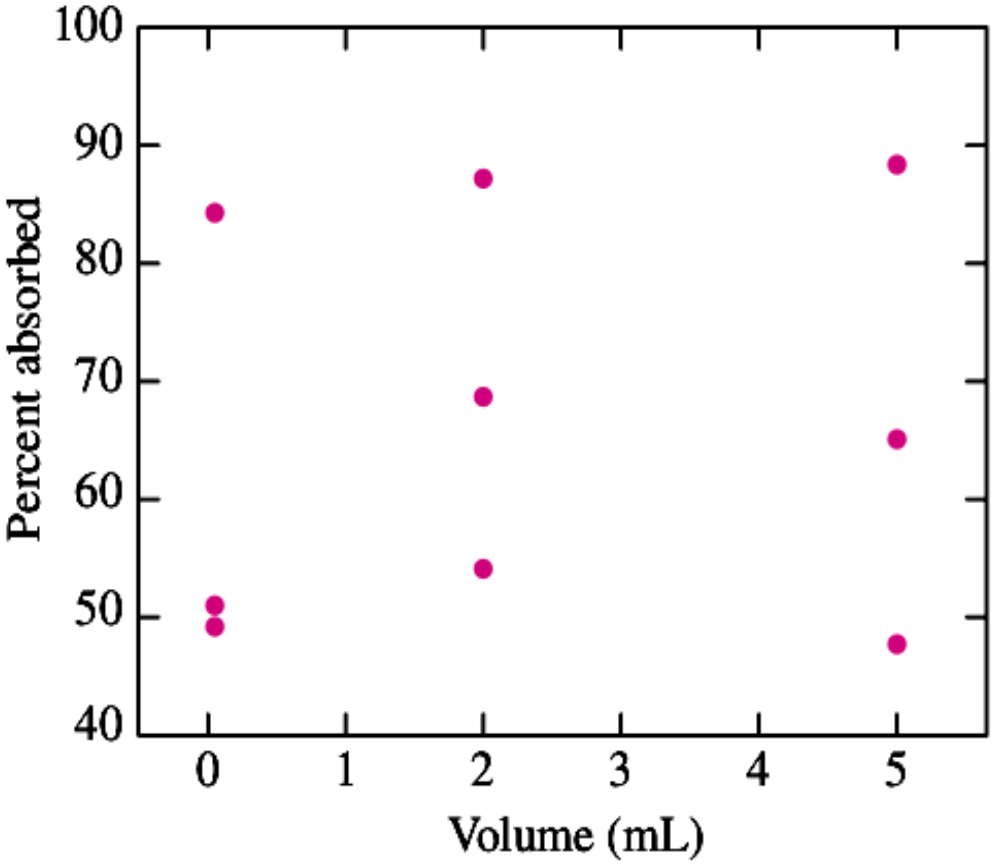
- If either time or volume affects the percentage absorbed, both will appear to do so, because they are highly correlated with each other.
- For this reason, it is impossible to determine whether it is the time or the volume that is having an effect.
- This relationship between time and volume resulted from the design of the experiment and should have been avoided.

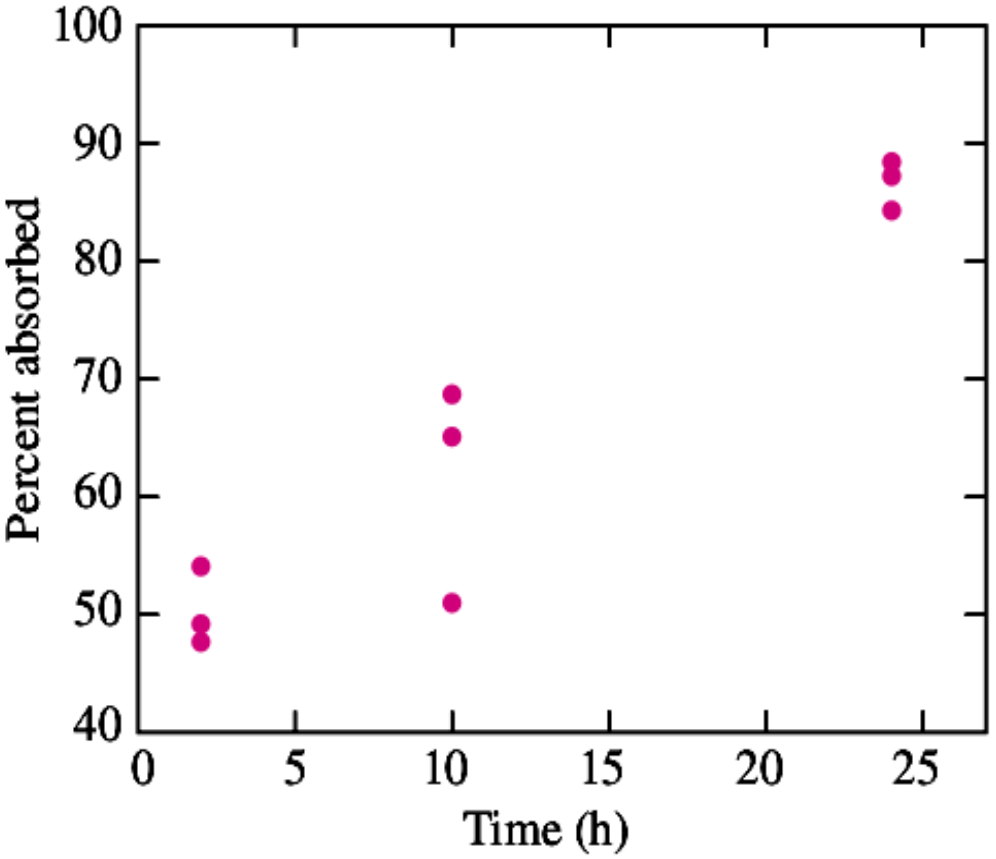
The scientist has repeated the experiment, this time with a new design.

Volume (mL)	Time (h)	Percent Absorbed
0.05	2	49.2
0.05	10	51.0
0.05	24	84.3
2.00	2	54.1
2.00	10	68.7
2.00	24	87.2
5.00	2	47.7
5.00	10	65.1
5.00	24	88.4

Time and Volume uncorrelated







r = 0.121

r = 0.952

She concludes that increasing the time that the skin is in contact with the chemical will cause the percentage absorbed to increase.

Is the conclusion justified?

- Before making a final conclusion that increasing the time actually causes the percentage absorbed to increase, the scientist must make sure that there are no other potential confounders around.

Controlled Experiments Reduce the Risk of Confounding

- In controlled experiments, confounding can often be avoided by choosing values for factors in a way so that the factors are uncorrelated.

- Observational studies are studies in which the values of factors cannot be chosen by the experimenter. Studies involving public health issues, such as the effect of environmental pollutants on human health, are usually observational, because experimenters cannot deliberately expose people to high levels of pollution. In these studies, confounding is often difficult to avoid.

Correlation and linearity

Importance of visual examination of the data

Correlation and linearity

- Correlation Coefficient Measures Only Linear Association, and its value **does not completely characterize the relationship.**

- Correlation coefficient, as a summary statistic, **cannot replace visual examination of the data.**

Importance of visual examination of the data

- Anscombe illustrated the importance of graphing data with these four data sets known as "**Anscombe's quartet**"

- He stressed that "a computer should make both calculations and graphs",

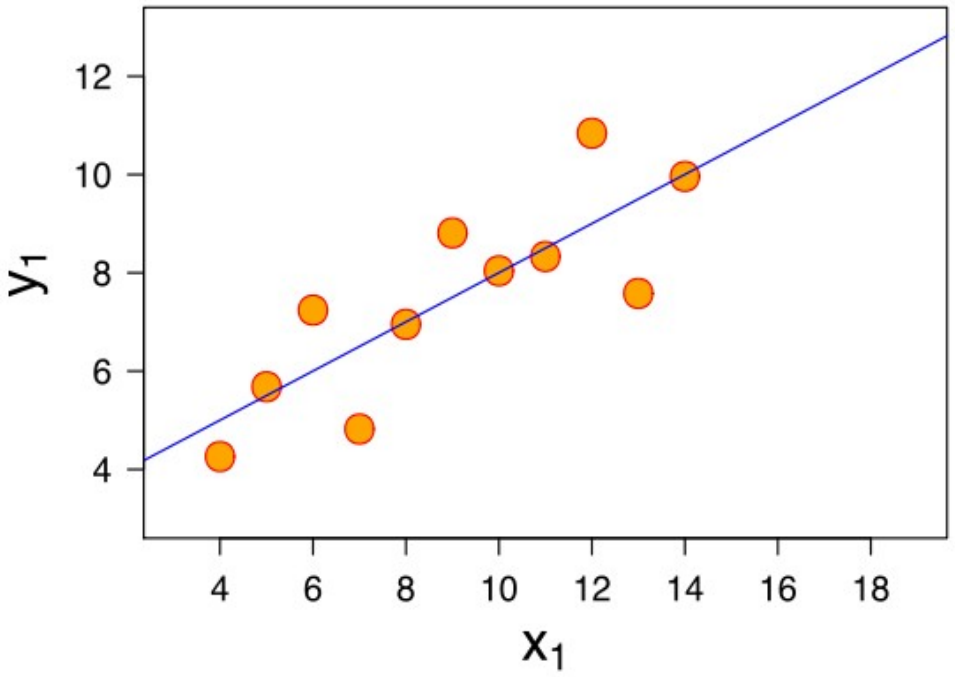
Francis Anscombe

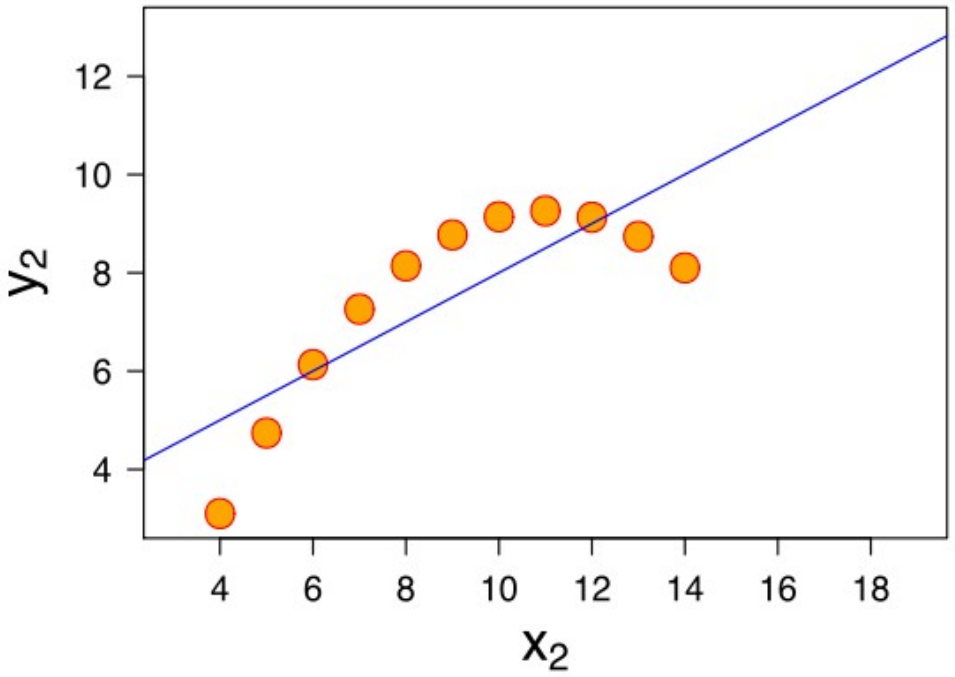
(May 1918 – October 2001)

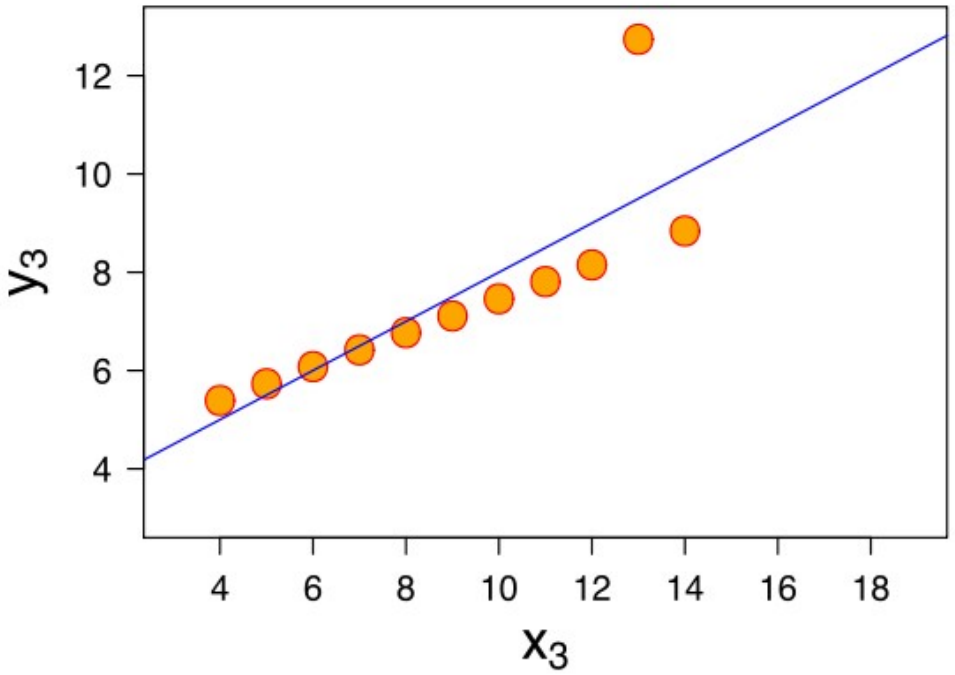
American Statistician

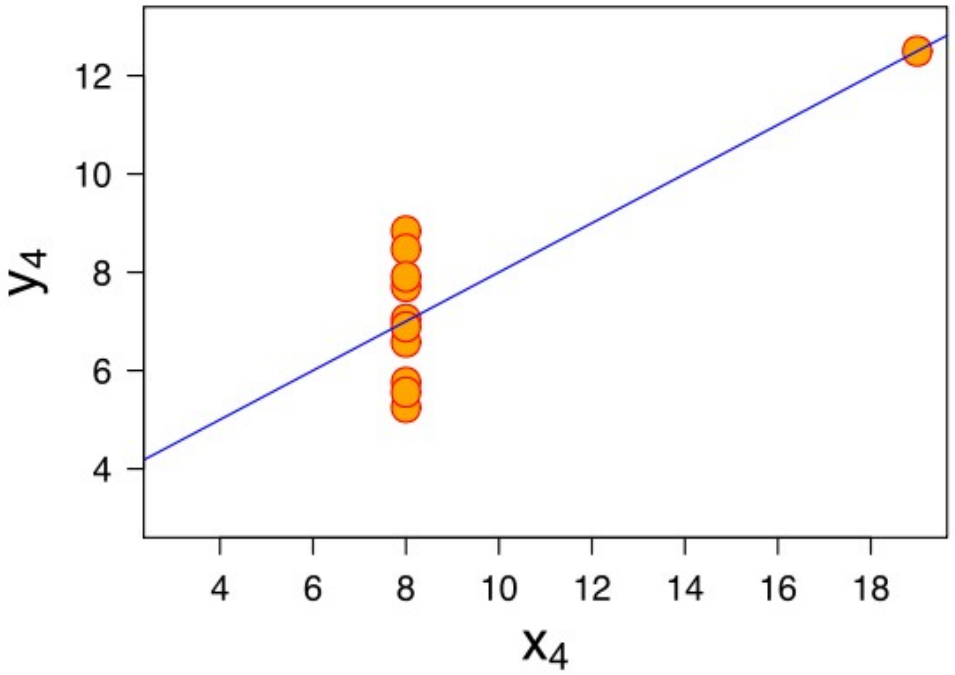
Anscombe's quartet

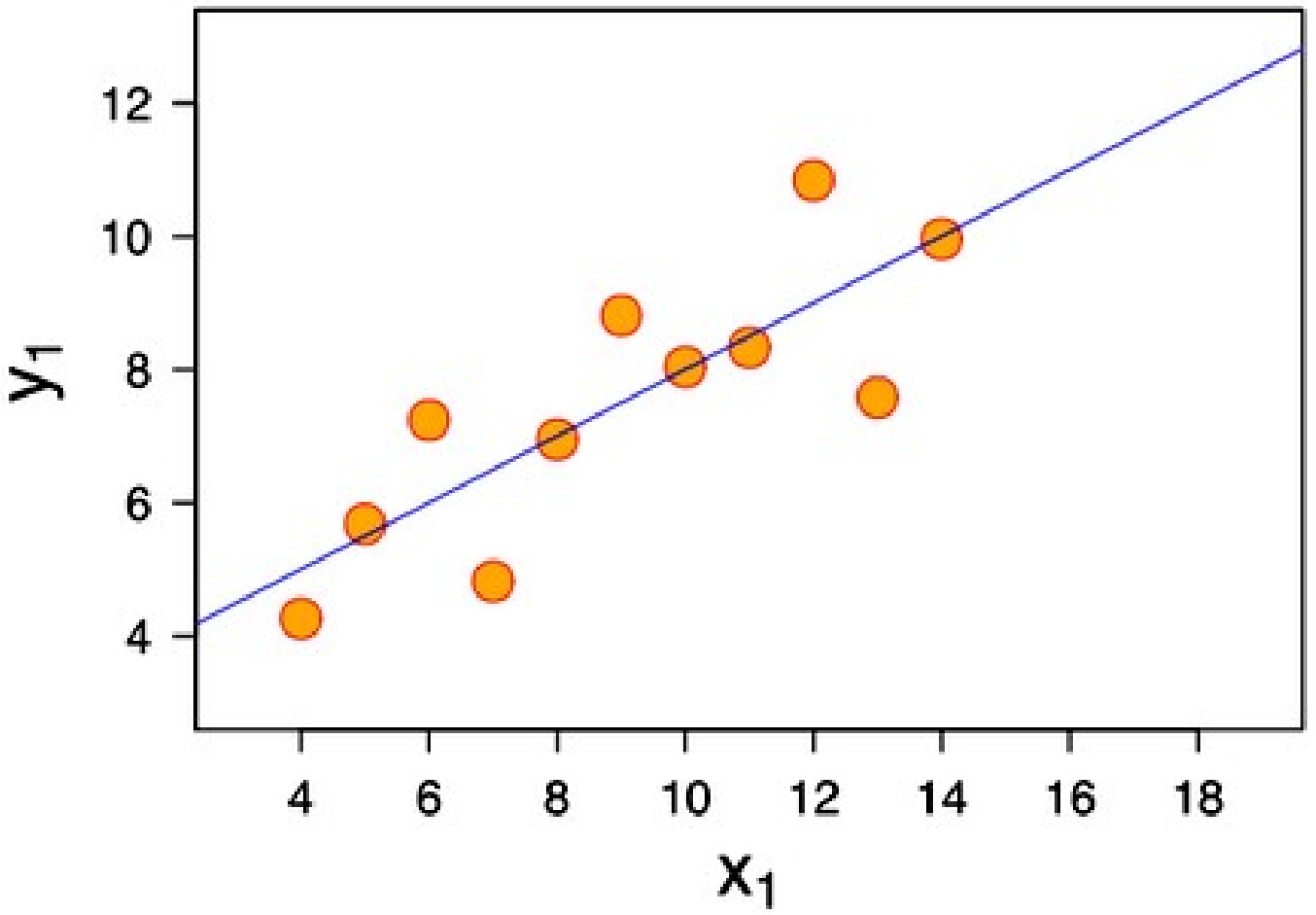
- The four y variables have the same mean (7.5), variance (4.12), correlation (0.816) and regression line (y = 3 + 0.5x).
- However, the distribution of the variables is very different.


12
10
8
6
4
y_1
4 6 8 10 12 14 16 18
x_1


12
10
8
6
4
y_2
4 6 8 10 12 14 16 18
x_2


12
10
8
6
4
y_3
4 6 8 10 12 14 16 18
x_3

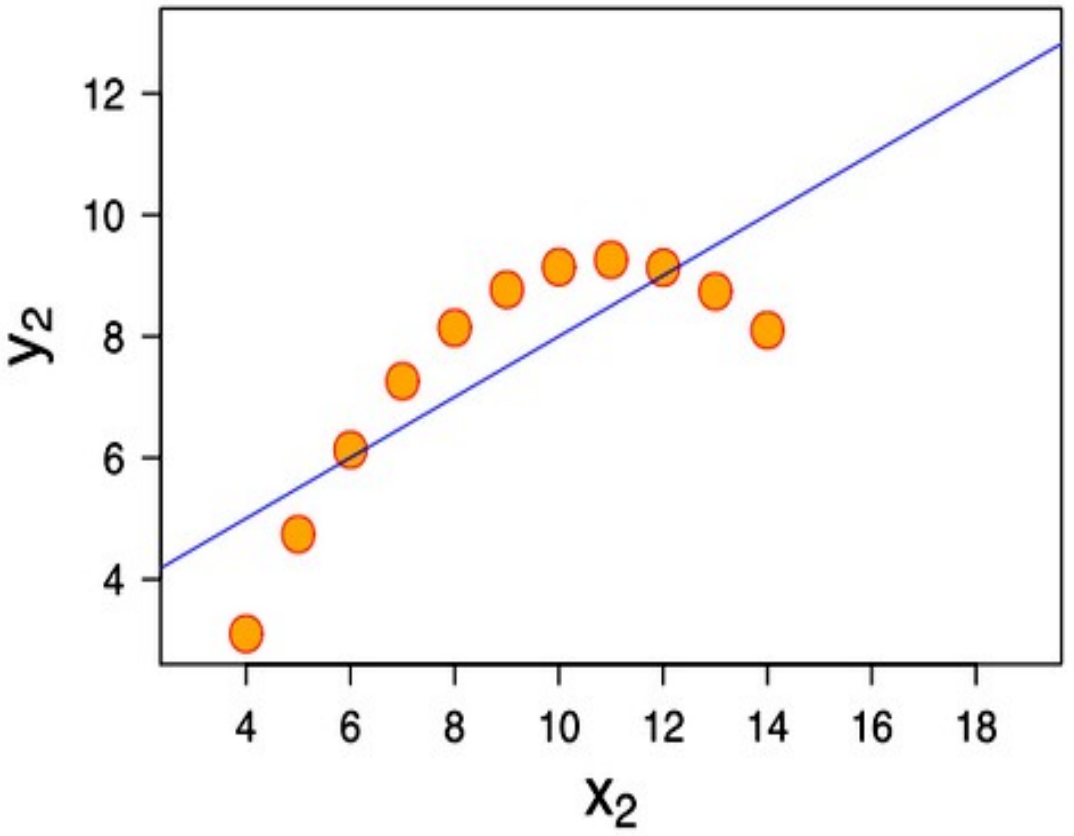

12
10
8
6
4
y_4
4 6 8 10 12 14 16 18
x_4



The first one (top left) seems to be distributed normally, and corresponds to what one would expect when considering two variables correlated and following the assumption of normality.

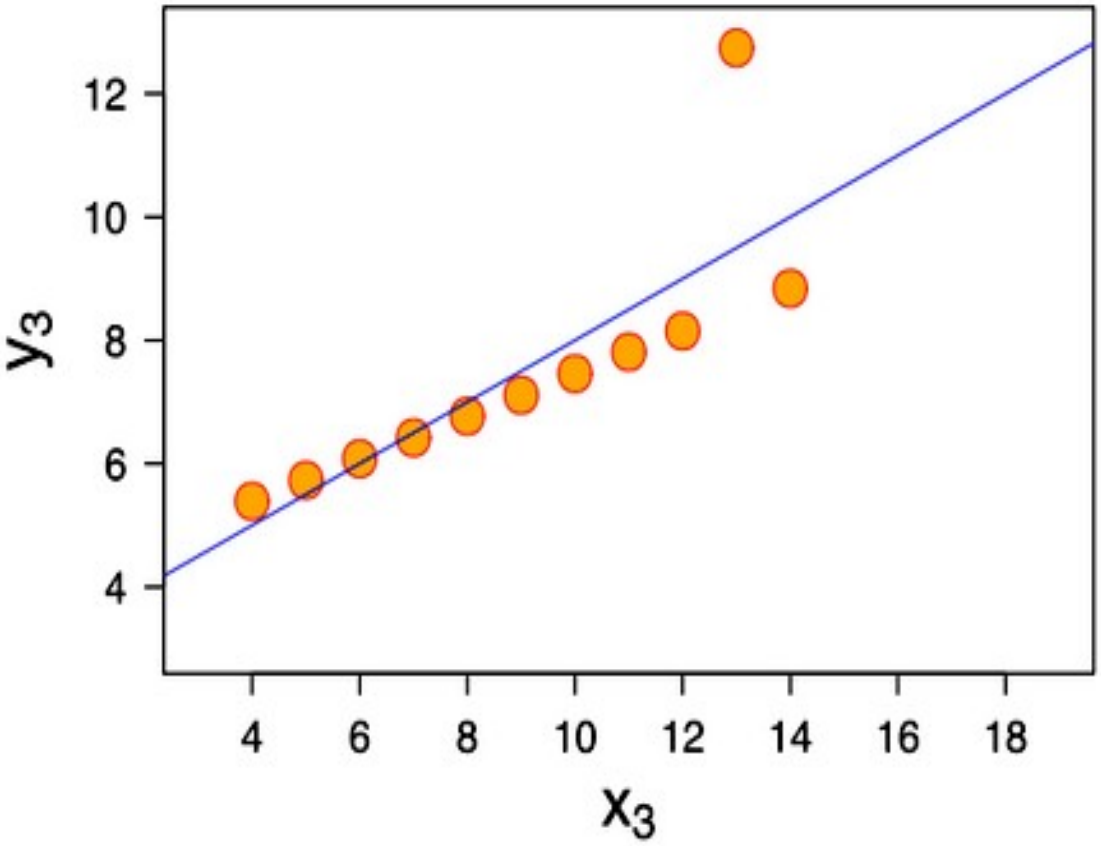
- The second one is not distributed normally; while an obvious relationship between the two variables can be observed, it is not linear.

- In this case the Pearson correlation coefficient does not indicate that there is an exact functional relationship: only the extent to which that relationship can be approximated by a linear relationship.

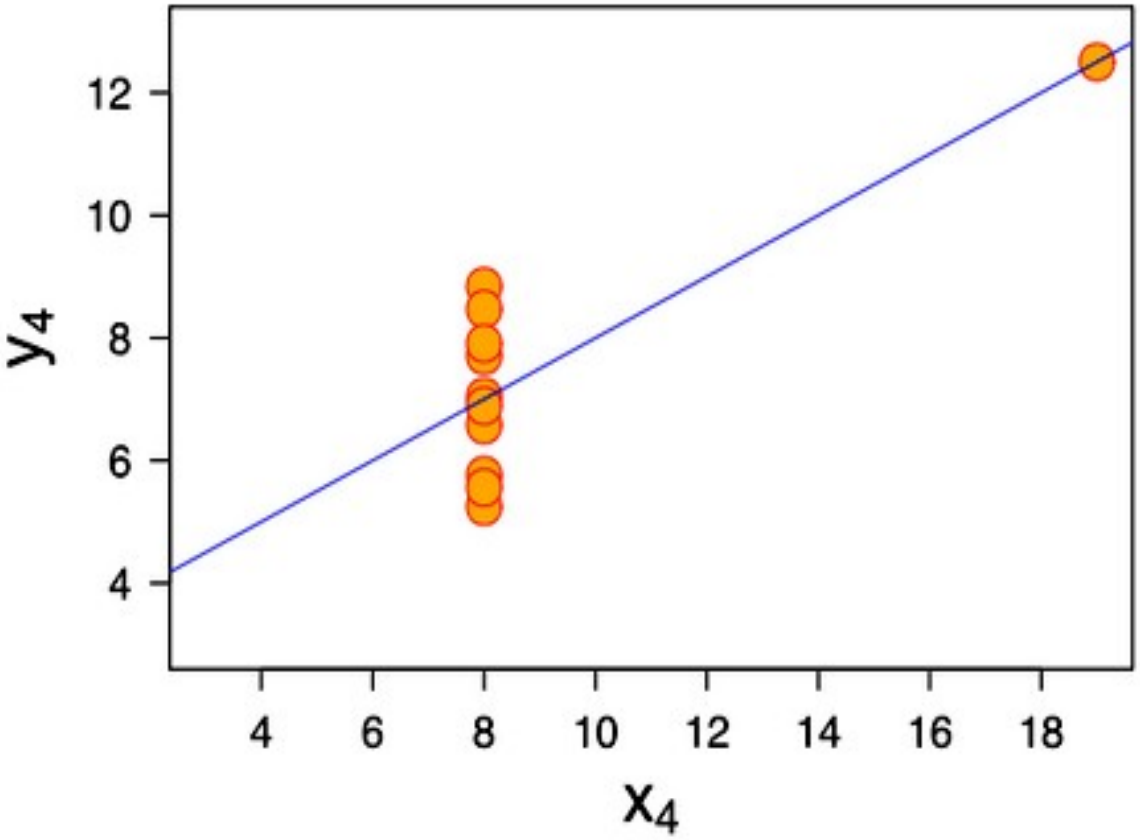


Correlation Coefficient - Misleading when outliers are present

In the third case (bottom left), the linear relationship is perfect, except for one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.



The fourth example (bottom right) shows another example when one outlier is enough to produce a high correlation coefficient, even though the relationship between the two variables is not linear.



Problem 2

- An article presents calculations of the ages (in calendar years before 1950) of several sediment samples taken at various depths (in cm) in Lago di Fimon, a lake in Italy. The results are presented in the following table.

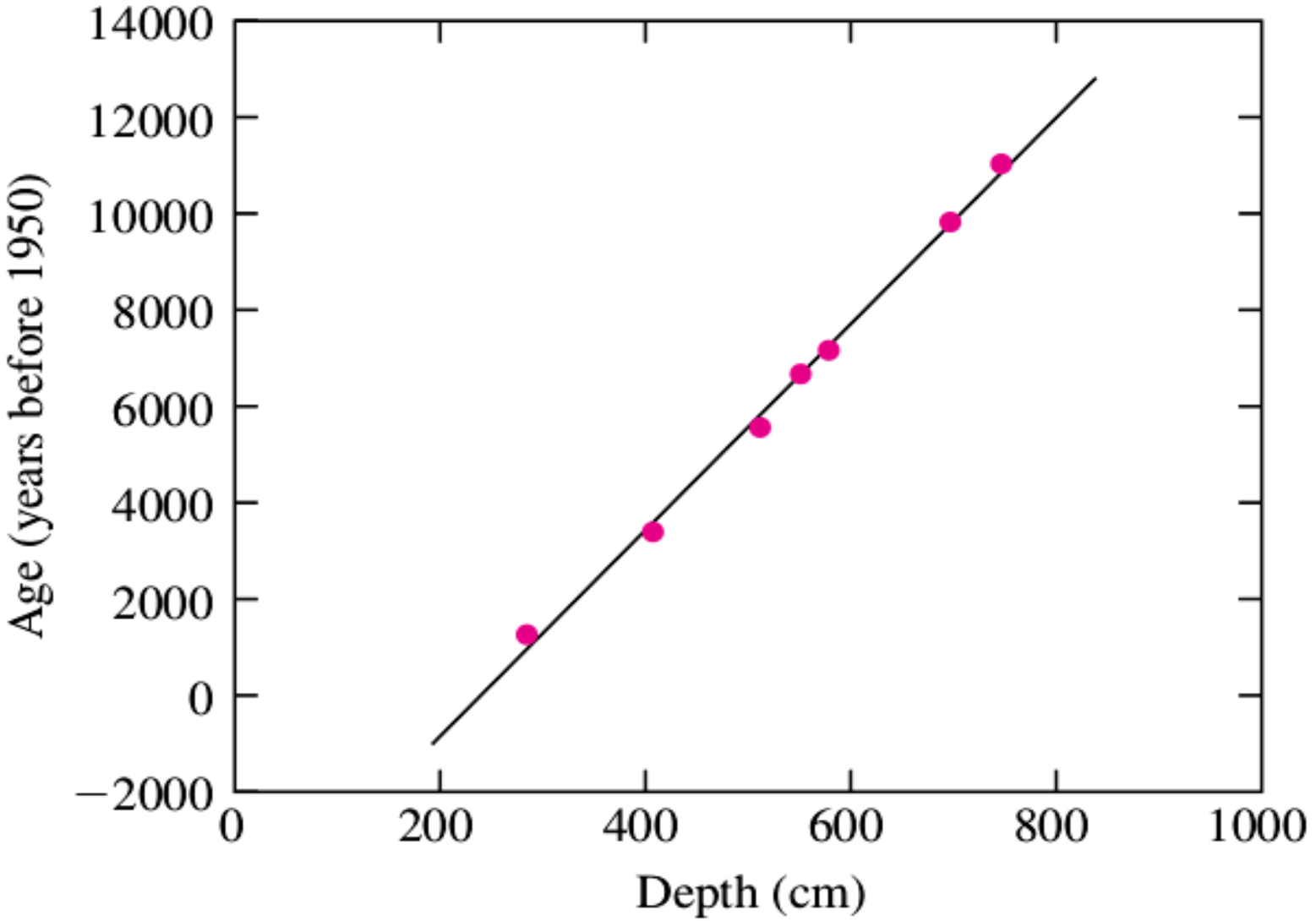
Depth	Age
284.5	1255
407.5	3390
512.0	5560
551.0	6670
578.5	7160
697.0	9820
746.5	11,030

1) Construct a scatterplot of age (y) versus depth (x).

2) Is the correlation coefficient an appropriate summary for these data? Explain why or why not.

3)Compute the correlation between depth(x) and age(y).

1) Construct a scatterplot of age (y) versus depth (x).



Problem 2 : Solution

2) Is the correlation coefficient an appropriate summary for these data? Explain why or why not.

Yes it is, as relationship between the two variables is linear.

3)Compute the correlation between depth(x) and age(y).

X_bar = 539.57 Sx = 159.43

Y_bar = 6412.1 Sy = 3416.1

3)Compute the correlation between depth(x) and age(y).

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$ = 152513.71

Sqrt(152513.71) = 390.53

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$ = 70019492.86

Sqrt(70019492.86) = 8367.77

3)Compute the correlation coefficient

$$\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - n\,\overline{x}\,\overline{y} = 3255209.921$$

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} = 0.996$$

$$= \frac{3255209.921}{390.53 * 8367.77} = 0.996$$

Inference on the Population Correlation

Statistical inference based on Pearson's correlation coefficient often focuses on one of the following two aims:

1) One aim is **to derive a confidence interval for** ρ.

2)The other aim is to **test the null hypothesis** about ρ, based on the value of the sample correlation coefficient r.

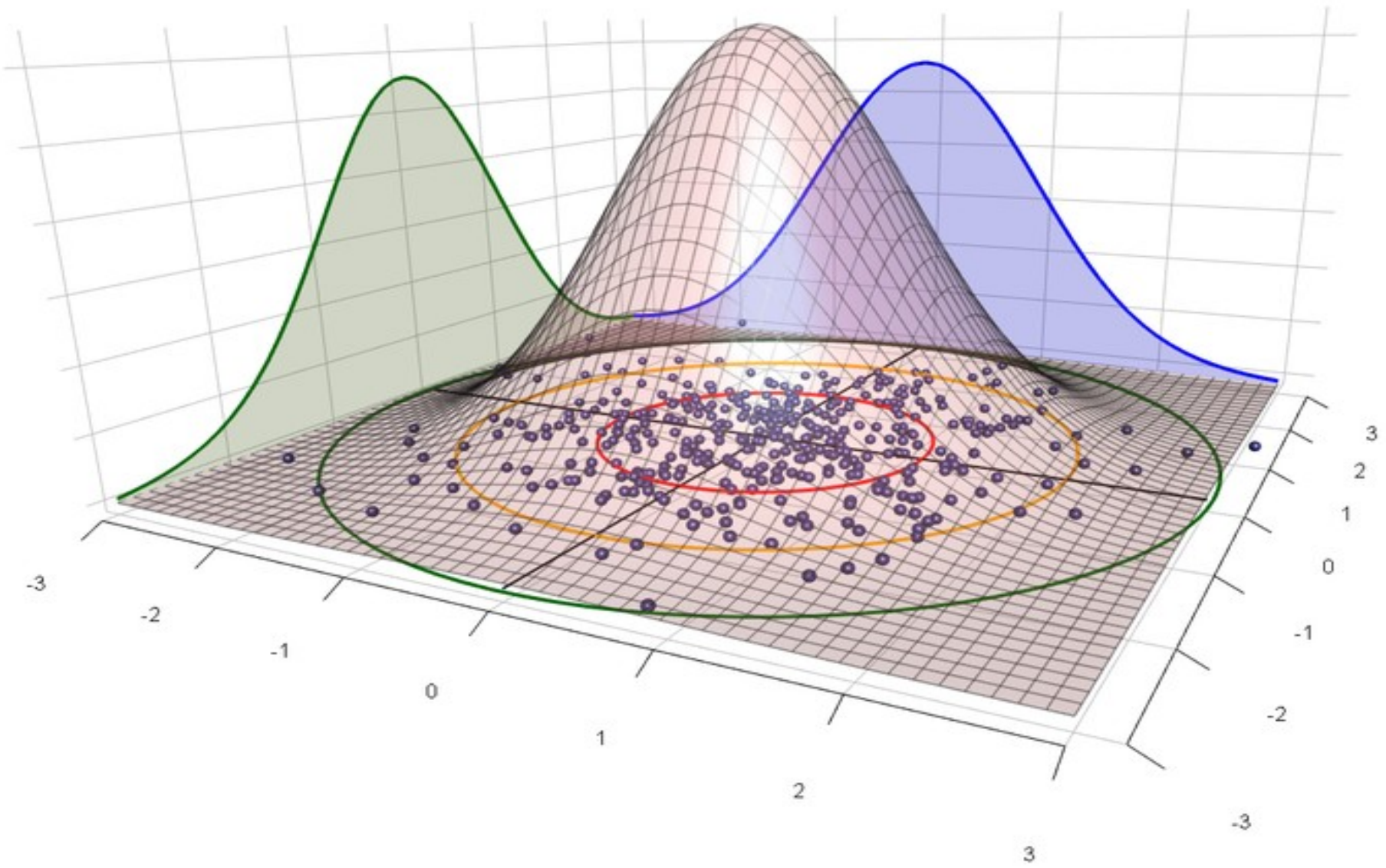
Bivariate Normal Distribution

- Sum of two independent normal random variables is normal.
- However, if the two normal random variables are not independent, then their sum is not necessarily normal.
- Two random variables X and Y are said to be bivariate normal, or jointly normal, if aX+bY has a normal distribution for all a, b $\in$ R.

PDF of Bivariate Normal Distribution

$$f_{XY}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}}\exp\{-\frac{1}{2(1-\rho^2)}\left[x^2 - 2\rho xy + y^2\right]\},$$

where $\rho \in (-1,1)$. If $\rho = 0$, then we just say X and Y have the standard bivariate normal distribution.



Prof. Preet Kanwal

Let X and Y be random variables with the bivariate normal distribution.
Let ρ denote the population correlation between X and Y.
Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a random sample from the joint distribution of X and Y.
Let r be the sample correlation of the n points.
Then the quantity

$$W = \frac{1}{2} \ln \frac{1+r}{1-r}$$

is approximately normally distributed, with mean given by

$$\mu_W = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$$

and variance given by

$$\sigma_W^2 = \frac{1}{n-3}$$

1) Computing confidence interval for ρ

- Find CI for μ_w as:

$$W \pm z_{\alpha/2} * \sigma_w$$

- Use upper and lower confidence bounds of μ_w to find CI for ρ using the following inequality:

$$\rho = \frac{e^{2\mu_w} - 1}{e^{2\mu_w} + 1}$$

Problem 3

- An article presents calculations of the ages (in calendar years before 1950) of several sediment samples taken at various depths (in cm) in Lago di Fimon, a lake in Italy. The results are presented in the following table.

Depth	Age
284.5	1255
407.5	3390
512.0	5560
551.0	6670
578.5	7160
697.0	9820
746.5	11,030

4) Find a 95% confidence interval for the correlation between the two reaction times.**(r = 0.996)**

Computing W

$$W = \frac{1}{2} \ln \frac{1+r}{1-r}$$ = 0.5 ln (1.996 / 0.004) = 3.1

- σ = sqrt(1 / (7 – 3)) = 0.5
- **95% confidence interval for μ_w**

$$W \pm 1.96(0.5) = W \pm 0.98$$

$$3.1 - 1.96(0.5) < \mu_w < 3.1 + 1.96(0.5)$$

$$2.12 < \mu_w < 4.08$$

95% confidence interval for ρ

- To obtain 95% confidence interval for ρ we transform the inequality as:

$$\frac{e^{2(2.12)} - 1}{e^{2(2.12)} + 1} < \frac{e^{2\mu_W} - 1}{e^{2\mu_W} + 1} < \frac{e^{2(4.08)} - 1}{e^{2(4.08)} + 1}$$

$$0.97 < \rho < 0.9994$$

2) Testing Null Hypothesis

- For testing null hypotheses of the form
  - $\rho = \rho_0$
  - $\rho \leq \rho_0$, and
  - $\rho \geq \rho_0$,
- where ρ_0 is a constant not equal to 0, the quantity W forms the basis of a test.

Problem 4

- An article presents calculations of the ages (in calendar years before 1950) of several sediment samples taken at various depths (in cm) in Lago di Fimon, a lake in Italy. The results are presented in the following table.

Depth	Age
284.5	1255
407.5	3390
512.0	5560
551.0	6670
578.5	7160
697.0	9820
746.5	11,030

5) Find the P-value for testing $H_0 : \rho \leq 0.3$ versus $H_1 : \rho > 0.3$. **(r = 0.996)**

Problem 4 : Solution

- Under H 0 we take ρ = 0.3

$$\mu_W = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$$ = 0.3095

- σ = sqrt(1 / (7 – 3)) = 0.5
- It follows that under H_0 ,

$$W \sim N(0.3095, 0.5^2)$$

- The observed value of W is W = 3.1
- Z- score = (3.1 – 0.3095)/0.5 = 5.581
- P= P(Z > 5.581) = 0 <0.05. Hence we can reject H_0

Testing Null Hypothesis ($\rho_0 = 0$)

- For testing null hypotheses of the form
  - $\rho = 0$,
  - $\rho \leq 0$, or
  - $\rho \geq 0$,

  a simpler procedure is available. When $\rho = 0$, the quantity

$$U = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

  has a Student's t distribution with $n - 2$ degrees of freedom.

Problem 5

- An article presents calculations of the ages (in calendar years before 1950) of several sediment samples taken at various depths (in cm) in Lago di Fimon, a lake in Italy. The results are presented in the following table.

Depth	Age
284.5	1255
407.5	3390
512.0	5560
551.0	6670
578.5	7160
697.0	9820
746.5	11,030

5) Test the hypothesis $H_0 : \rho \leq 0$ versus $H_1 : \rho > 0$.

(r = 0.996)

Problem 5 : Solution

$$U = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

- P-value will be approx 0.

$$= \frac{0.996\ \sqrt{\quad 7-2}}{\sqrt{1-\quad 0.996^2}}$$

$$= \quad 24.92$$

Problem 6

- Tire pressure (in kPa) was measured for the right and left front tires on a sample of 10 automobiles.

Right Tire Pressure	Left Tire Pressure
184	185
206	203
193	200
227	213
193	196
218	221
213	216
194	198
178	180
207	210

$r = 0.930698$.

a. Find a 95% confidence interval for ρ.

b. Can you conclude that $\rho > 0.9$?

c. Can you conclude that $\rho > 0$?

Computing W

$$W = \frac{1}{2} \ln \frac{1+r}{1-r}$$ = 0.5 ln (1.931 / 0.069) = 1.67

- σ = sqrt(1 / (10 – 3)) = 0.378
- **95% confidence interval for μ_w**

$$W \pm 1.96(0.378) = W \pm 0.741$$

$$1.67 - 1.96(0.378) < \mu_w < 1.67 + 1.96(0.378)$$

$$0.929 < \mu_w < 2.411$$

95% confidence interval for ρ

- To obtain 95% confidence interval for ρ we transform the inequality as:

$$\frac{e^{2(0.929)} - 1}{e^{2(0.929)} + 1} < \frac{e^{2\mu_w} - 1}{e^{2\mu_w} + 1} < \frac{e^{2(2.411)} - 1}{e^{2(2.411)} + 1}$$

$$0.73 < \rho < 0.984$$

Can you conclude that ρ > 0.9?

- Under H 0 we take ρ = 0.9

$$\mu_W = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$$ = 1.47

- σ = sqrt(1 / (10 – 3)) = 0.378
- It follows that under H_0 ,

$$W \sim N(1.47, 0.378^2)$$

- The observed value of W is = 1.67
- Z- score = (1.67 – 1.47)/0.378 = 0.53
- P= P(Z > 0.53) = 0.2981 > 0.05. Hence we cannot reject H_0

c. Can vou conclude that ρ > 0?

$$U = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$= \frac{0.931\ \sqrt{10-2}}{\sqrt{1-\quad 0.931^2}}$$

$$= \quad 7.214$$

- Under H 0 , U has a Student's t distribution with 10 − 2 = 8 degrees of freedom.
- From the t table,

  P < 0.001.
- We conclude that ρ > 0 and reject H0

Homework question

- An article evaluates the use of tree ring widths to estimate changes in the masses of glaciers.
- For the Sentinel glacier, the net mass balance (change in mass between the end of one summer and the end of the next summer) was measured for 23 years.
- During the same time period, the tree ring index for white bark pine trees was measured, and the sample correlation between net mass balance and tree ring index was $r = -0.509$.
- Can youconclude that the population correlation ρ differs from 0?

Regression Analysis – Simple Linear Regression

Prof. Preet Kanwal

Introduction

- Regression analysis is an important tool for modelling and analyzing data.

- Regressions are used to quantify the relationship between one variable(dependent variable) and the other variables(independent variables) that are thought to explain it.

- Here, we fit a curve / line to the data points, in such a manner that the differences between the distances of data points from the curve or line is minimized.

Benefits of Using Regression Analysis

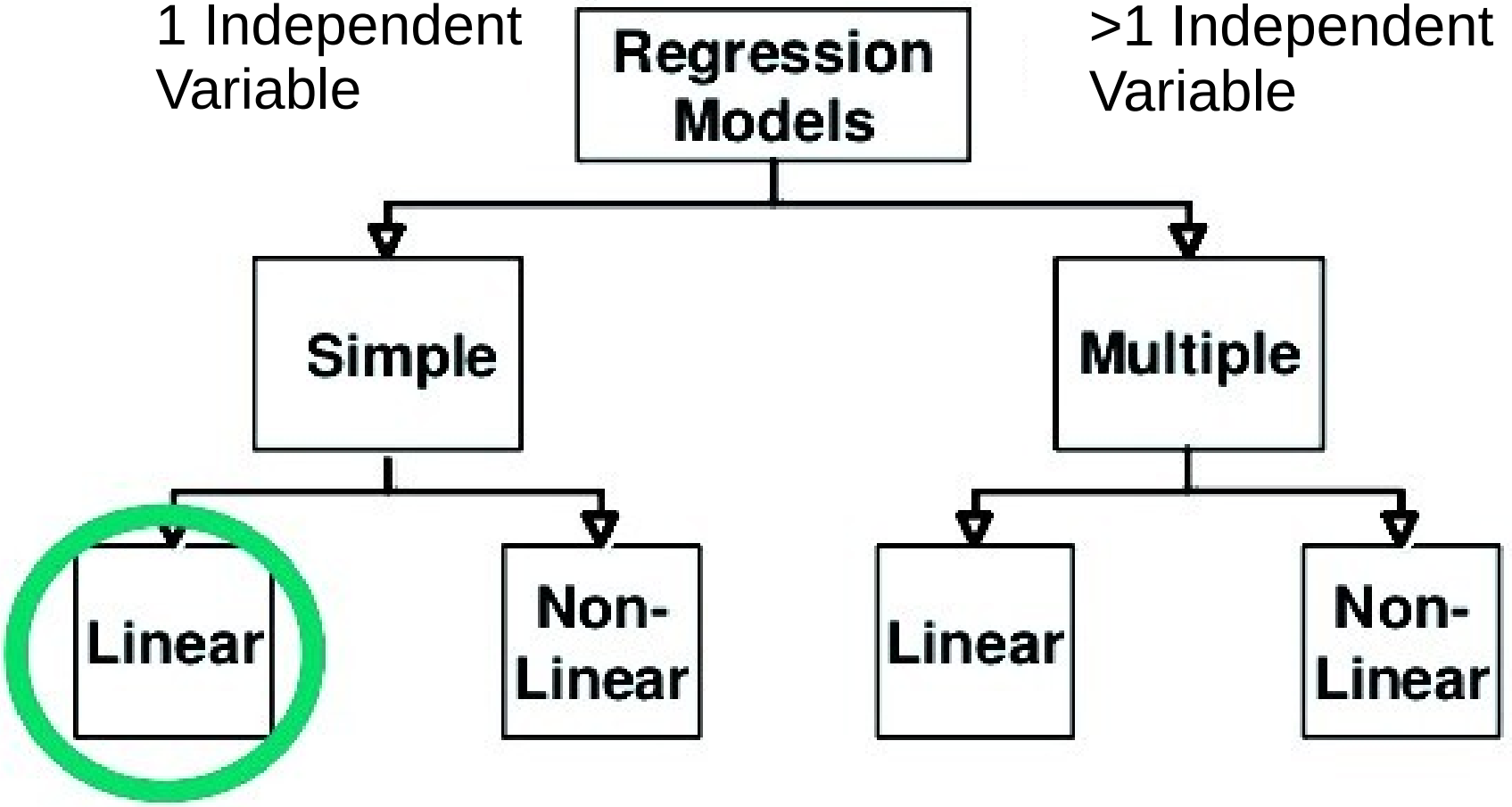
- It indicates the significant relationships between dependent variable and independent variable.

- It indicates the strength of impact of (multiple) independent variable(s) on a dependent variable.

Types of Regression Analysis

- There are various kinds of regression techniques available to make predictions.
- These techniques are mostly driven by three metrics :
  - Number of independent variables
  - Type of dependent variables(Continuous or Binary)
  - Shape of regression line.

Types of Regression Analysis



Types of Regression Analysis

1) Linear Regression :

- In this technique, the dependent variable is continuous, independent variable(s) can be continuous or discrete, and nature of regression line is linear.
- Power of independent variable = 1.
- Linear equation has one basic form.

2) Non Linear Regression :

- If the variables do not share a linear relationship it can be quantified using Non Linear Regression.
- Power of independent variable > 1
- Nonlinear equations can take many different forms.

Types of Linear Regression

1) Simple Linear Regression :

There is only one independent and one dependent variable.

$$Y = a + bX$$

- Here, X : Independent Variable and Y: dependent variable.

2) Multiple Linear Regression:

There are multiple independent and one dependent variable.

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + ...+ b_nX_n$$

- Here, X1, X2, ... Xn are independent variables which affect Y(dependent variable).

Points to Ponder

- There must be linear relationship between independent and dependent variables

- Linear Regression is very sensitive to Outliers. It can terribly affect the regression line and eventually the forecasted values.

Applications of Regression Analysis

The two primary uses for regression in business are forecasting and optimization.

1) Forecasting :

- Helps managers predict future demand for their products. Demand analysis, for example, predicts how many units consumers will purchase.
- Helps fine-tune manufacturing and delivery processes.

2) Optimization of business processes:

- A factory manager might, for example, build a model to understand the relationship between oven temperature and the shelf life of the cookies baked in those ovens.
- A company operating a call center may wish to know the relationship between wait times of callers and number of complaints.Today, managers considers regression an indispensable tool.

Simple Linear Regression

- The scatterplot tends to be clustered around a line known as the least-squares line.
- Linear Model for two variables:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

y_i - Dependent Variable

x_i - Independent Variable

β_0, β - Regression Coefficients

ε_i - Error

Example

TABLE 7.1 Measured lengths of a spring under various loads

Weight (lb) x	Measured Length (in.) y	Weight (lb) x	Measured Length (in.) y
0.0	5.06	2.0	5.40
0.2	5.01	2.2	5.57
0.4	5.12	2.4	5.47
0.6	5.13	2.6	5.53
0.8	5.14	2.8	5.61
1.0	5.16	3.0	5.59
1.2	5.25	3.2	5.61
1.4	5.19	3.4	5.75
1.6	5.24	3.6	5.68
1.8	5.46	3.8	5.80

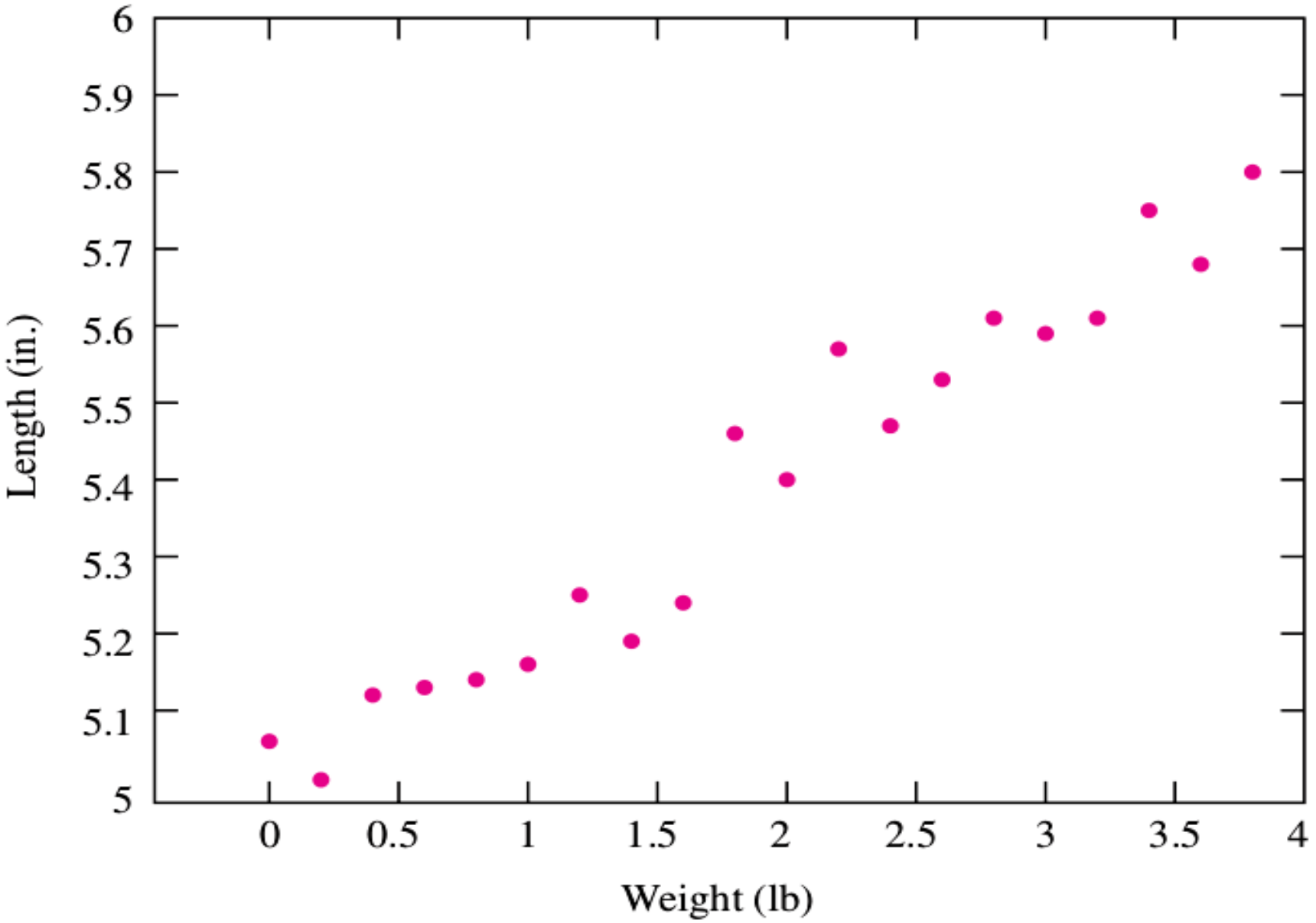


FIGURE 7.9 Plot of measured lengths of a spring versus load.

Measurement error (ε_i)

- If there were no measurement error, the points would lie on a straight line with slope β 1 and intercept β 0 , and these quantities would be easy to determine.

- Because of measurement error, β 0 and β 1 cannot be determined exactly, but they can be estimated by calculating the **least-squares line.**

ε_i : accumulation of error from many sources

- In practice, ε_i represents the accumulation of error from many sources.
- In an experminent about checking the elasiticity of a spring, ε_i can be affected by errors in:
  - measuring the length of the spring,
  - errors in measuring the weights of the loads placed on the spring,
  - variations in the elasticity of the spring due to changes in ambient temperature or metal fatigue, and so on.

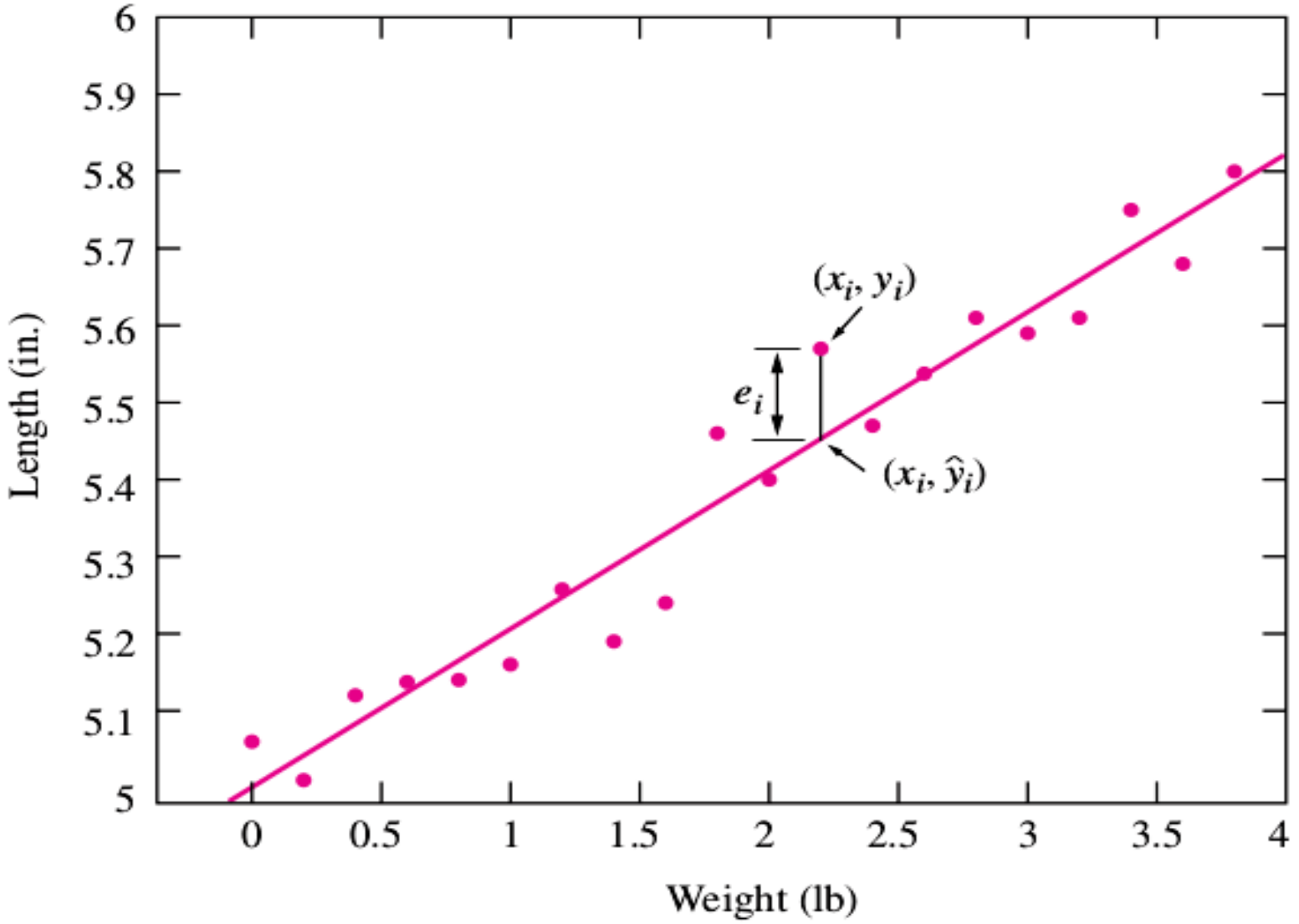
Equation of least-squares line

Equation of least-squares line

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

$\hat{\beta}_0$ and $\hat{\beta}_1$ are called least-squares coefficients, estimates of β 0 and β 1

The least-squares line superimposed



Residuals

Fitted value $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Residual = Observed – Predicted

$$e_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)$$

The least-squares line is the line that minimizes the sum of the squared residuals, i.e.,

$$\sum_{i=1}^{n} e_i^2 \text{ is minimized}$$

Computing the Equation of the Least-Squares Line

$$e_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)$$

Therefore $\hat{\beta}_0$ and $\hat{\beta}_1$ are the quantities that minimize the sum

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

These quantities are

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Problem

- An article presents calculations of the ages (in calendar years before 1950) of several sediment samples taken at various depths (in cm) in Lago di Fimon, a lake in Italy. The results are presented in the following table.

Depth	Age
284.5	1255
407.5	3390
512.0	5560
551.0	6670
578.5	7160
697.0	9820
746.5	11,030

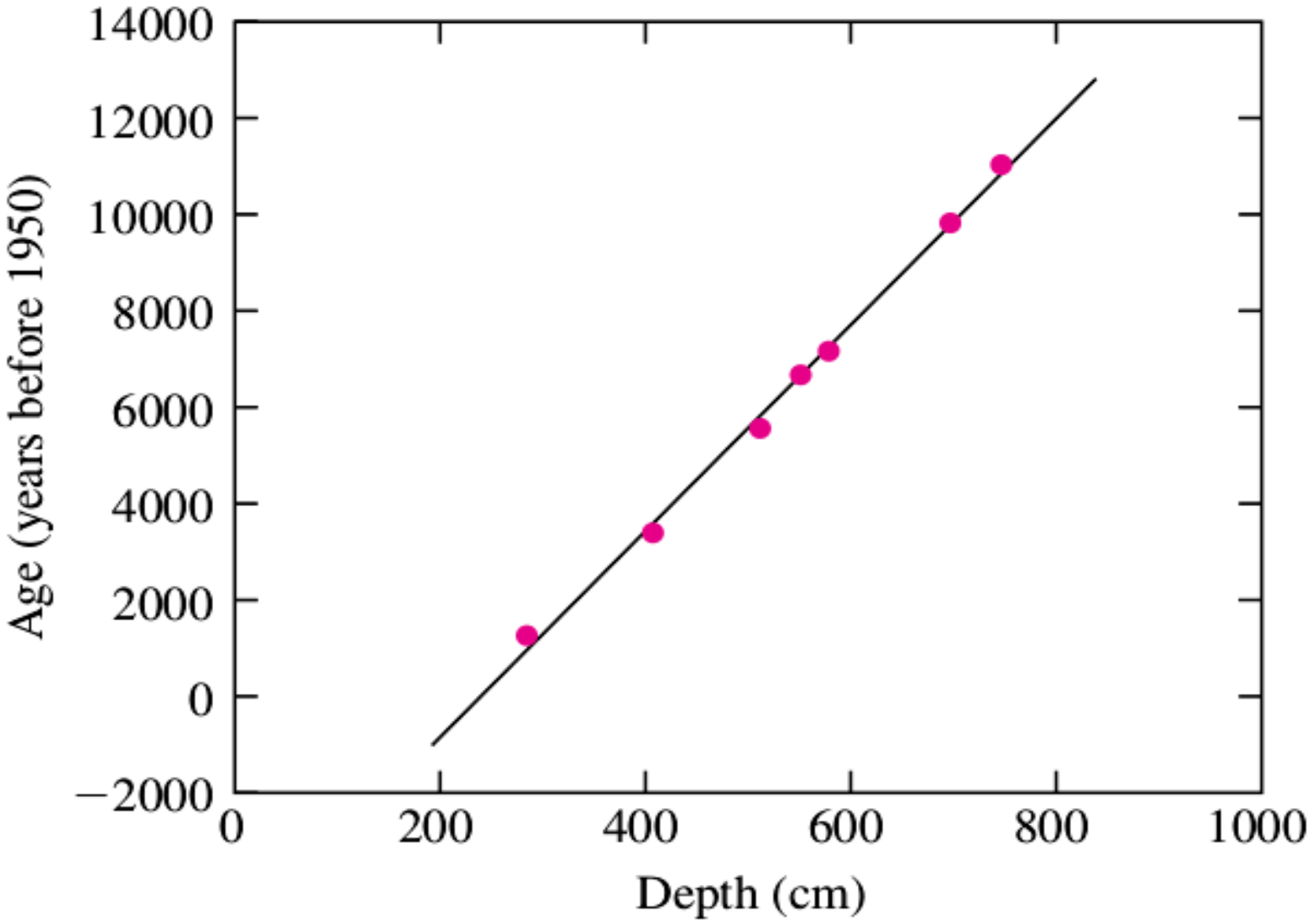
1) Compute the least-squares line for predicting age from depth.

2) If two samples differ by 100 cm in depth, by how much would you predict their ages to differ?

3) Predict the age for a specimen whose depth is 600 cm.

4) For what depth would you predict an age of 5000?

Scatterplot of age (y) versus depth (x).



Solution

1) $y = -5129.11+21.3924x$

2) $y_1 - y_2 = -5129.11 + 21.39\ x_1 - (-5129.11 + 21.39x_2)$

- $y_1 - y_2 = 21.39\ (x_1 - x_2)$
- $y_1 - y_2 = 21.39 * 100$
- $y_1 - y_2 = 2139$

3) $y = -5129.11 + 21.39\ (600) = 7704.89$

4) $5000 = -5129.11 + 21.39\ x$

- $x = 473.544$

Regression using Python

- numpy.polyfit:
  - For Least-Squares Polynomial Fit
  - For fitting a linear model, use deg = 1

- Demo

Points to Ponder

- Don't Extrapolate Outside the Range of the Data :
  - We cannot predict y when x lies outside the range $(x_{max} - x_{min})$
  - No outliers.

- Don't Use the Least-Squares Line When the Data Aren't Linear

Problem

- An article presents calculations of the ages (in calendar years before 1950) of several sediment samples taken at various depths (in cm) in Lago di Fimon, a lake in Italy. The results are presented in the following table.

Depth	Age
284.5	1255
407.5	3390
512.0	5560
551.0	6670
578.5	7160
697.0	9820
746.5	11,030

1) Should the least-squares line be used to predict the age for a depth of 50 cm? If so, predict the age. If not, explain why not.

Solution

The least-squares line should not be used to predict the age for a depth of 50 cm as It is outside the range of the data (Outlier)

Interpreting the slope of the least-squares line in terms of the correlation coefficient

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} = r$$

Multiply both sides by

$$\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}/\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2} = s_y/s_x$$

we get,

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = r\frac{s_y}{s_x} \quad \Rightarrow \quad \hat{\beta}_1 = r\frac{s_y}{s_x}$$

Another form of least-squares line

Substituting $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ in least-squares line

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

we get,

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

=> $$\hat{y} - \bar{y} = r \frac{s_y}{s_x} (x - \bar{x})$$

Note

Thus the least-squares line is the line that passes through the center of mass of the scatterplot $(\overline{x}, \overline{y})$, with slope

$$\hat{\beta}_1 = r \frac{s_y}{s_x}$$

Problem

In a study relating the degree of warping, in mm, of a copper plate (y) to temperature in ◦ C (x), the following summary statistics were calculated: n = 40,

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = 98,775$$

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = 19.10,$$

$$\overline{x} = 26.36, \overline{y} = 0.5188,$$

$$\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = 826.94.$$

Problem

a. Compute the correlation r between the degree of warping and the temperature.

b. Compute the least-squares line for predicting warping from temperature.

c. Predict the warping at a temperature of 40 ◦C.

d. At what temperature will we predict the warping to be 0.5 mm?

a. Compute the correlation r between the degree of warping and the temperature.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = 0.602052.$$

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 826.94.$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 98,775$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = 19.10,$$

b. Compute the least-squares line for predicting warping from temperature.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 0.008371956$$

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 826.94.$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 98,775$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 0.298115.$$

$$\bar{x} = 26.36, \bar{y} = 0.5188,$$

Least-squares line : $y = 0.298115 + 0.008371956x$

c. Predict the warping at a temperature of 40 ◦C.

Least-squares line :

$$y = 0.298115 + 0.008371956x$$

$$0.298115 + 0.008371956(40) = 0.633 \text{ mm}$$

d. At what temperature will we predict the warping to be 0.5 mm?

Least-squares line :

$$y = 0.298115 + 0.008371956x$$

y = 0.5mm

Let x be the required temperature.

$$0.5 = 0.298115 + 0.008371956x$$

=> x = 24.1◦ C

Note

Linear regression calculates an equation that minimizes the distance between the fitted line and all of the data points. Technically, it minimizes the sum of the squared residuals.

That is why the name least-squares line.

In general, a model fits the data well if the differences between the observed values and the model's predicted values are small and unbiased.

How well a model explains a given set of data?

Measuring Goodness-of-Fit

- A goodness-of-fit statistic is a quantity that measures how well a model explains a given set of data.
- The correlation coefficient r measures the strength of the linear relationship between x and y.
- Therefore r is a goodness-of-fit statistic for the linear model.
- Visual Examination of goodness-of-fit is done using Residual Plots that can indicate biased results more effectively than statistical measures.

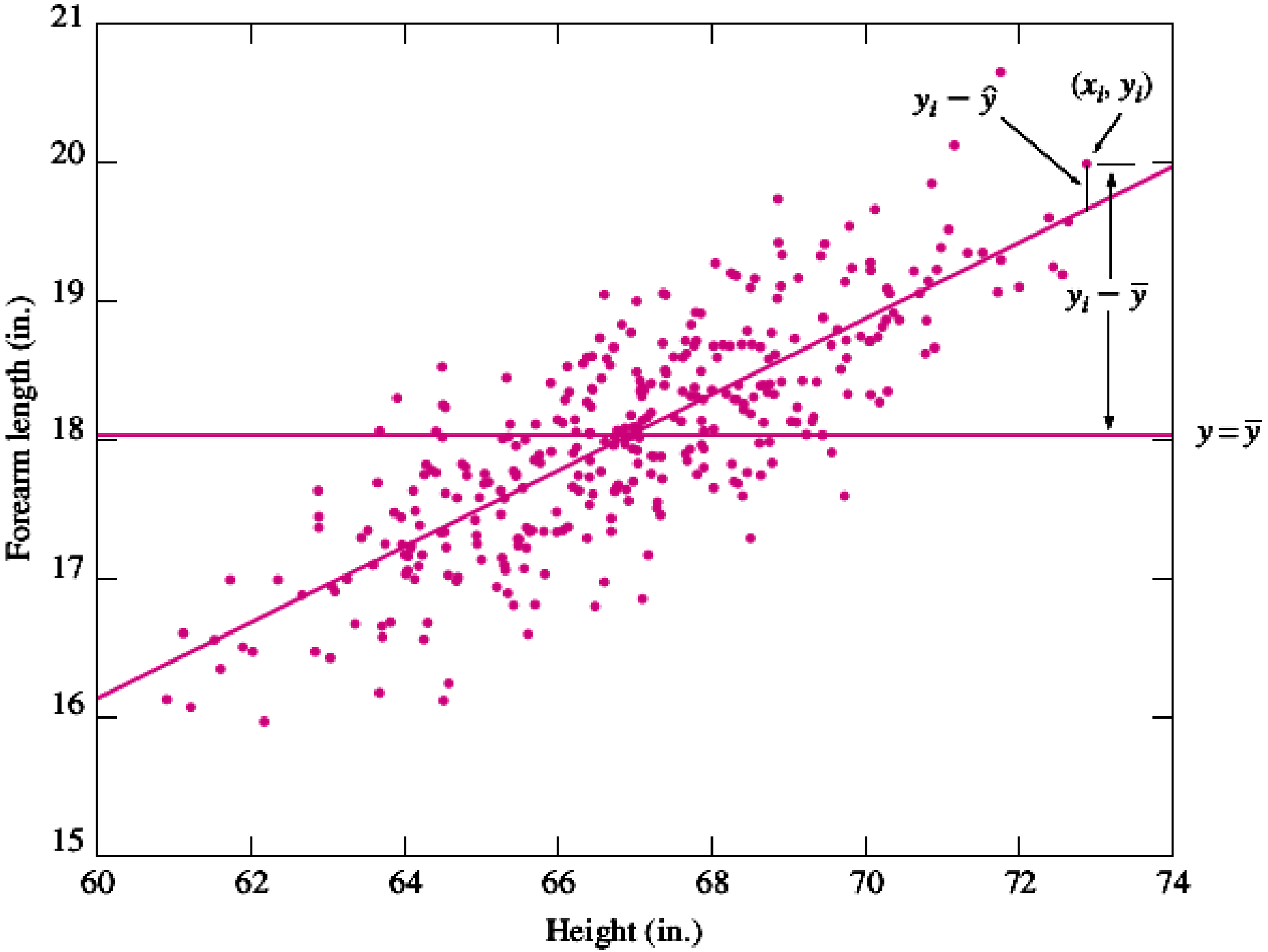


FIGURE 7.12 Heights and forearm lengths of men. The least-squares line and the horizontal line $y = \bar{y}$ are superimposed.

Goodness-of-fit statistic

- The strength of the linear relationship can be measured by computing the reduction in sum of squared prediction errors obtained by using $\hat{y}_i$ rather than $\bar{y}$
- $\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is a **goodness-of-fit statistic.**
- The bigger this difference is, the more tightly clustered the points are around the least-squares line and the stronger the linear relationship is between x and y.

Goodness-of-fit statistic – has units

- The quantity:

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

  has units, namely the squared units of y.
- We could not use this statistic to compare the goodness-of-fit of two models fit to different data sets, since the units would be different.
- To measure goodness-of-fit on an absolute scale, we use a different statistic which is unitless, based on r.

Coefficient of Determination - Unitless

$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ the **error sum of squares**

$\sum_{i=1}^{n}(y_i - \bar{y})^2$ the **total sum of squares.**

$\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ the **regression sum of squares.**

$$\text{Total sum of squares} = \text{Regression sum of squares} + \text{Error sum of squares}$$

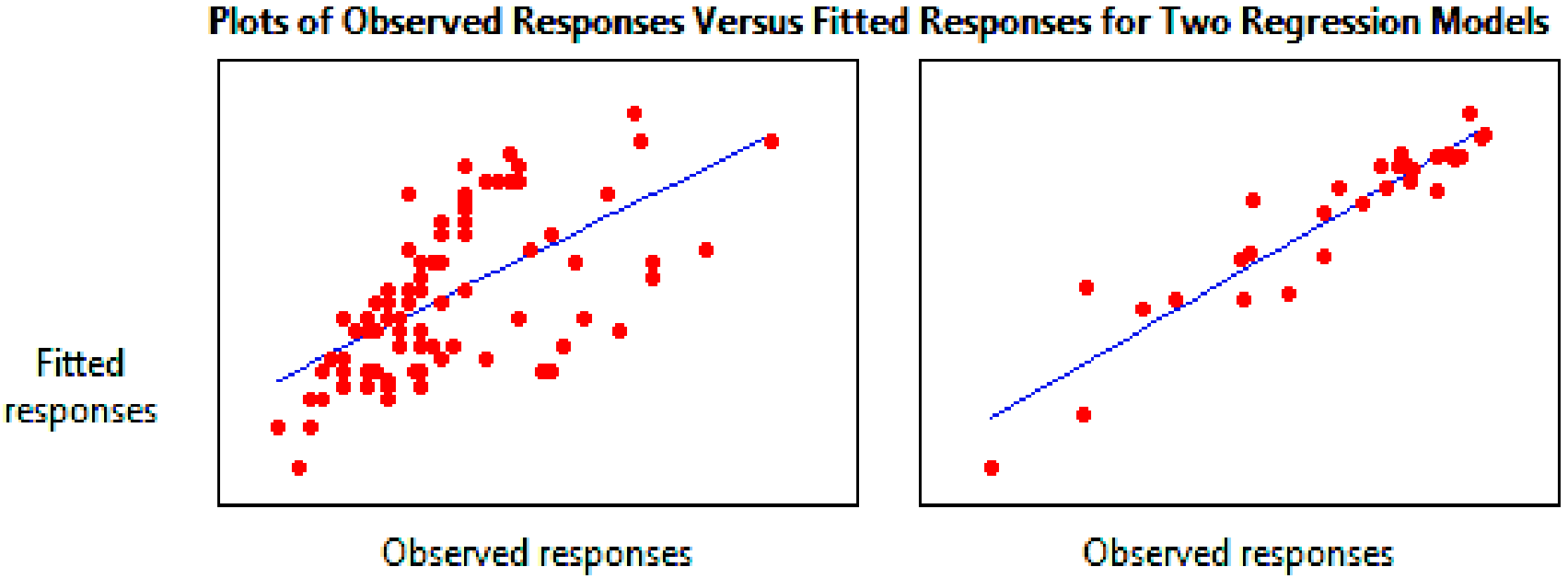
$$r^2 = \frac{\text{Regression sum of squares}}{\text{Total sum of squares}}$$

It measures the proportion of variation in the dependent variable that can be attributed to the independent variable.

r^2

- **r^2** is a statistical measure of how close the observed data are to the fitted regression line.
- **It is the percentage of the response variable variation that is explained by a linear model.**
- **r^2** is always between 0 and 100%:
  - 0% indicates that the model explains none of the variability of the response data around its mean.
  - 100% indicates that the model explains all the variability of the response data around its mean.
- In general, the higher the **r^2**, the better the model fits your data. However, **there are important conditions for this guideline.**

Graphical Representation of r^2



- Plotting fitted values by observed values graphically illustrates different **r²** values for regression models.
- The regression model on the left accounts for 38.0% of the variance while the one on the right accounts for 87.4%.

Key Limitations of R-squared

- **r^2** cannot determine whether the coefficient estimates and predictions are biased, which is why you must assess the residual plots.
- **Low R-squared values are not always bad and high R-squared values are not always good!**
- **r^2** does not indicate whether a regression model is adequate. You can have a low R-squared value for a good model, or a high R-squared value for a model that does not fit the data!

Compute the error sum of squares, the regression sum of squares, and the total sum of squares.

In a study relating the degree of warping, in mm, of a copper plate (y) to temperature in ◦ C (x), the following summary statistics were calculated: n = 40,

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = 98{,}775$$

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = 19.10,$$

$$\overline{x} = 26.36,\ \overline{y} = 0.5188,$$

$$\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = 826.94.$$

Solution

$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ the **error sum of squares**

$\sum_{i=1}^{n}(y_i - \bar{y})^2$ the **total sum of squares.** = 19.10

the **regression sum of squares.** = Total sum of squares – error sum of squares

$$r^2 = \frac{\text{Regression sum of squares}}{\text{Total sum of squares}}$$ = 0.602052^2

0.602052^2 = (19.10 - error sum of squares)/ 19.10

=> error sum of squares = 19.10 - (0.3625 * 19.10) = 12.177

=> Regression sum of squares = 19.10 – 12.177. = 6.923

Uncertainties in the Least-Squares Coefficients

Prof. Preet Kanwal

Uncertainties in the Least-Squares Coefficients

- Each time the process is repeated, the values of ε_i, and thus the values of $\hat{\beta}_0$ and $\hat{\beta}_1$, will be different.
- In other words, ε_i, $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables.
- if the ε_i tend to be small in magnitude, the points will be tightly clustered around the line, and the uncertainty in the least-squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ will be small.

Assumptions for ε_i

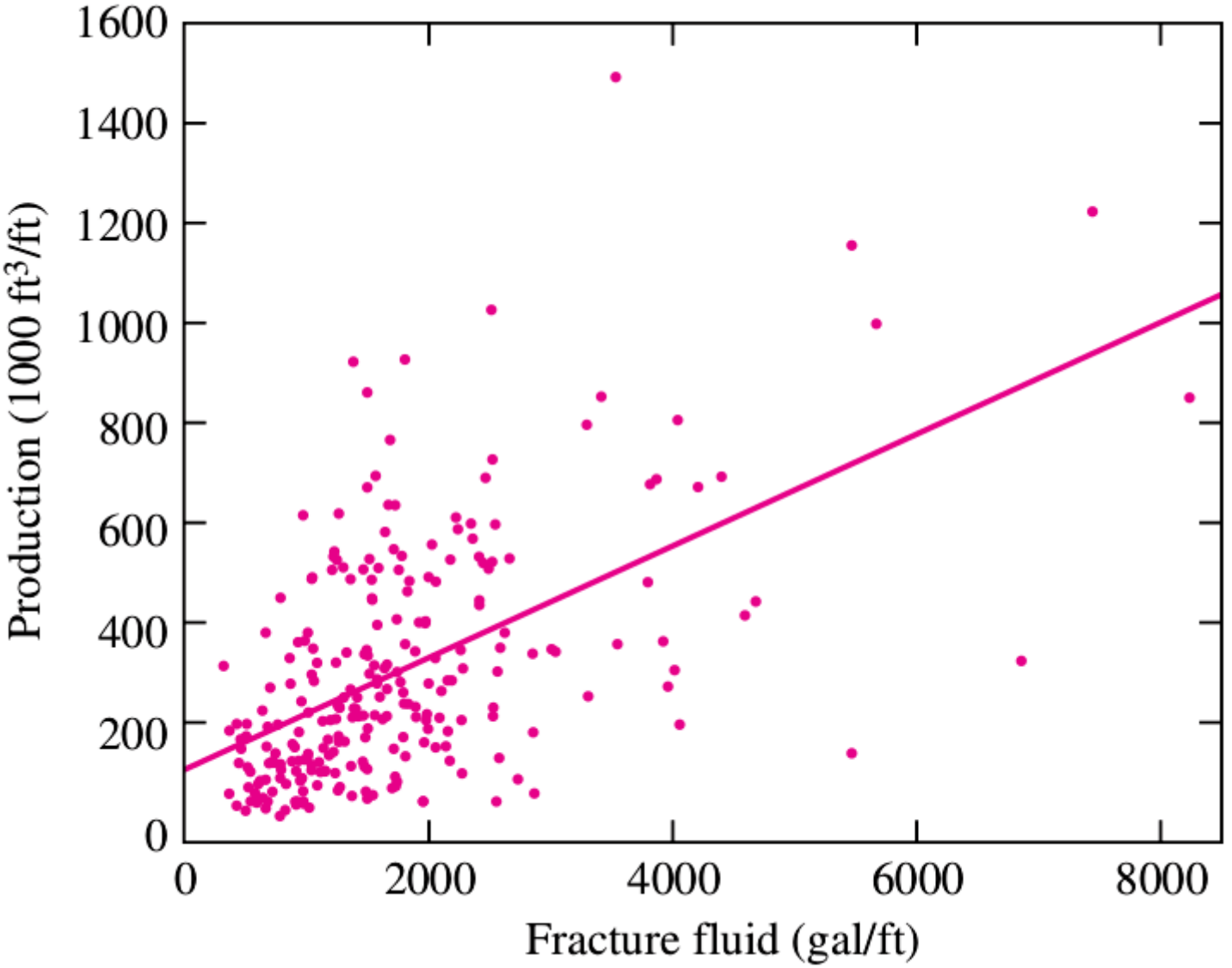
- In order for the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ to be useful, we need to estimate how large their uncertainties are.
- In order to do this, we need to know something about the nature of the errors ε_i .

Assumptions for Errors in Linear Models

In the simplest situation, the following assumptions are satisfied:

1. The errors $\varepsilon_1, \ldots, \varepsilon_n$ are random and independent. In particular, the magnitude of any error ε_i does not influence the value of the next error ε_{i+1}.
2. The errors $\varepsilon_1, \ldots, \varepsilon_n$ all have mean 0.
3. The errors $\varepsilon_1, \ldots, \varepsilon_n$ all have the same variance, which we denote by σ^2.
4. The errors $\varepsilon_1, \ldots, \varepsilon_n$ are normally distributed.

What feature of the scatterplot indicates that assumption 3 is violated?



There is a greater amount of vertical spread on the right side of the plot than on the left.

Magnitude of the variance σ^2

- Under these assumptions, the effect of the ε_i is largely governed by the magnitude of the variance σ^2 , since it is this variance that determines how large the errors are likely to be.

- Since the magnitude of the variance is reflected in the **degree of spread of the points around the least-squares line**, it follows that by measuring this spread, we can estimate the variance.

s^2 : estimate of the error variance σ^2

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

or

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2} = \frac{(1-r^2)\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-2}$$

Mean and Variance of y_i

In the linear model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, under assumptions 1 through 4, the observations $y_1, \ldots, y_n$ are independent random variables that follow the normal distribution. The mean and variance of y_i are given by

$$\mu_{y_i} = \beta_0 + \beta_1 x_i$$

$$\sigma^2_{y_i} = \sigma^2$$

The slope β_1 represents the change in the mean of y associated with an increase of one unit in the value of x.

$$\hat{\beta}_0 \sim N(\beta_0, s_{\hat{\beta}_0})$$

$$\hat{\beta}_1 \sim N(\beta_1, s_{\hat{\beta}_1})$$

Under assumptions 1 through 4

- The quantities $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed random variables.
- The means of $\hat{\beta}_0$ and $\hat{\beta}_1$ are the true values β_0 and β_1, respectively.
- The standard deviations of $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimated with

$$s_{\hat{\beta}_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

and

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

where $s = \sqrt{\frac{(1 - r^2)\sum_{i=1}^{n}(y_i - \bar{y})^2}{n - 2}}$ is an estimate of the error standard deviation σ.

More Spread in the x Values, the Better

When one is able to choose the x values, it is best to spread them out widely. The more spread out the x values, the smaller the uncertainties in $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

Specifically, the uncertainty $\sigma_{\widehat{\beta}_1}$ in $\widehat{\beta}_1$ is inversely proportional to $\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}$, or equivalently, to the sample standard deviation of $x_1, x_2, \ldots, x_n$.

Caution: If the range of x values extends beyond the range where the linear model holds, the results will not be valid.

Inferences on the Slope and Intercept

- Using $\hat{\beta}_0$ and $\hat{\beta}_1$ estimates to find confidence intervals for, and to test hypotheses about, the true values β 1 and β 0 .
- Under assumptions 1 through 4,

$$\hat{\beta}_0 \sim N(\beta_0, s_{\hat{\beta}_0})$$

$$\hat{\beta}_1 \sim N(\beta_1, s_{\hat{\beta}_1})$$

- Under the assumptions 1 to 4, the quantities

$$\frac{\widehat{\beta}_0 - \beta_0}{s_{\widehat{\beta}_0}} \text{ and } \frac{\widehat{\beta}_1 - \beta_1}{s_{\widehat{\beta}_1}}$$

have Student's t distributions with n − 2 degrees of freedom.

Constructing CI for β 0 and β 1

Level $100(1-\alpha)\%$ confidence intervals for β_0 and β_1 are given by

$$\widehat{\beta}_0 \pm t_{n-2,\alpha/2} \cdot s_{\widehat{\beta}_0} \qquad \widehat{\beta}_1 \pm t_{n-2,\alpha/2} \cdot s_{\widehat{\beta}_1}$$

where

$$s_{\widehat{\beta}_0} = s\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}} \qquad s_{\widehat{\beta}_1} = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

Problem

A chemical reaction is run 12 times, and the temperature x_i (in ◦ C) and the yield y_i (in percent of a theoretical maximum) is recorded each time. The following summary statistics are recorded:

$$\bar{x} = 65.0 \qquad \bar{y} = 29.05 \qquad \sum_{i=1}^{12}(x_i - \bar{x})^2 = 6032.0$$

$$\sum_{i=1}^{12}(y_i - \bar{y})^2 = 835.42 \qquad \sum_{i=1}^{12}(x_i - \bar{x})(y_i - \bar{y}) = 1988.4$$

Problem – Continued

Let β 0 represent the hypothetical yield at a temperature of 0 ◦ C, and let β 1 represent the increase in yield caused by an increase in temperature of 1 ◦ C. Assume that assumptions 1 through 4 hold.

1) Compute the least-squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

2) Compute the error variance estimate s^2.

3) Find 95% confidence intervals for β 0 and β 1 .

4) A chemical engineer claims that the yield increases by more than 0.5 for each 1 ◦ C increase in temperature. Do the data provide sufficient evidence for you to conclude that this claim is false?

Solution – Computing least-squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 0.329642$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 7.623276$$

Solution – Computing error variance estimate s^2

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{(1-r^2)\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-2}$$

$$r^2 = \frac{[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})]^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2} = 0.784587$$

$$s^2 = \frac{(1-r^2)\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-2} = 17.996003$$

Solution – Computing 95% confidence intervals for β 0 and β 1

$$\hat{\beta}_0 \sim N(\beta_0, s_{\hat{\beta}_0})$$

$$\hat{\beta}_1 \sim N(\beta_1, s_{\hat{\beta}_1})$$

$$s = \sqrt{17.996003} = 4.242170.$$

$$s_{\hat{\beta}_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = 3.755613$$

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{\sum_{i-1}^{n}(x_i - \bar{x})^2}} = 0.0546207.$$

Solution – Computing 95% confidence intervals for β 0 and β 1

- There are n − 2 = 10 degrees of freedom.

  $t_{10,.025} = 2.228$.

  - Therefore a 95% confidence interval for β 0 is

    $$7.623276 \pm 2.228(3.755613)$$

    or (−0.744, 15.991)

- The 95% confidence interval for β 1 is

  $$0.329642 \pm 2.228(0.0546207)$$

  or (0.208, 0.451)

H 0 : β 1 ≥ 0.5 versus H 1 : β 1 < 0.5

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 0.329642$$

$$s_{\widehat{\beta}_1} = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = 0.0546207.$$

- t = (0.329642 − 0.5)/0.0546207 = −3.119.
- n = 12 -2 = 10
- Since the alternate hypothesis is of the form β 1 < b, the P-value is the area to the left of t = −3.119.
- From the t table, 0.005 < P < 0.01.
- Since P < 0.05 : We can conclude that the claim is false.

Problem: Two engineers are independently estimating the spring constant of a spring, using the linear model specified by Hooke's law.

Engineer A measures the length of the spring under loads of 0, 1, 3, 4, and 6 lb, for a total of five measurements.

Engineer B uses loads of 0, 2, 6, 8, and 12 lb, measuring once for each load.

The engineers all use the same measurement apparatus and procedure. Each engineer computes a 95% confidence interval for the spring constant.

If the width of the interval of engineer A is divided by the width of the interval of engineer B, the quotient will be approximately_____.

Solution

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2} \cdot s_{\hat{\beta}_1}$$

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

n = 5 , $t_{5-2,\ 0.025}$

Assuming s to be constant, the width of a confidence interval is proportional to $1/\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Let S_A and S_B denote the values of $\sum_{i=1}^{n}(x_i - \bar{x})^2$ for Engineers A and B respectively.

Then $S_A = 22.8$, $S_B = 4S_A = 91.2$.

Solution

- Let W_A and W_B denote the widths of the confidence intervals for Engineers A and B, respectively.

$$W_A / W_B \approx \sqrt{S_B / S_A} = 2$$

Checking Assumptions and Transforming Data

Prof. Preet Kanwal

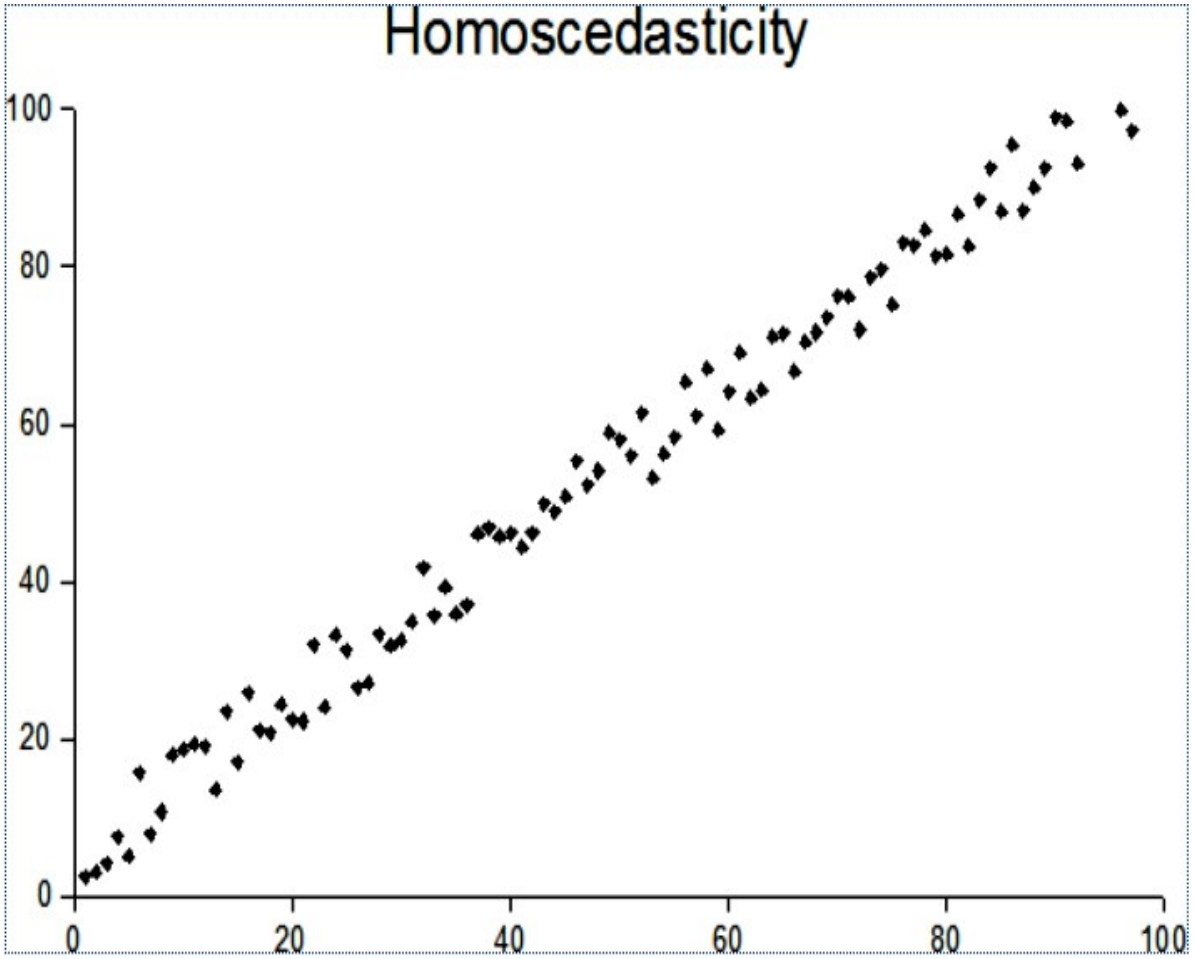
Checking Assumptions

- The methods for inference are valid only if the four Assumptions for Errors in Linear Models hold.
- Hence, we need ways to check these assumptions to assure ourselves that our methods are appropriate.
- There have been innumerable diagnostic tools proposed for this purpose.
- We will restrict ourselves here to a few of the most basic procedures.
  - The Plot of Residuals versus Fitted Values

Assumption 3

- The standard deviations of the error terms are constant and do not depend on the x-value.
- Consequently, each probability distribution for y (response variable) has the same standard deviation regardless of the x-value (predictor).

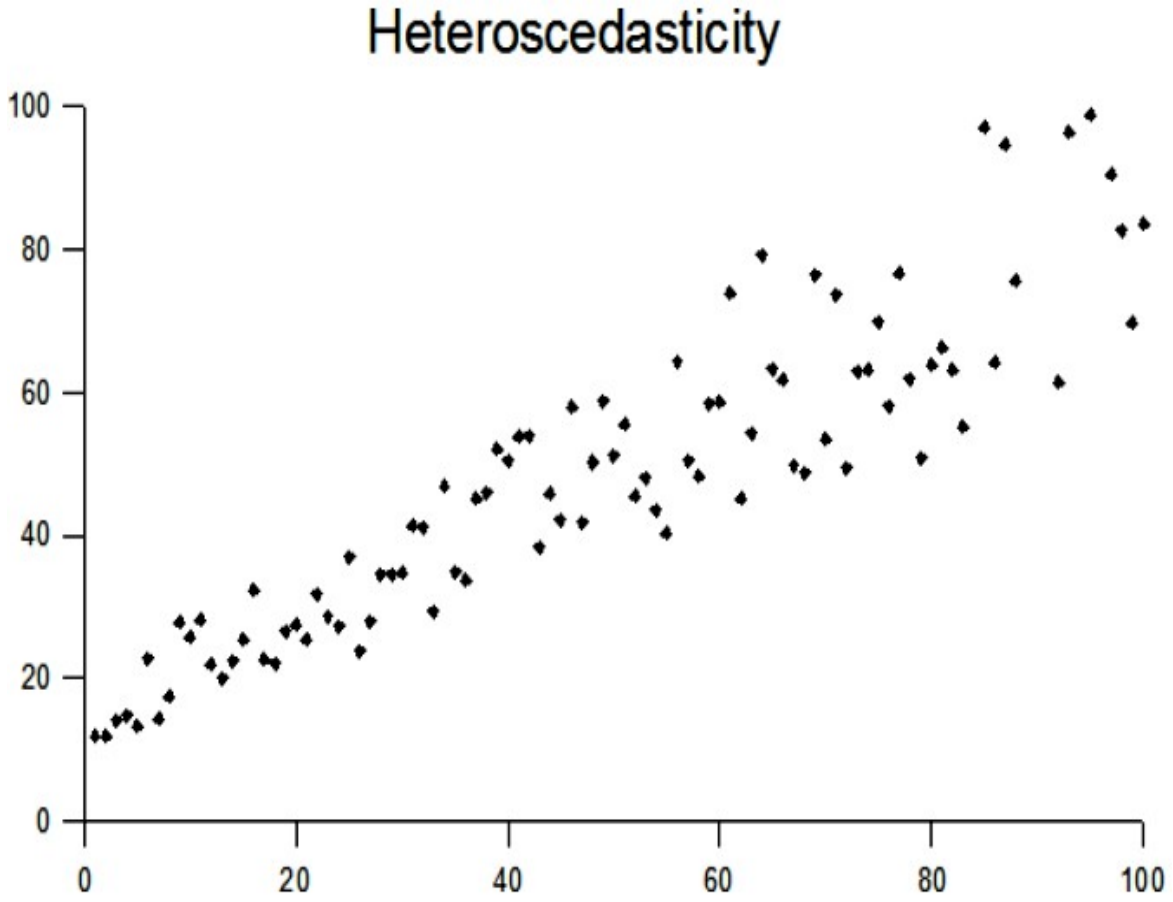
Homoscedastic Scatterplot



[homogeneity of variance]

- When the vertical spread in a scatterplot doesn't vary too much.
- Having data values that are scattered, or spread out, to about the same extent.
- Violations in homoscedasticity results in overestimating the goodness of fit as measured by the Pearson coefficient

Heteroscedastic Scatterplot



- Heteroscedasticity is the absence of homoscedasticity.
- When the vertical spread in a scatterplot varies too much.

Residual Plots

Prof. Preet Kanwal

Introduction

- Randomness and unpredictability are crucial components of any regression model.
- If you don’t have those, your model is not valid.
- **The Residual plot is used to detect non-linearity, unequal error variances, and outliers.**
- Helps to check the validity of a regression model.

Various Residual Plots

- **Residuals versus fitted**
  - Use the residuals versus fits plot to verify the assumption that the residuals have a constant variance.
- **Residuals versus predictors(x)**
  - This plot should show a random pattern of residuals on both sides of 0. Non-random patterns may indicate that predictor variables are unrelated to the residuals.
- **Residuals versus order of data**
  - Use the residuals versus order plot to verify the assumption that the residuals are uncorrelated with each other.

Residual plot

Residual = Observed - Predicted(fitted)

- A residual plot is a graph that shows the residuals on the vertical axis and the fitted value (or x) on the horizontal axis.

- If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.

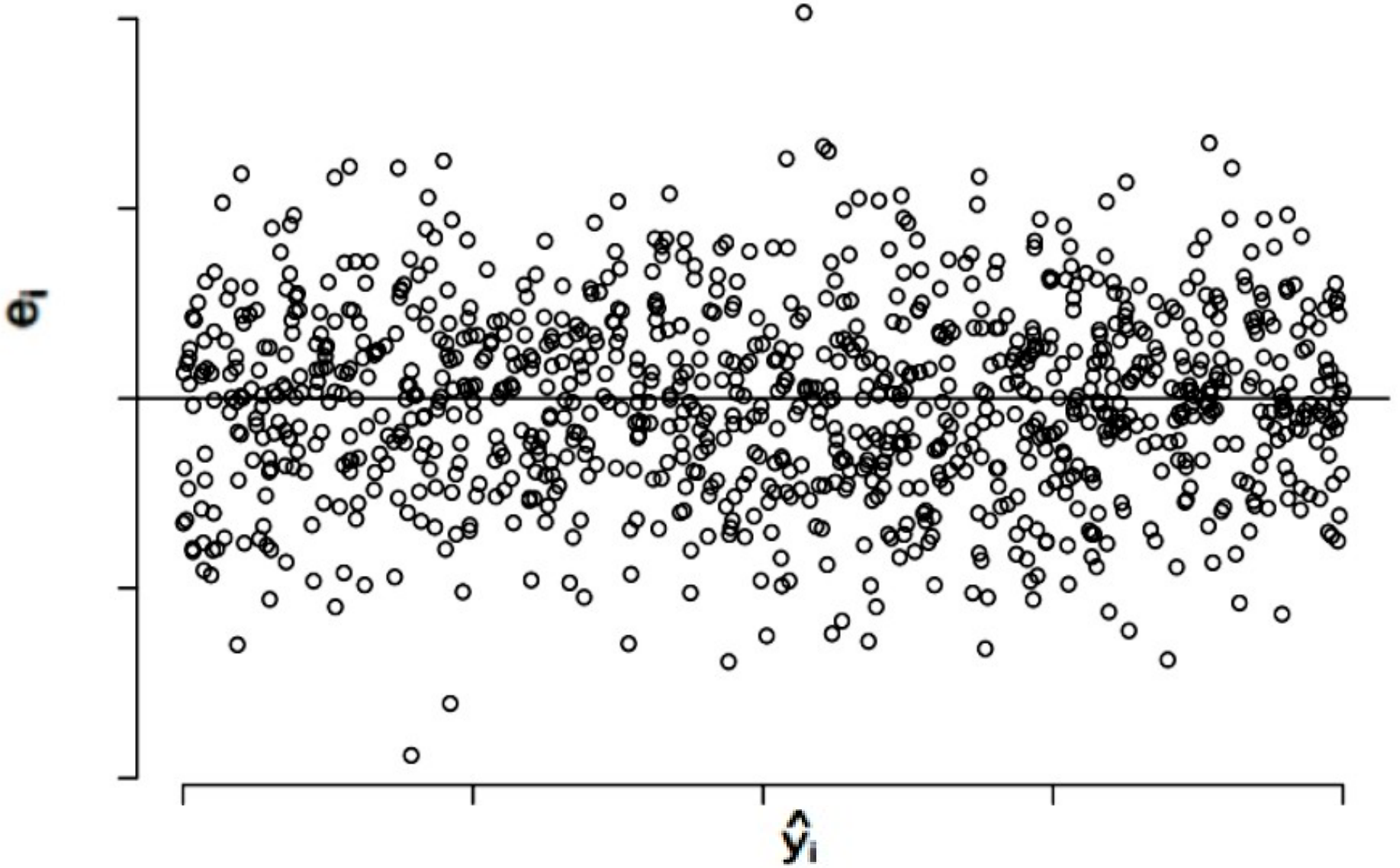
- Positive values for the residual (on the y-axis) mean the prediction was too low, and negative values mean the prediction was too high; 0 means the guess was exactly correct.

Residual plot

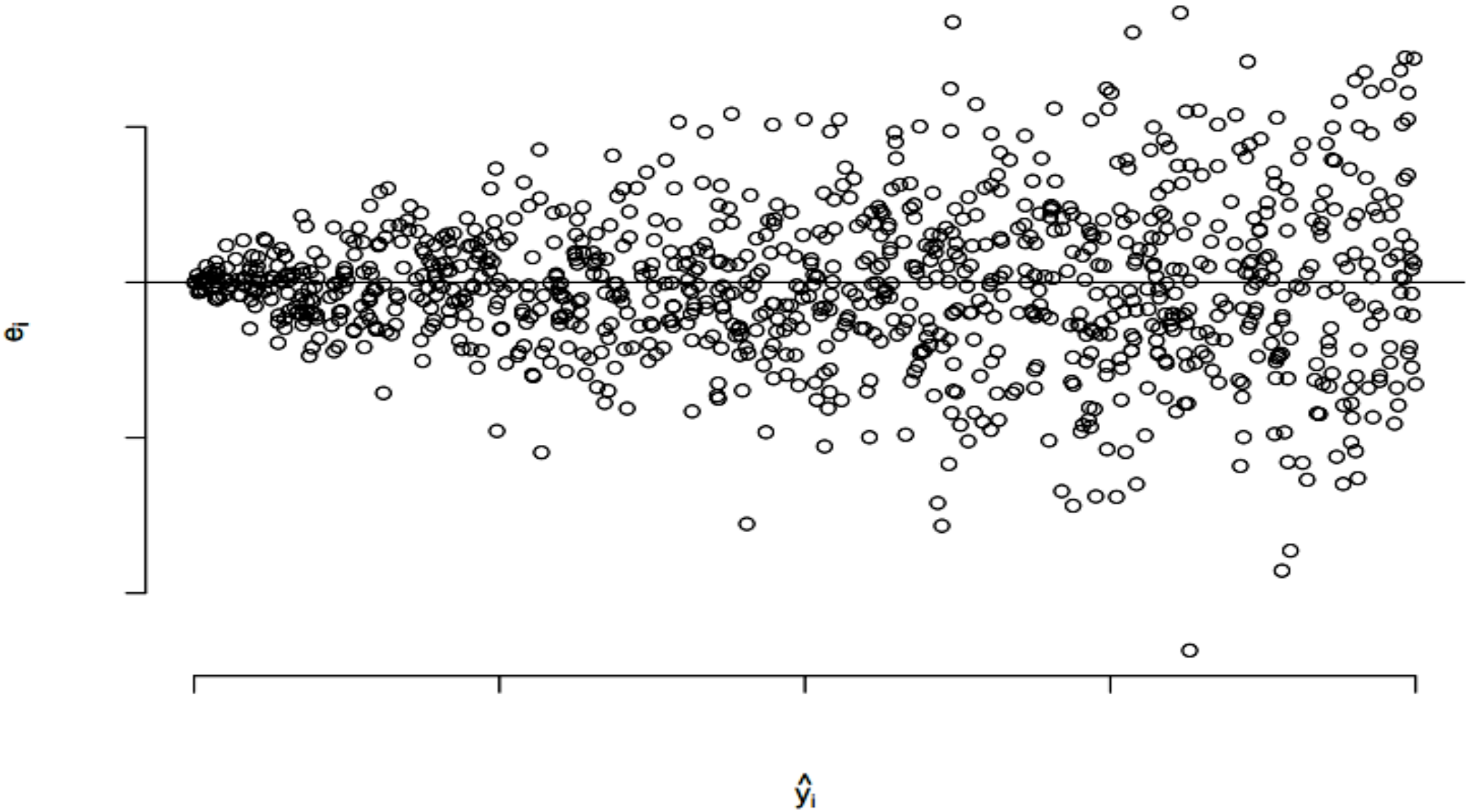
- By mathematical necessity:
  - The residuals have mean 0, and
  - The correlation between the residuals and fitted values is 0 as well.
  - The least-squares line is therefore horizontal, passing through 0 on the vertical axis.

- When the linear model is valid, and assumptions 1 through 4 are satisfied, the plot will show **no substantial pattern.**(You shouldn't be able to predict the error for any given observation.)

Plot of perfect residual distribution



Plot of residual distribution which has unequal variance



Summary

If the plot of residuals versus fitted values

- Shows no substantial trend or curve, and
- Is **homoscedastic**, that is, the vertical spread does not vary too much along the horizontal length of plot, except perhaps near the edges,

then it is *likely*, but not *certain*, that the assumptions of the linear model hold.

However, if the residual plot *does* show a substantial trend or curve, or is **heteroscedastic**, it is certain that the assumptions of the linear model do *not* hold.

Residuals vs. order plot

- Use the residuals versus order plot to verify the assumption that the residuals are uncorrelated with each other.
- If the data are obtained in a time (or space) sequence, a residuals vs. order plot helps to see if there is any correlation between the error terms that are near each other in the sequence.
- If the residuals are randomly distributed around zero, it means that there is no drift in the process.

  Note: The plot is only appropriate if you know the order in which the data were collected!

Problem

- An article presents calculations of the ages (in calendar years before 1950) of several sediment samples taken at various depths (in cm) in Lago di Fimon, a lake in Italy. The results are presented in the following table.

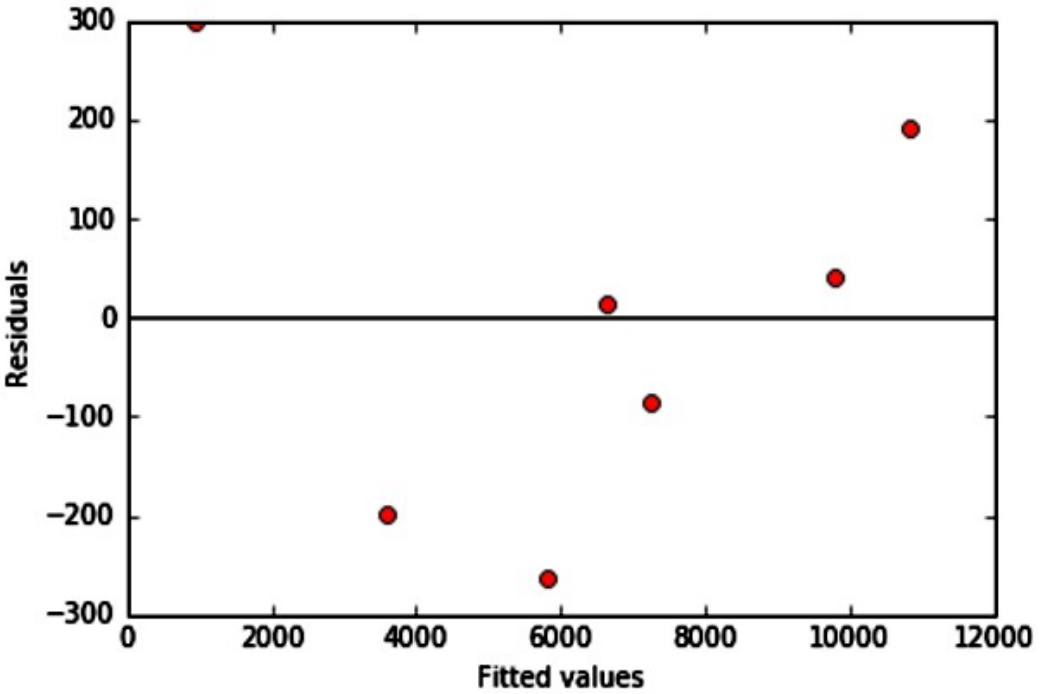
Depth	Age
284.5	1255
407.5	3390
512.0	5560
551.0	6670
578.5	7160
697.0	9820
746.5	11,030

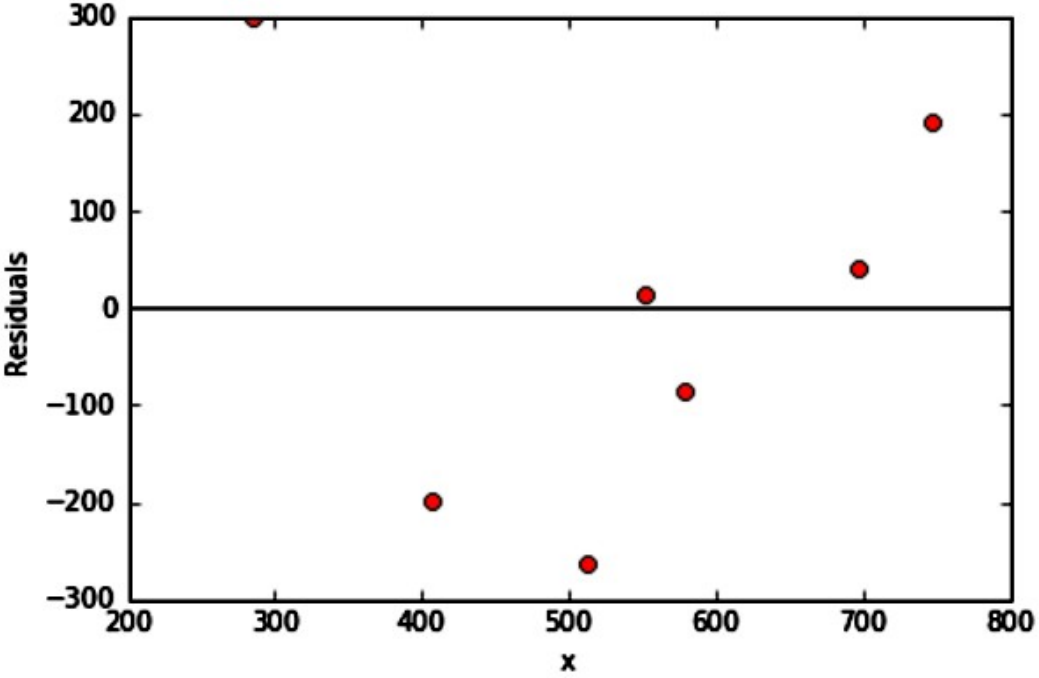
Construct Residual Plot.

Residuals

x(depth)	y(age : observed)	y_hat(predicted) (**-5129.026712 + 21.389512*x**)	Residuals(Obsv - predicted)
284.5	1255	956.289452	298.710548
407.5	3390	3587.199428	-197.199428
512.0	5560	5822.403432	-262.403432
551.0	6670	6656.5944	13.4056
578.5	7160	7244.80598	-84.80598
697.0	9820	9779.463152	40.536848
746.5	11030	10838.243996	191.756004
		Sum of all Residuals	0.00015

Residual Plots





Residual Plots with Only a Few Points Can Be Hard to Interpret

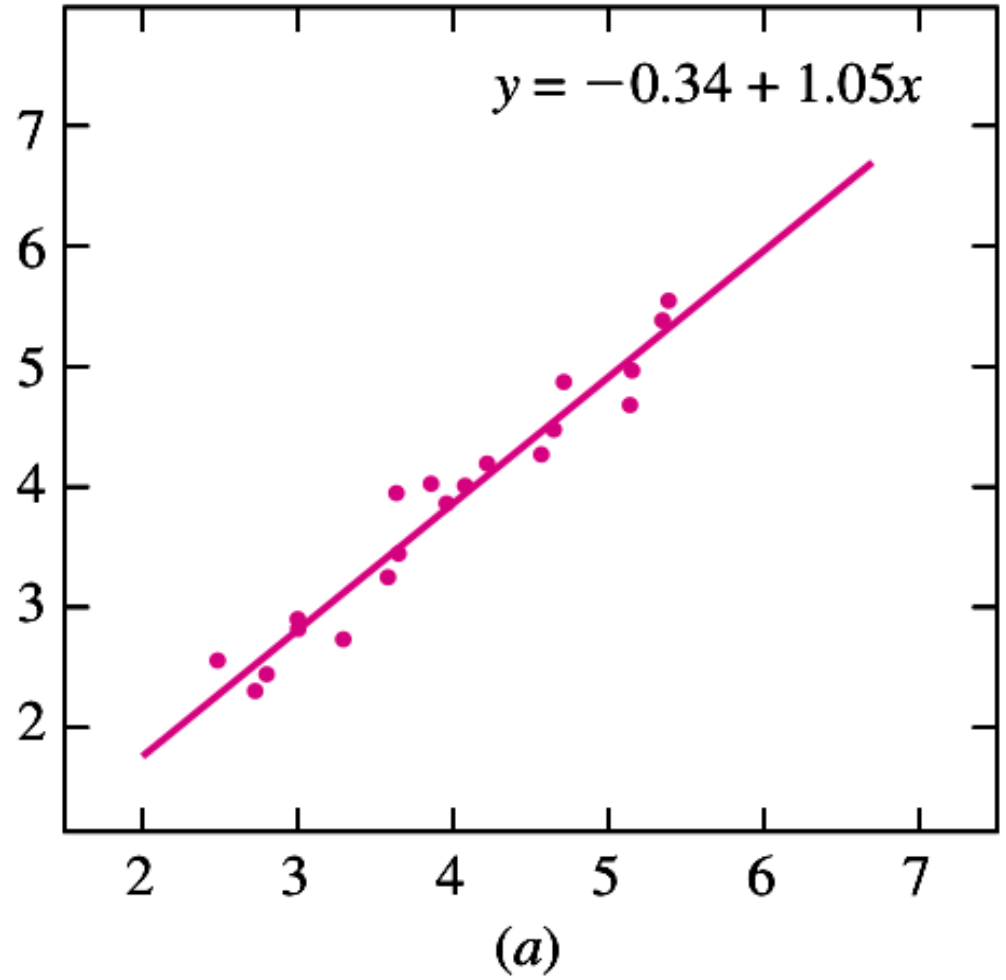
- When one is faced with a sparse residual plot that is hard to interpret, a reasonable thing to do is to fit a linear model, but to consider the results tentative, with the understanding that the appropriateness of the model has not been established.

- If and when more data become available, a more informed decision can be made.

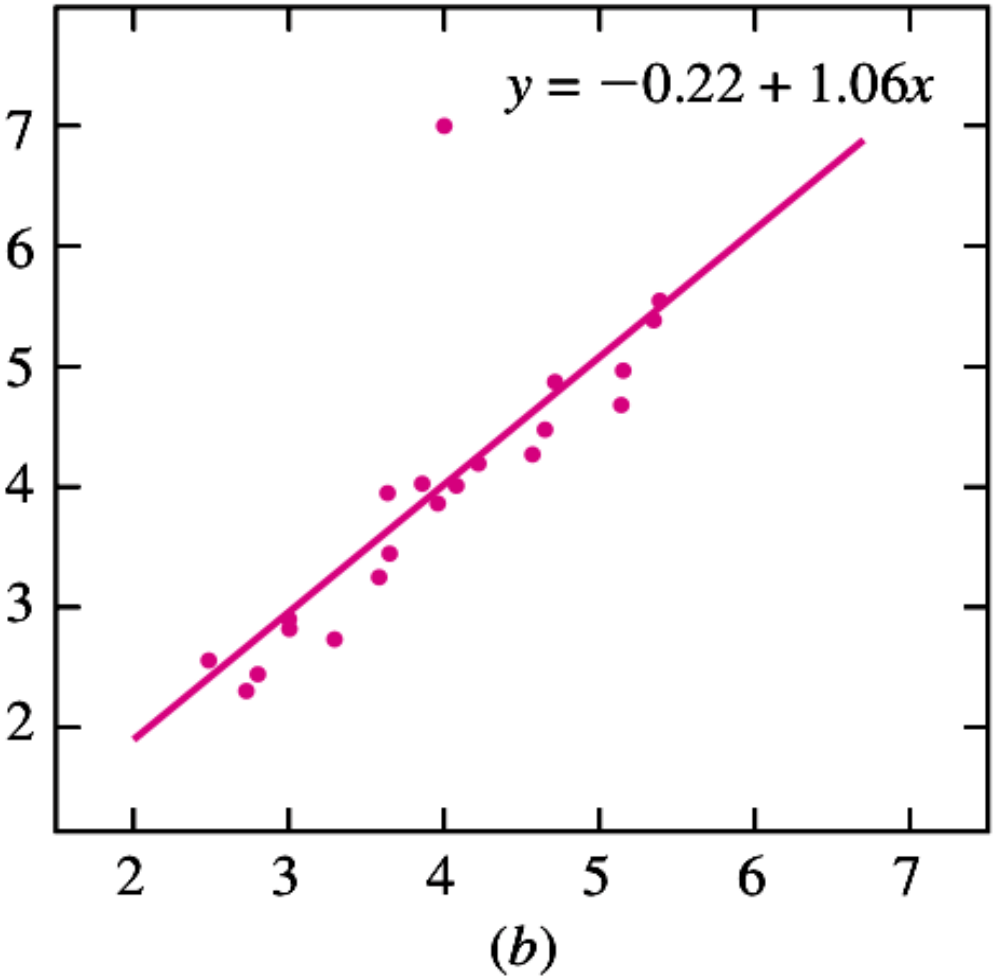
Outliers and Influential Points

- Both the scatterplot and the residual plot should be examined for outliers.
- The first thing to do with an outlier is to try to determine the cause for its existence.
- Outliers can often be identified by visual inspection.
- Sometimes transforming the variables will eliminate outliers by moving them nearer to the bulk of the data.
- **An outlier that makes a considerable difference to the least-squares line when removed is called an influential point.**
- In general, outliers with unusual x values are more likely to be influential than those with unusual y values, but every outlier should be checked.

Not Influential Outlier

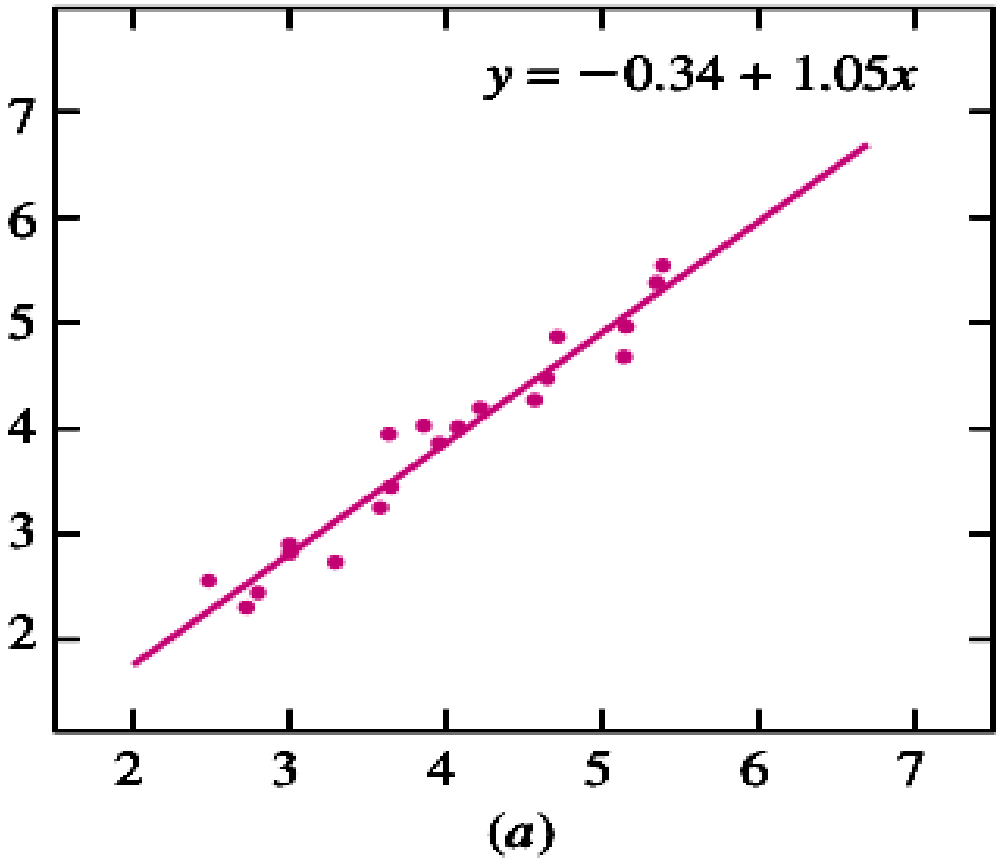


a) No outlier

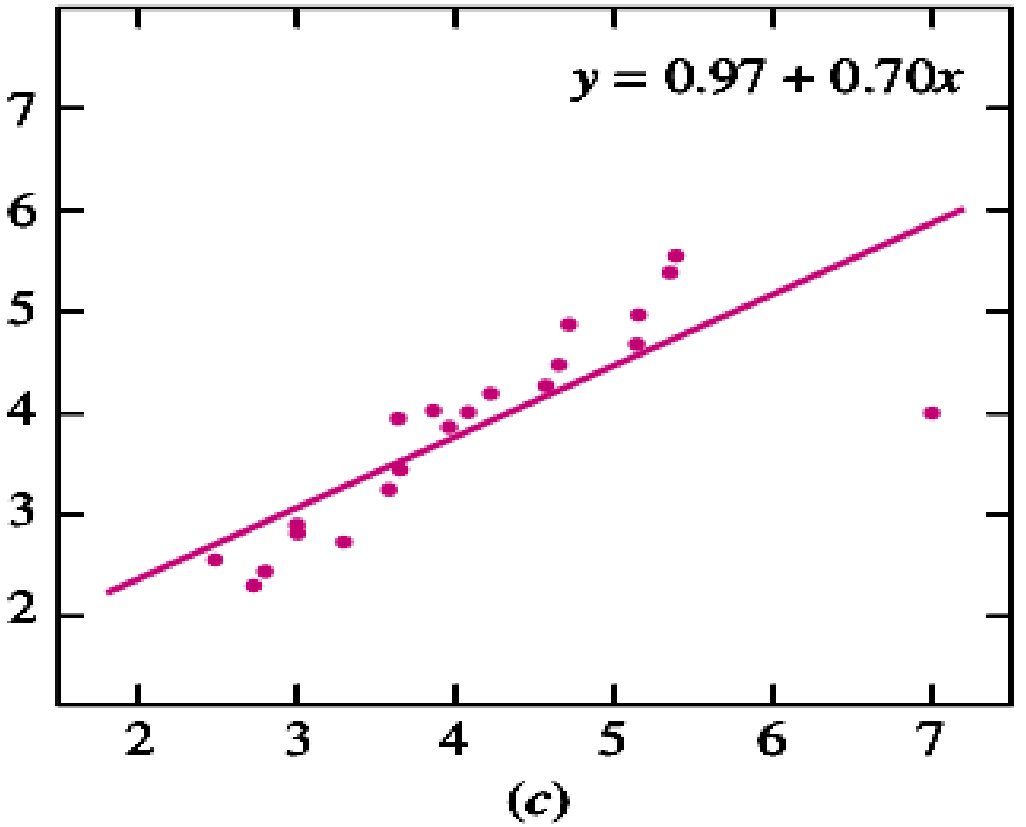


b) With outlier(not influential as little change in least-squares line) – hence, can be removed.

Influential Outlier



a) No outlier



c) Outlier Added - considerable change in the least-squares line, so this point is influential.

Checking Independence of errors(ε_i) - (Assumption 1) : plot of Residual vs. time

- Plot the residuals against the order in which the observations were made(time).
- If there are trends in the plot, it indicates that the relationship between x and y may be varying with time.
- In this case time must be used as another independent variable and multiple regression should be performed.
- This indicates that the value of each error is influenced by the errors in previous observations, so therefore the **errors are not independent.**

Checking Normality of errors(ε_i) – (Assumption 4)

- A normal probability plot of the residuals can be made to chck normaility of errors.
- It can be a good idea to make a probability plot when variables are transformed.
- The assumption of normality is not so important when the number of data points is large.

Problem [Python Demo]

A windmill is used to generate direct current. Data are collected on 45 different days to determine the relationship between wind speed in mi/h (x) and current in kA (y). The data are provided in the **“Residuals.csv”**

Problem

- Compute the least-squares line for predicting:

  1) x vs. y

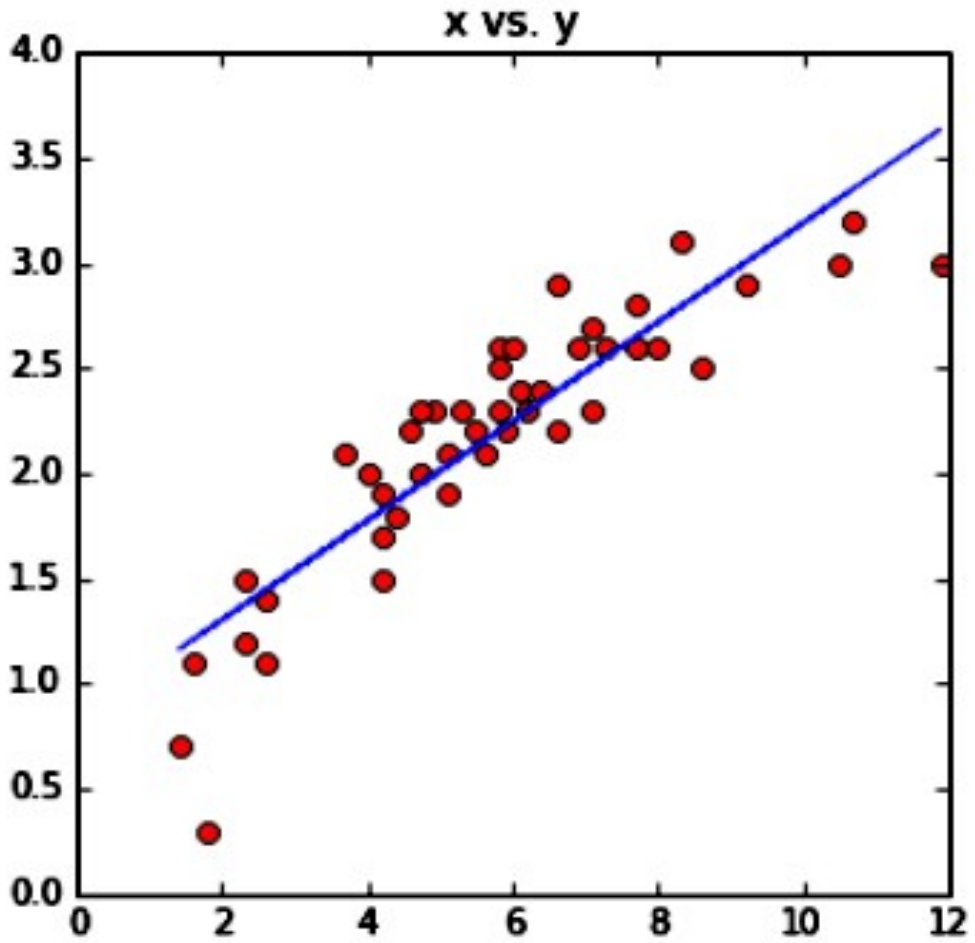
  2) ln x vs. y

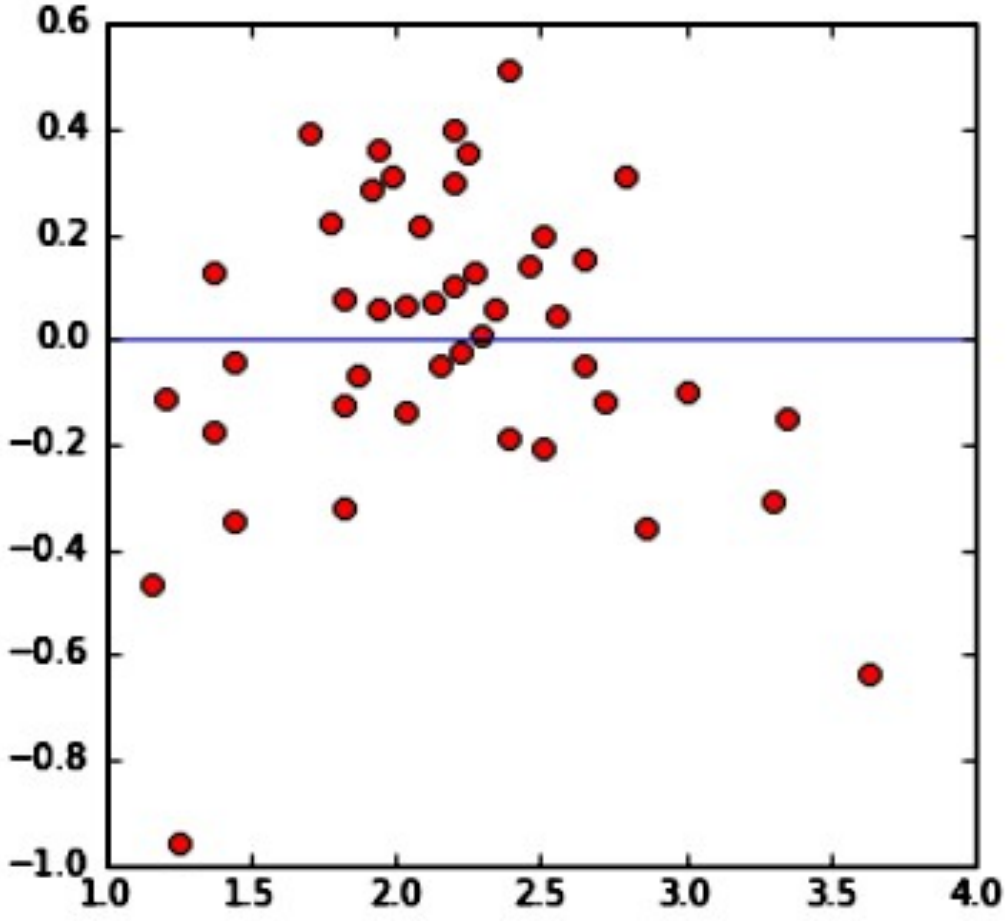
  3) x vs. ln y

  4) x vs. sqrt(y)
- Print intercept(0) and slope (1) value in each case.
- Make a plot of residuals versus fitted values in each case.
- Which of the four models (1) through (4) fits best? Explain.

x vs. y

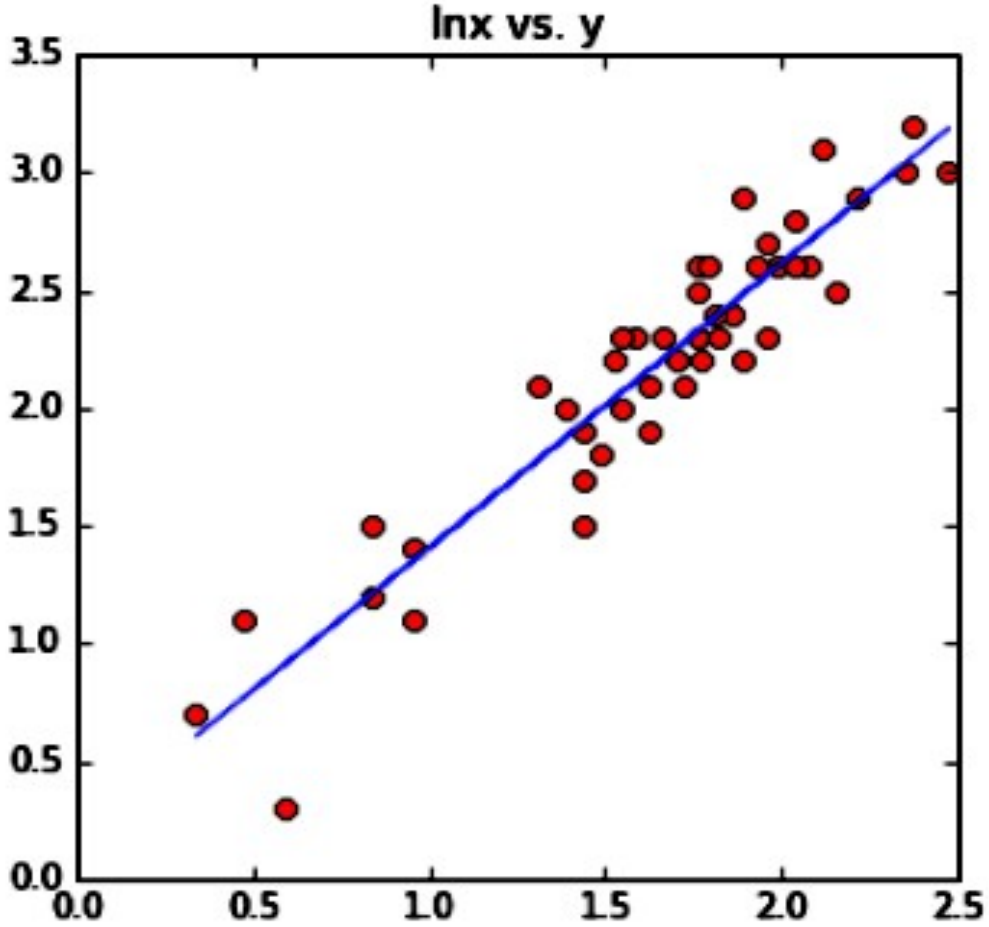
```
b0 :   0.833248548766
b1 :   0.235423405858
```

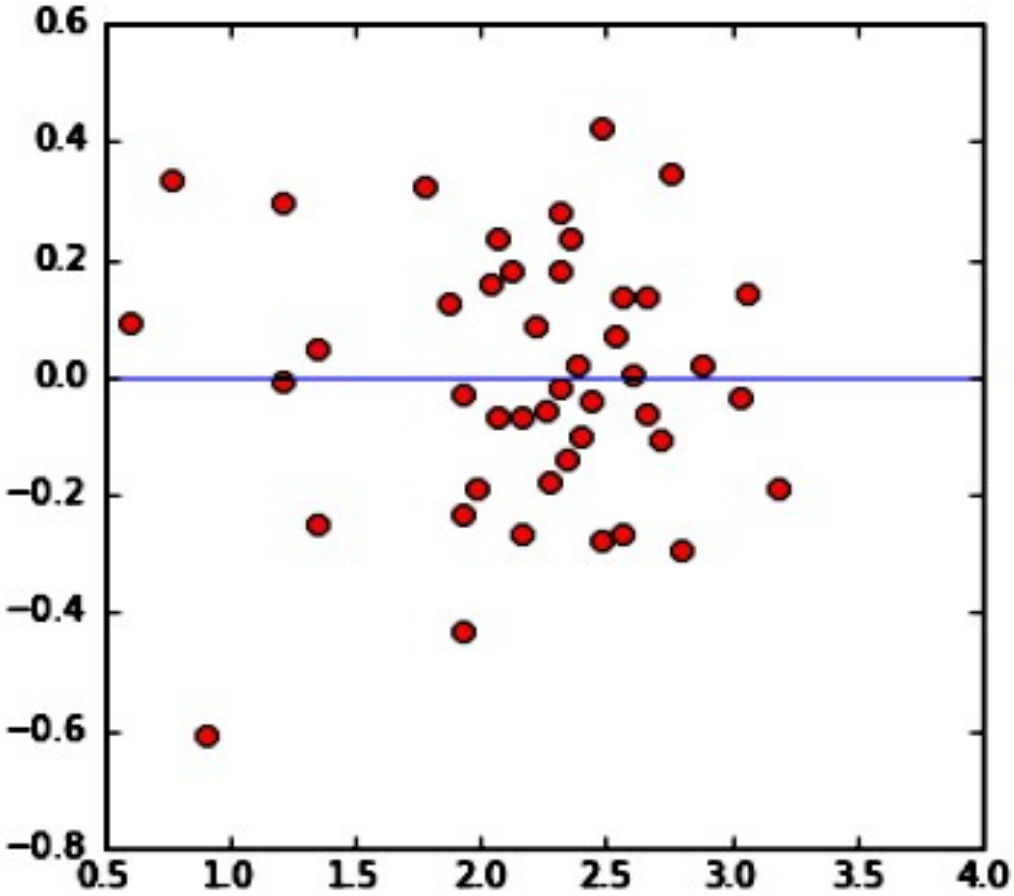




ln x vs. y

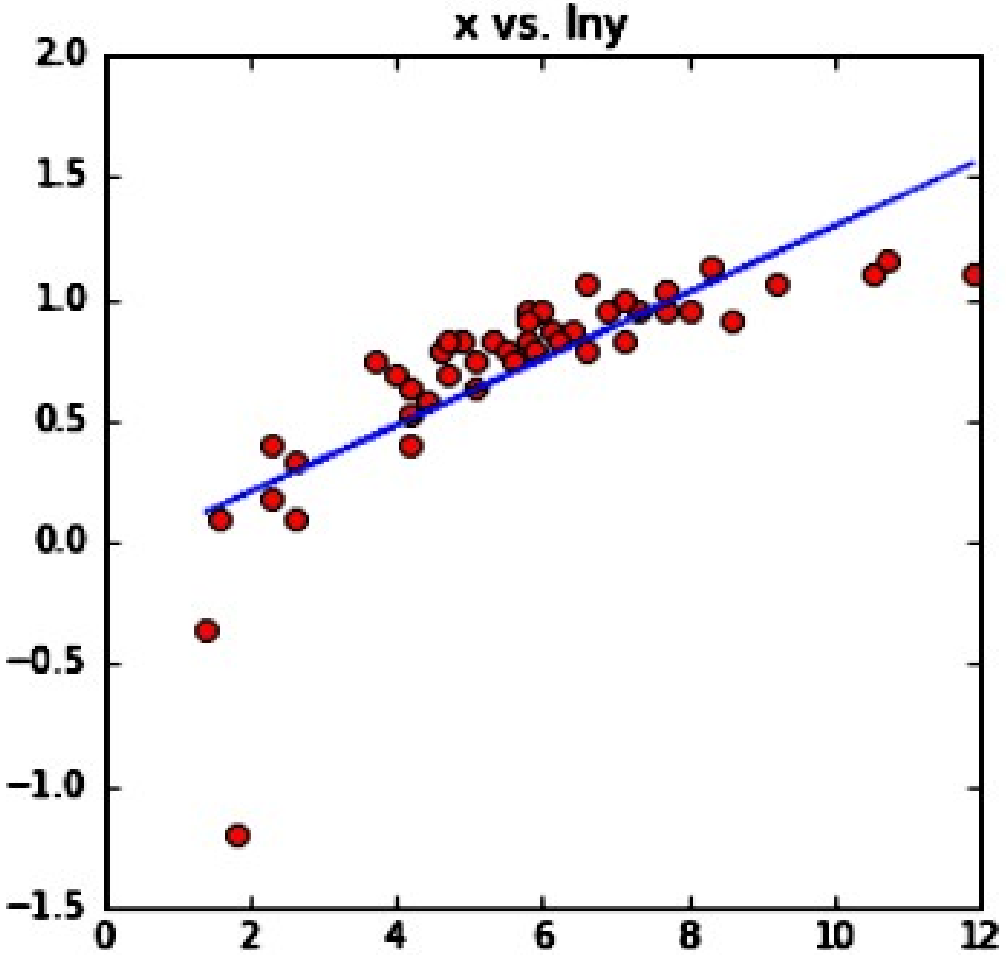
```
b0 :   0.198782099188
b1 :   1.20658835366
```

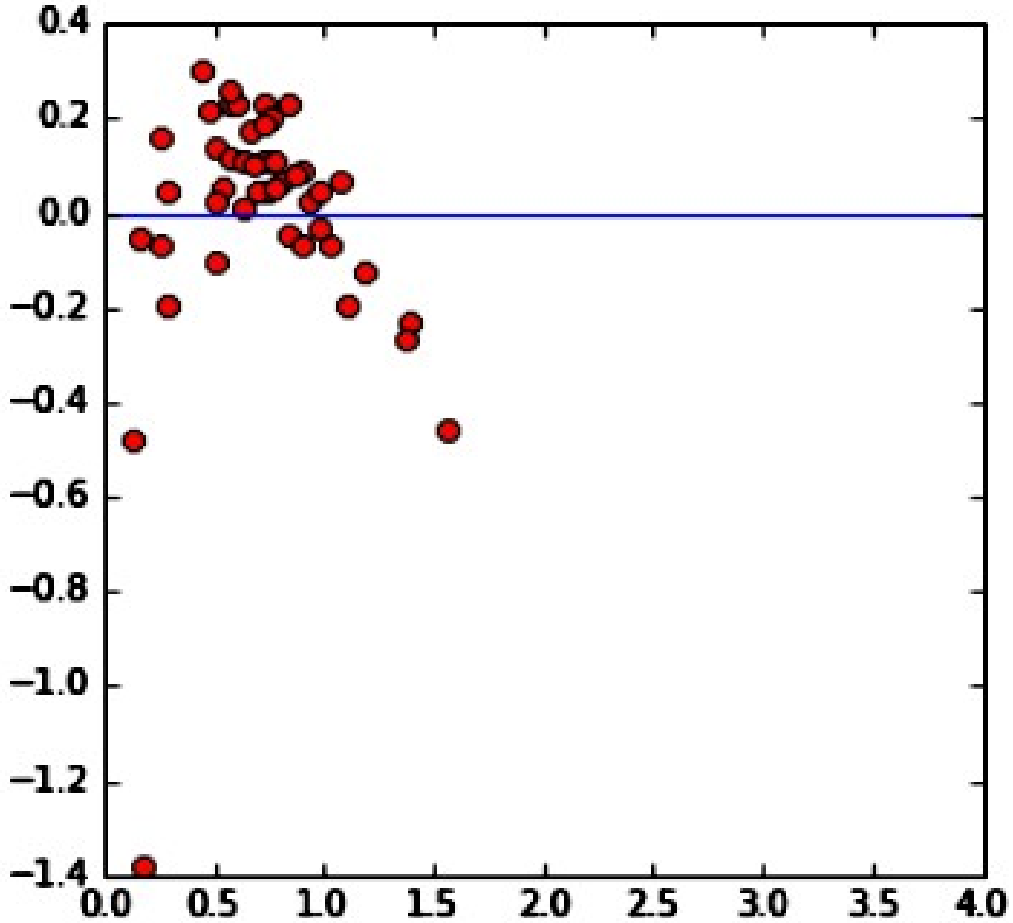




x vs. ln y

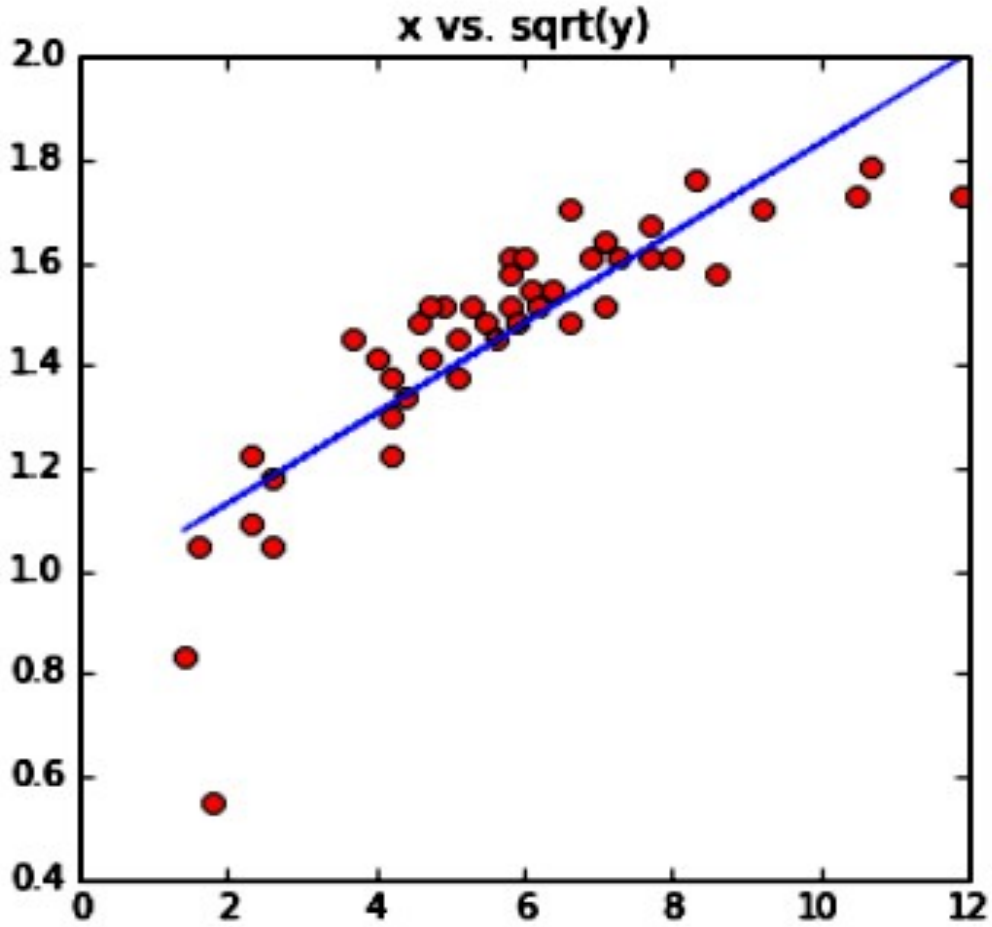
```
b0 :   -0.0678688289177
b1 :   0.136731381153
```

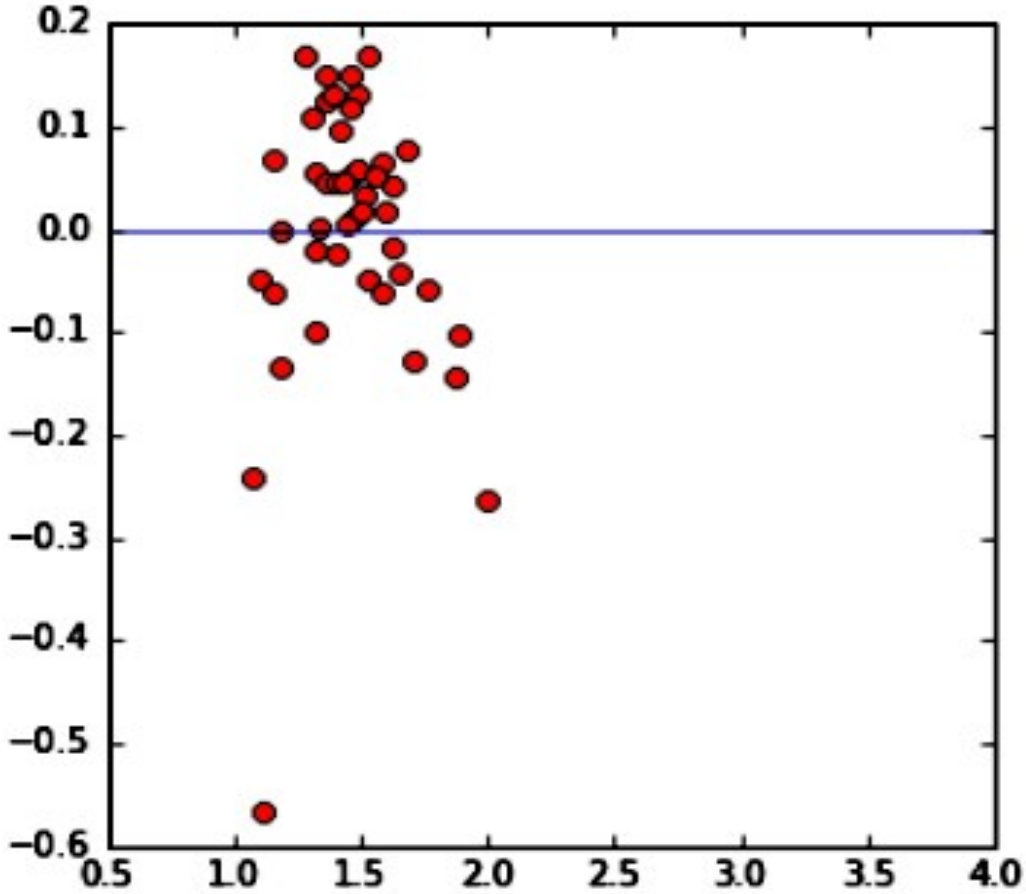




x vs. sqrt(y)

```
b0 :   0.956039085838
b1 :   0.0874331090987
```





Methods to Fix Voilations

1) Transforming the variables
2) Weighted least-squares
3) Multiple Regression
4) Non linear Regression

Transforming the variables

- Fixes violations of assumptions and allow the linear model to be used.
- Replacing a variable with a function of itself is called transforming the variable. Specifically, raising a variable to a power is called a power transformation.
- Taking the logarithm of a variable is also considered to be a power transformation, even though the logarithm is not a power.

Determining Which Transformation to Apply

- Proceed by trial and error.

Transformations Don't Always Work

- Other methods should be used.

Methods Other Than Transforming Variables

1) Weighted least-squares
2) Multiple Regression
3) Non linear Regression

Weighted least-squares

In this method, the x and y coordinates of each point are multiplied by a quantity known as a weight.

- Points in regions where the vertical spread is large are multiplied by smaller weights
- Points in regions with less vertical spread are multiplied by larger weights. The effect is to make the points whose error variance is smaller have greater influence in the computation of the least-squares line.

Multiple Regression

- When the residual plot shows a trend, this sometimes indicates that more than one independent variable is needed to explain the variation in the dependent variable.
- In these cases, more independent variables are added to the model, and multiple regression is used.

Nonlinear regression

Some relationships are inherently nonlinear. For these, a method called nonlinear regression can be applied.

Empirical Models and Physical Laws

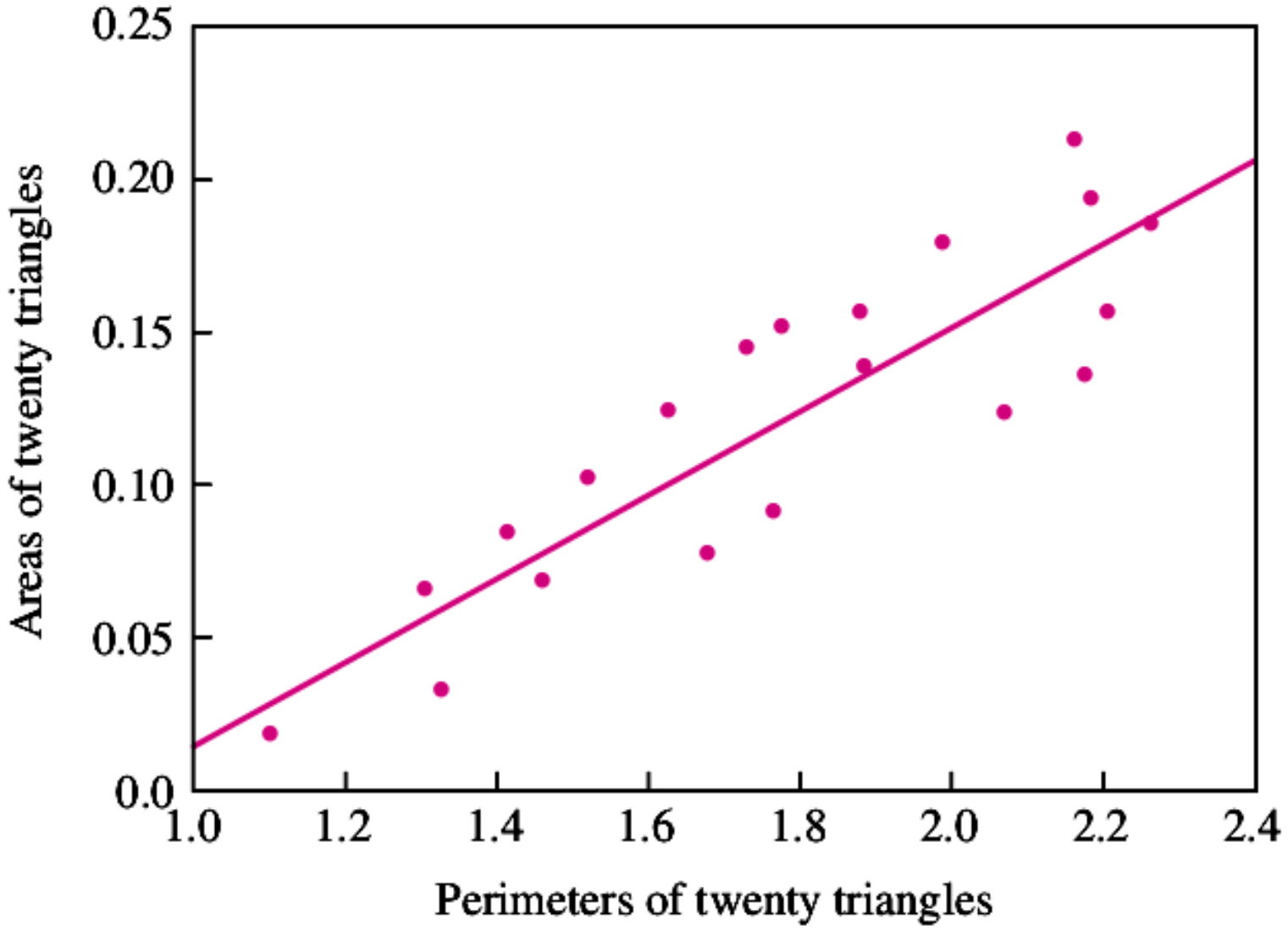
- Data analysis can be based on empirical models or physical laws(such as Hooke's law).
- A model that is chosen because it appears to fit the data, in the absence of physical theory, is called an empirical model.
- In real life, most data analysis is based on empirical models.

Interpretation of results based on empirical models and Physical laws

- A physical law may be regarded as true. For example, in the Hooke's law data, we can be sure that the relationship between the load on the spring and its length is truly linear.

- The best we can hope for from an empirical model is that it is useful. For example : scatterplot of area versus perimeter of a triangle.

Interpretation of results based on Empirical models



Empirical models : Example (Area vs. Perimeter of a Triangle)

- We notice, that triangles with larger perimeters seem to have larger areas, so we fit a linear model:

$$\text{Area} = \beta\ 0 + \beta\ 1\ (\text{Perimeter}) + \varepsilon$$

- The correlation between area and perimeter is r = 0.88, which is strongly positive. The linear model appears to fit well.

$$\text{Area} = -1.232 + 1.373\ (\text{Perimeter})$$

Empirical models : Example (Area vs. Perimeter of a Triangle)

- While this linear model may be useful, it is not true.
- In the absence of a better method, it may be of some use in estimating the areas of triangles.
- The true mechanism, of course, is given by the law

$$\text{Area} = 0.5 \times \text{base} \times \text{height}$$

- A collection of triangles could be designed in such a way that :
  - the ones with the larger perimeters had smaller areas.
  - the area might appear to be proportional to the square of the perimeter

Summary

- Physical laws are applicable to all future observations.
- An empirical model is valid only for the data to which it is fit. It may or may not be useful in predicting outcomes for subsequent observations.
- Determining whether to apply an empirical model to a future observation requires scientific judgment rather than statistical analysis.

Thank you !