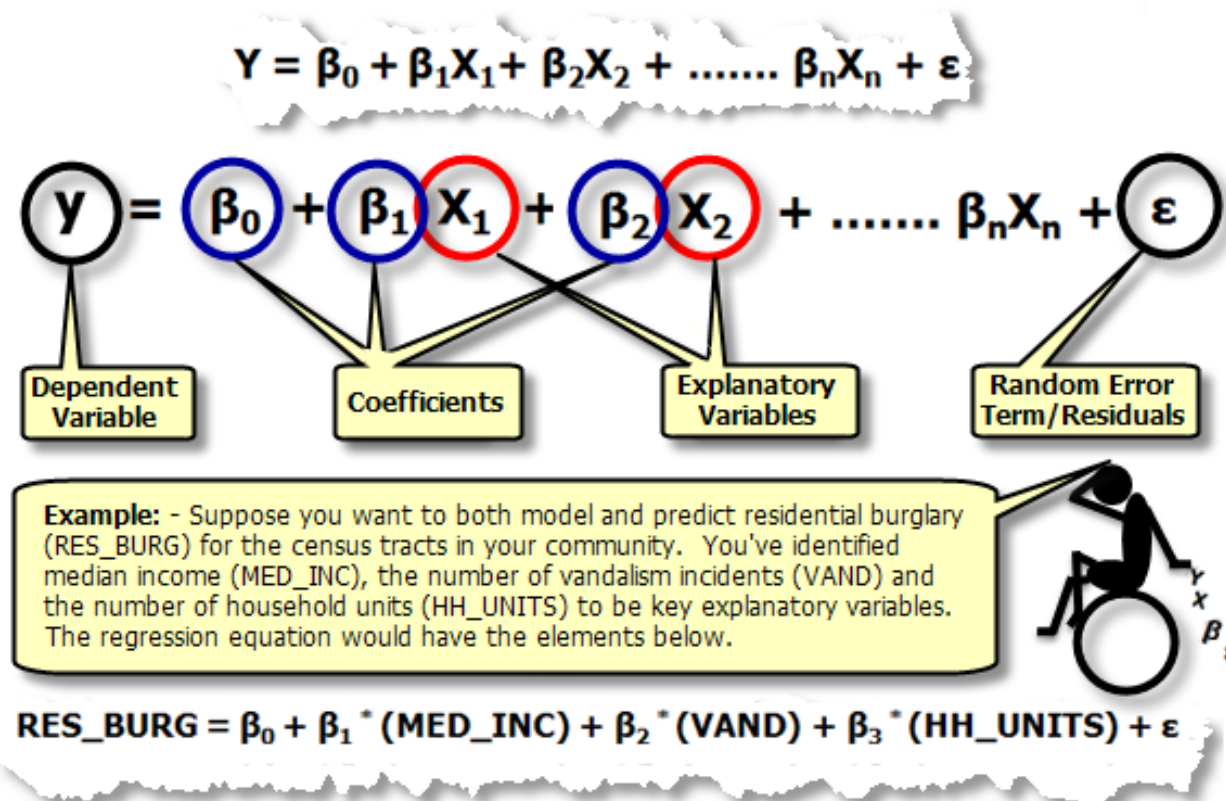


Topic 4: Simple Correlation and Regression Analysis



Question: *“How do we determine how the changes in one variable are related to changes in another variable or variables?”*

Answer: REGRESSION ANALYSIS

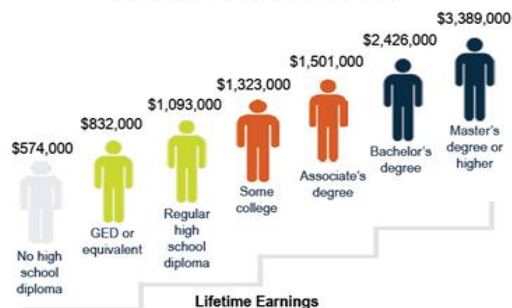
↳ Which derives a description of the functional nature of the relationship between two or more variables.

Examples:

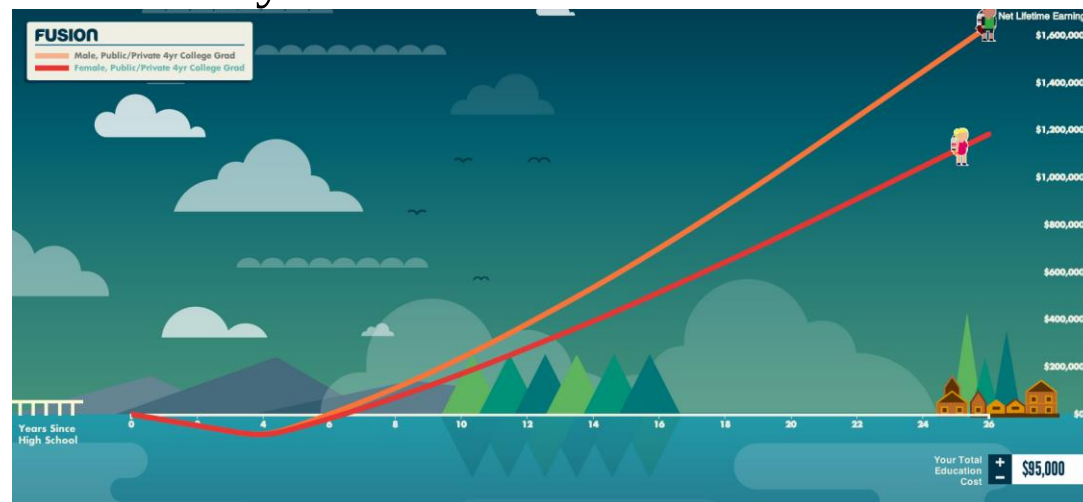
(i) *Life-time earnings* can be explained by:

- ◆ Educational level
- ◆ Job experiences ◆ occupation
- ◆ Gender ◆ country of birth

Over their lifetimes, Illinoisans who drop out of high school make hundreds of thousands, even millions, of dollars less than their graduating and higher-educated peers.



Sum, A., & Khatiwada, I. (2012, November). *High school dropouts in Chicago and Illinois and their persistent labor market problems*. Chicago: Center for Labor Market Studies, Northeastern University. Reflects 18 to 64 year olds.



(ii) *Age of Death* can be explained by

◆ Parent's age of death

◆ Race

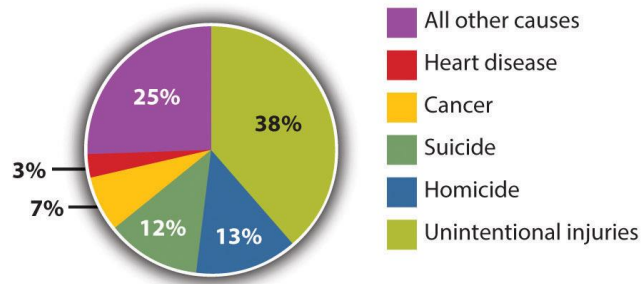
◆ Healthcare accessibility

◆ weight

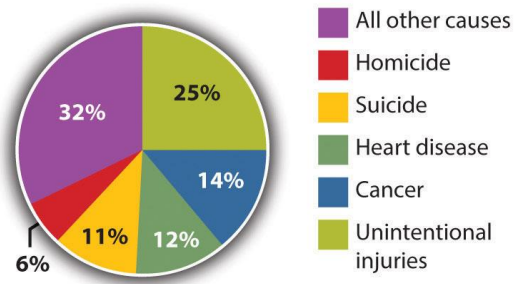
◆ political situation

◆ diet

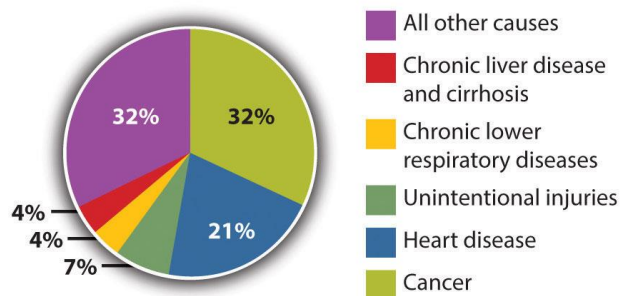
Aged 1–24 years
Number of deaths = 39,086



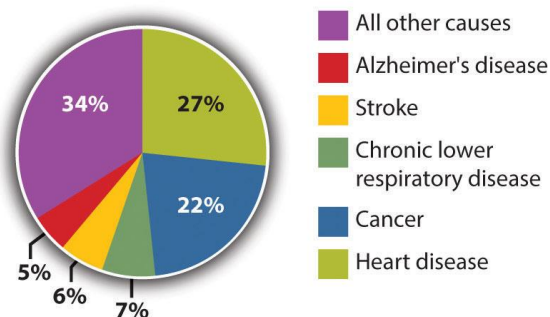
Aged 25–44 years
Number of deaths = 112,117



Aged 45–64 years
Number of deaths = 493,376

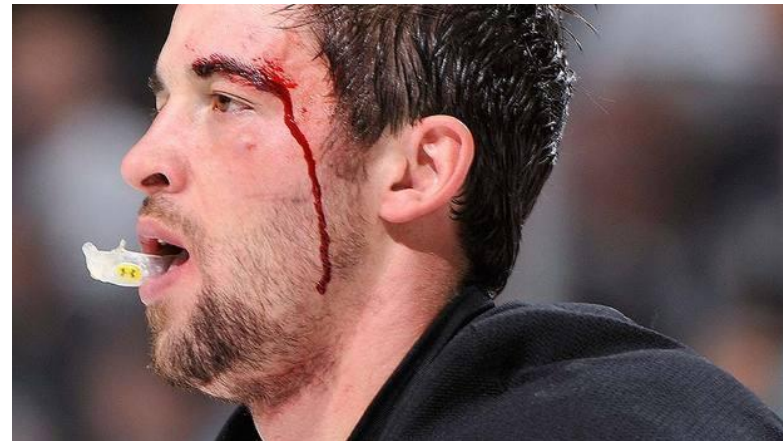
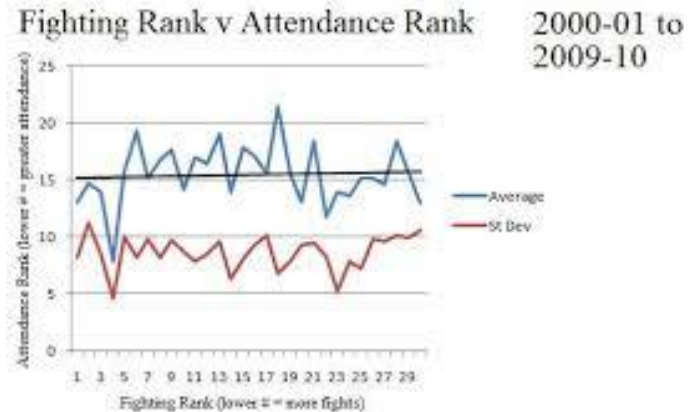


Aged 65 and over
Number of deaths = 1,796,620



(iii) ***Attendance at hockey games*** can be explained by

- ◆ Rank of team
- ◆ player salary
- ◆ History of penalties
- ◆ violence during the game

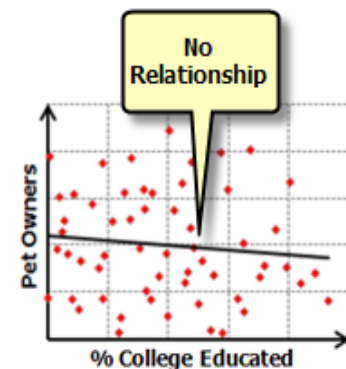
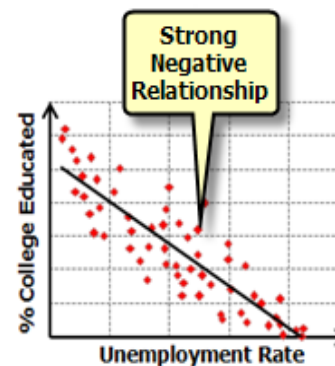
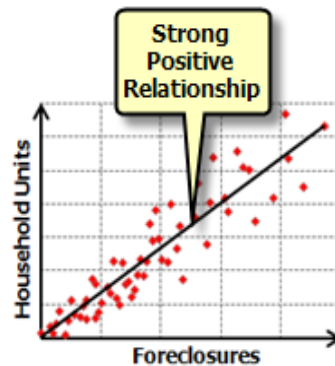
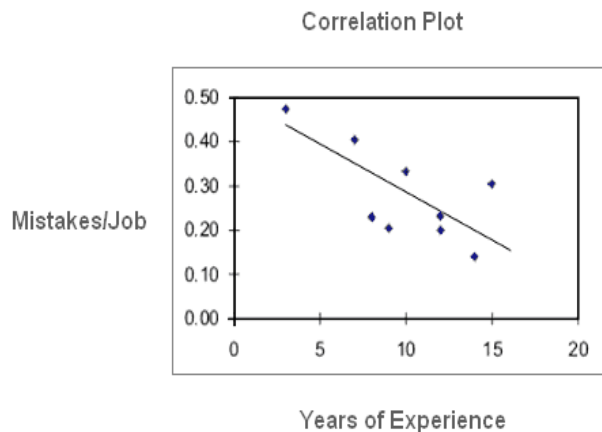


(II) Question:

“How do we determine the strength of the relationship between two or more variables?”

Answer: Correlation Analysis

Which determine the “strength” of such relationships between two or more variables.



This topic will cover the fundamental ideas of **econometrics.**

☆ We will transform theoretical economic relationships into specific functional forms by:

- (i) Gathering the data to estimate the parameters of these functions
- (ii) Test hypotheses about these parameters and
- (iii) Do predictions.

We will bring together:

- (i) Economic theory. i.e. demand theory
- (ii) Empirical data. i.e. price index, GDP
- (iii) Statistical tools.

With regression analysis we estimate the value of one variable (dependent variable) on the basis of one or more other variables (independent or explanatory variables.)

Examples: Demand Function

Suppose the demand for Good A can be expressed by the following:

$$Q_A = f(P_A, P_B, M) \quad \text{“multi-variate” relationship.}$$

The quantity of Good A demanded is a function of its own price (P_A), the price of another good (P_B) and disposable income, (M).

According to the Law of Demand, there is an inverse relationship between the quantity of A demanded (Q_A) and its price, (P_A), given that everything else is constant (ceteris paribus).

This is a partial relationship.

Q_A is referred to as the dependent variable, while P_A , P_B and M are independent or explanatory variables.

 More Happy Emotions and Less Stress associated with...	 Less Happy and More Sad Emotions associated with...	 More Stress associated with...
✓ Being religious ✓ Being older than 60 ✓ Being the weekend! ✓ Being married ✓ Having health insurance	✗ Being alone ✗ Having a health condition ✗ Being a smoker ✗ Being female ✗ Being a caregiver for an elderly/disabled relative	✗ Having headaches ✗ Being alone ✗ Having a health condition ✗ Being a caregiver for an elderly/disabled relative ✗ Having children at home ✗ Being a college graduate

- Economic theory dictates the variable or variables whose values determine the behaviour of a variable of interest.

i.e. Economic theory make claims about the variables the affect the value of the variable of interest.

- For example, the price of a new car will determine the quantity of cars demanded.

Or the interest rate on mortgages, CPI and the GDP may partially determine the demand for new housing.

But:

- (i) Economic theory does not always provide the expected signs for all explanatory variables.

Example: P_B : its sign is the demand function depends on whether the good is a complement or substitute good.

- (ii) Economic theory does not always provide clear and precise information about the functional form of the function.

The functional form of these type of relationships must be specified before estimation can take place. Otherwise, estimation of economic theory can be misleading.

Example: A linear demand function is specified

$$Q_A = \beta_1 + \beta_2 P_A + \beta_3 P_B + \beta_4 M$$

- Where the β_i 's are the parameter of the model.
- Parameters are usually unknown.
- We often wish to estimate them.

Note: Q_A , P_A , P_B , and M may be transformations of non-linear data:

For example: $Q_A = \frac{1}{Q_C}$ or $Q_C^2 = Q_A$.

These non-linear relationships have been transformed into a linear format and hence, expressed in a linear regression model.

$$Q_{Sales} = (Q_{Sales})^2$$

For example: $Q_{Sales} = \left[\frac{1}{Q_{Sales}} \right]$

Once we have determined the functional form of the regression, we can address questions such as:

.....

- If we change the income tax laws, will there be an effect on the quantity of new homes demanded, which works through disposable income?

- Are goods A and B substitutes or complements (which is determined by cross price elasticities)? If the price of B rises, does the quantity of A increase or decrease?
- If P_B increases and it is a complement: Q_A demanded decreases.

Example: Quantity of large vehicles demanded decreases when the price of gasoline increases.

Example: the price of batteries and the demand for battery powered toys.

If P_B increases and it is a substitute: Q_A demanded increases.

Example: Quantity of apples demanded increases when the price of oranges increases.

- Apples and oranges are substitutes.
- Above we had examples of exact relationships between the dependent variable and the explanatory variables.

The functional relationships were deterministic (i.e. no uncertainty involved.)

However, in the real world, such relationships are not as simple or straight forward.

For illustration, if we look at two businesses that build the same product, and face the same production costs and factor prices, their demands are not usually the same.

There are many other individual factors that have been ignored. (Firm's ideology, skill level of employees, marketing experience, etc.)

There will be some uncertainty: “**stochastic**” function.

Example: Suppose we have several households with the same income, facing the same prices. The typical demands from each household will differ because other individual factors have been ignored by the model.

Such as: tastes
Beliefs
Health
Influence of advertising
Age
Gender

These factors project uncertainty into the function.

This stochastic element is incorporated into the function by including a “stochastic” error term into the model.

$$Q_Y = \beta_1 + \beta_2 P_Y + \beta_3 P_X + \beta_4 M + \varepsilon$$


Otherwise, we could simply regard the deterministic relationship as explain the average relationship between the dependent variable and a set of explanatory variables:

$$E(Q_Y) = \beta_1 + \beta_2 P_Y + \beta_3 P_X + \beta_4 M$$

For this topic, we will restrict our models to the following:

(1) Simple Regression

One dependent variable(Y) and one independent variable (X). This can be extended to a multi-variate case.

(2) Linear Parameters:

We assume the relationship between the dependent variable and the independent variable is linear in terms of parameters:

$$E(Y) = \alpha + \beta X$$

Population regression line is linear in terms of parameters α and β , which depict the functional relationship between X and Y .

(3) $E(Y) = \alpha + \beta X$ Means for a given value of X , the expected value of Y (average) is given by $\alpha + \beta X$.

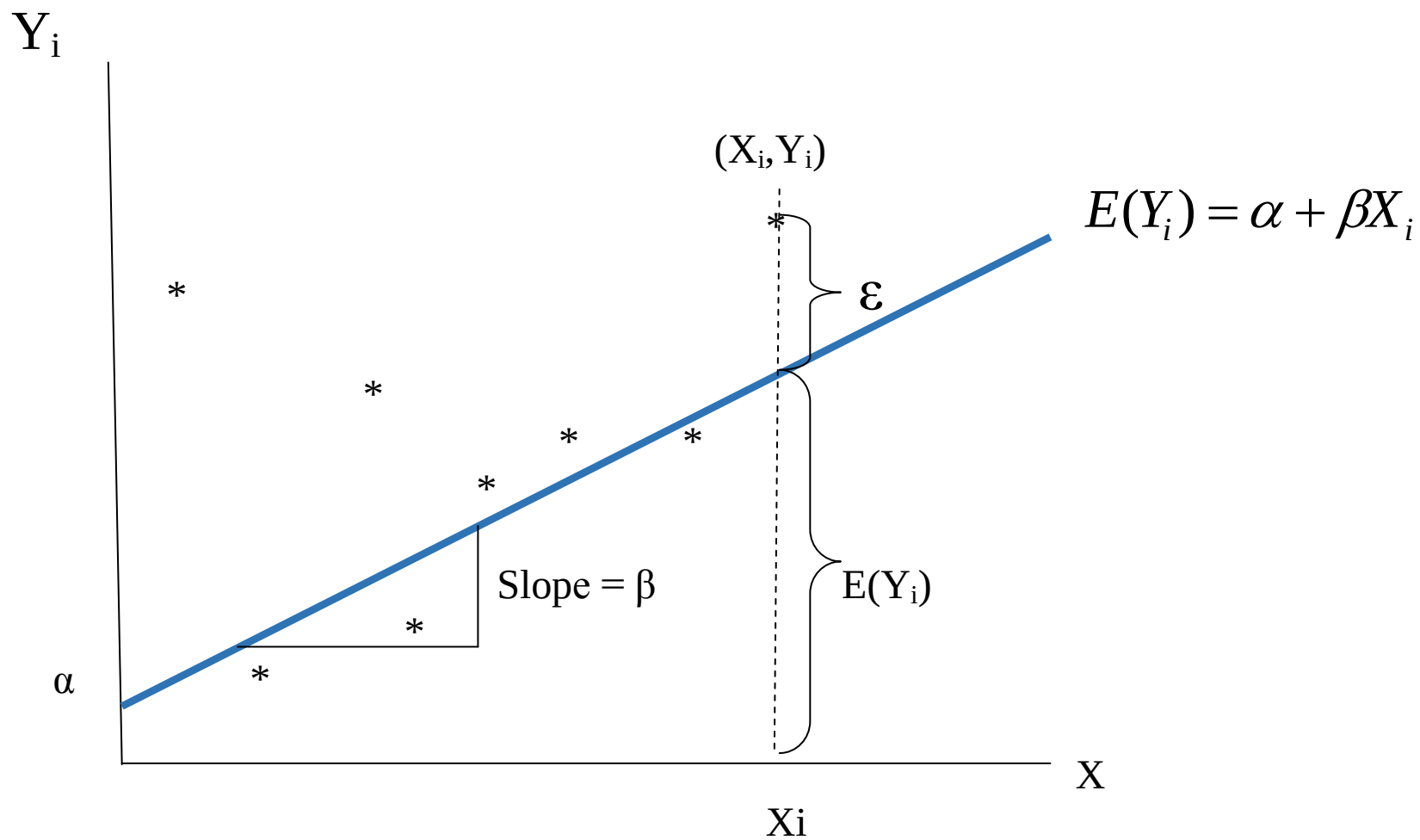
We expect Y , $[E(Y)]$, to change as X changes.

Use subscripts to denote which observation of X and Y we are looking at: $E(Y_i) = \alpha + \beta X_i$ for observation i .

Parameters:

- α is referred to as the (population) Y -intercept term

β denotes the (population) slope term which is the derivative of Y with respect to X : $\left(\frac{\partial y}{\partial x}\right)$.



Example: Let Y be the crime rate and X represent the unemployment rate. Then for the i^{th} observation, the expected value of crime is:

$E(\text{crime}) = \alpha + \beta(\text{Unemployment rate})$ for the i^{th} observation.

$E(Y_i) = \alpha + \beta (X_i)$ for the i^{th} observation.

For each city with the same X (unemployment rate) actual crime (Y) varies because of the “other” factors involved.

For observation i :

$$Y_i = \alpha + \beta X_i + \varepsilon$$



Population regression model

where $\varepsilon = Y_i - E(Y_i)$

which represents the difference between the actual observed Y_i (crime rate) and the population regression line, $E(Y_i)$.

β – the slope – measures the marginal product of a change in the unemployment rate which is implicit in a change in the crime rate.

Another Example:

Let Y_i be the number of years of post secondary education for individual i .

Let X_i be the number of years of post secondary education by one of the parents (either by the father or mother, whoever has more education,) of individual i .

Suppose we know that $\alpha = 1.3$ and $\beta = 0.8$. Then the population regression line is:

$$Y_i = 1.3 + 0.8X_i + \varepsilon$$

Or

$$E(Y_i) = 1.3 + 0.8X_i$$

Notes on this example:

- (i) Suppose the individual 'i' is Jason. One of his parents has four years of post secondary schooling. How many years of post secondary education do we expect Jason to have, *ceteris paribus*?

$X_i = 4$, so:

$$E(Y_i) = 1.3 + 0.8X_i$$

$$E(Y_i) = 1.3 + 0.8(4)$$

$$E(Y_i) = 4.5 \text{ years}$$

On average, we expect an individual with a parent that has 4 years of post secondary schooling to complete 4.5 years of post secondary schooling.

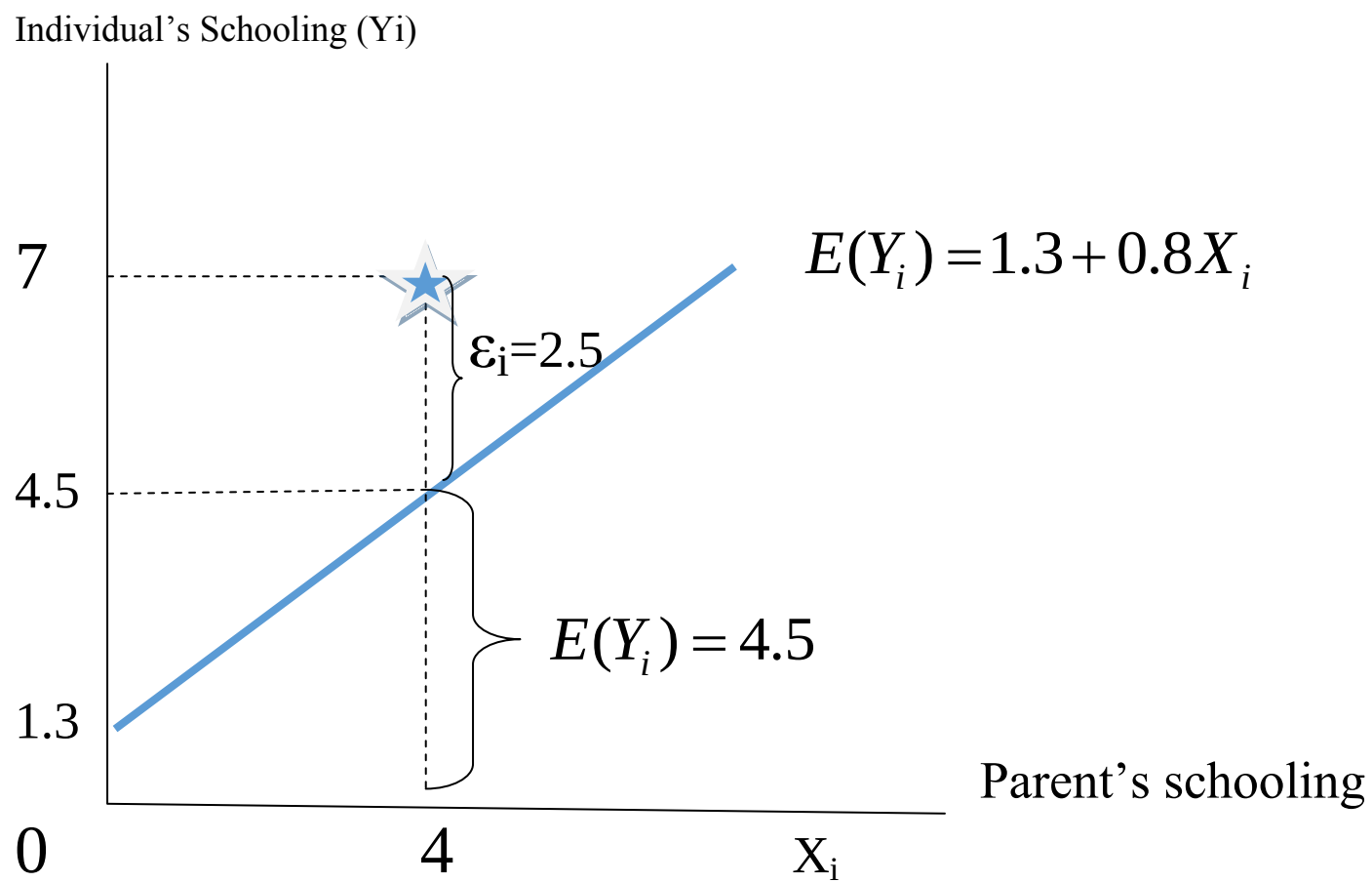
(ii) Other factors affect Y_i for $i = \text{Jason}$.

If the number of years of schooling is in fact 4.5, the random error term, ε_i , equals zero: $\varepsilon_i = 0$.

But, if the actual number of years of school completed is seven years, $\varepsilon_i = 2.5$ for Jason.

Obviously, regression analysis is useful since it allows us to quantify the association or relationship among variables.

We can determine how variable affect each other numerically!



The Sample Regression Model (α and β Unknown)

In the previous examples, we assumed that we knew the value of α and β – the population parameters.

This is unrealistic, and we usually must estimate α and β using sample data.

I.e. this is analogous to using $\bar{X}$ as an estimator of μ , or s^2 as an estimator of σ^2 .

Questions:

- 1) How do we decide which estimators to use for estimating α and β ?
- 2) How can we use sample information?

Example: House Prices

House Price in \$1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Estimated Regression Equation:

$$\widehat{\text{house price}} = 98.25 + 0.1098 (\text{sq.ft.})$$

Predict the price for a house
with 2000 square feet

The sample regression line takes the form:

$$\hat{Y}_i = a + bX_i$$

Where:

- (i) The “hat” indicates a “fitted value” or a calculated value.
- (ii) ‘ a ’ is the estimator of α .
- (iii) ‘ b ’ is the estimator of β .
- (iv) The ‘ i ’ subscript indicate the i^{th} observation and includes all observations in the sample from 1 to n (n =sample size.)

As with the population regression line, $\hat{Y}_i \neq Y_i$, typically.

The difference between $\hat{Y}_i$ and Y_i , the sample error, is denoted e_i : $\Rightarrow e_i = Y_i - \hat{Y}_i$

Substituting:

$$\hat{Y}_i = a + bX_i \Rightarrow e_i = Y_i - a - bX_i$$

Hence, we can express the sample regression model as:

$$Y_i = a + bX_i + e_i$$

Note:

e_i is the sample error.

$\mathcal{E}_i$ is the population error.

They are not the same thing.

e_i is observable.

$\mathcal{E}_i$ is not observable.

$e_i = Y_i - a - bX_i \Rightarrow a, b$ are known sample estimates of α and β .

$\mathcal{E}_i = Y_i - \alpha - \beta X_i \Rightarrow \alpha$ and β are unknown population parameters.

In general, $a \neq \alpha$ and $b \neq \beta$. }

↳ The usual distinction between a sample and a population.

Let e_i (the sample error) be referred to as the residual term.

Let $\mathcal{E}_i$ (the population error) be referred to as the disturbance term.

Let 'a' be our estimator of α derived from a sample.

Let 'b' be our estimator of β derived from a sample.

Since an estimator is a formula or rule, we must determine 'a' and 'b' from the sample information.

Which method should be employed, since there are many possible estimators?

The Method of Least Squares

Recall we have restricted our discussion in this topic to a two-variable (bivariate) model which is linear in parameters.

$$Y_i = \alpha + \beta X_i + \varepsilon$$

This model is linear in variables (Y and X) as well as parameters.

But, it is not necessary to restrict our variables this way.



Simple Linear Regression Model

The population regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Diagram labels and components:

- Dependent Variable: Y_i
- Population Y intercept: β_0
- Population Slope Coefficient: β_1
- Independent Variable: X_i
- Random Error term: ε_i
- Linear component: $\beta_0 + \beta_1 X_i$
- Random Error component: ε_i

Statistics for Managers Using
Microsoft Excel, 4e © 2004
Prentice-Hall, Inc.

Chap 12-10

Example:

Regression analysis can handle:

(i) Log-linear models: $\text{Log}(Y_i) = \alpha + \beta X_i + \varepsilon$

or

(ii) Log-Log models: $\text{Log}(Y_i) = \alpha + \beta \text{LOG}(X_i) + \varepsilon$

These models are linear in parameters, but non-linear in variables.

By redefining the variables: $\begin{matrix} Y_i^* = \log(Y_i) \\ X_i^* = \log(X_i) \end{matrix}$, the models are linear in the new variables:

$$\boxed{\text{Log}(Y_i) = Y_i^* = \alpha + \beta X_i + \varepsilon} \quad \text{and} \quad \boxed{\text{Log}(Y_i) = \alpha + \beta \text{LOG}(X_i) + \varepsilon}$$

respectively.

What is important for the following analysis is that our parameters are linear.

$$Y_i = b_o + b_1 X_1 + b_2 X_2 + \dots + b_n X_n + e_i$$

where

Y_i is the dependent variable (estimated streamflow statistic) for site i ,

b_o to b_n are the regression model coefficients determined in the analysis,

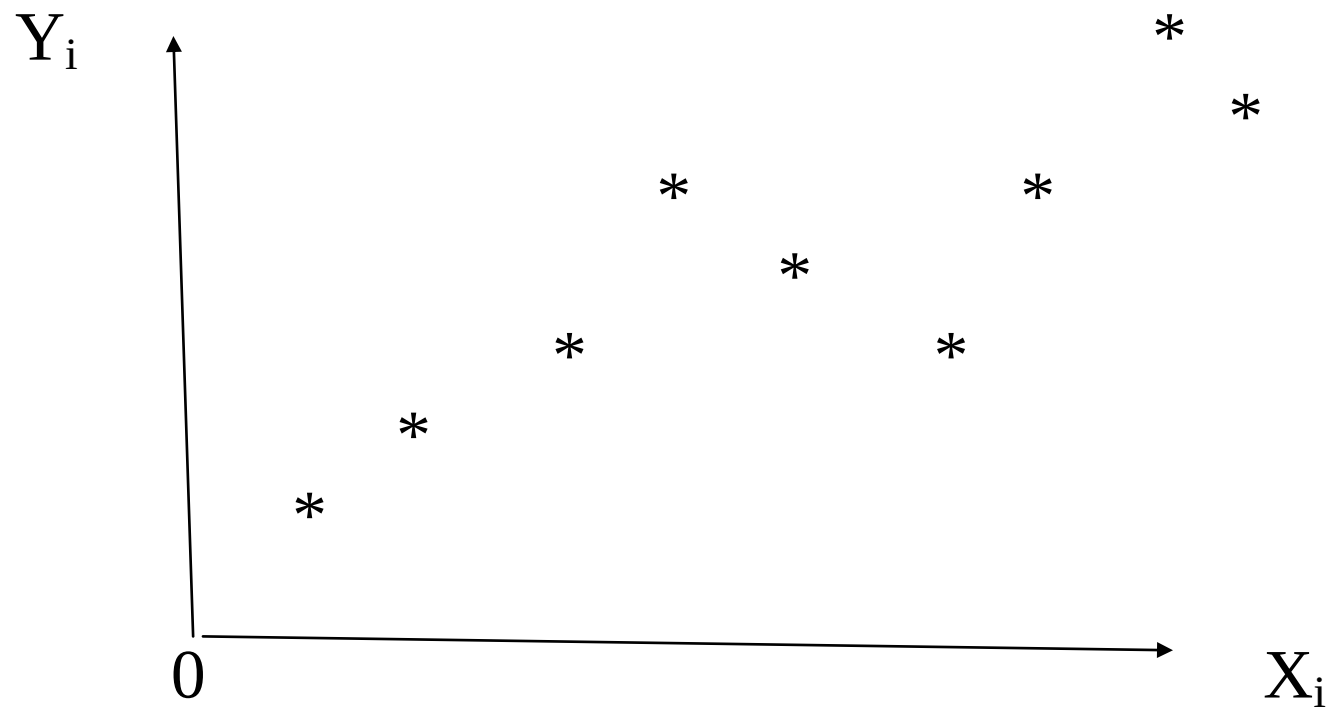
X_1 to X_n are the independent variable (basin characteristics) for site i ,

e_i is the residual error or difference between the observed and estimated dependent variable for site i .

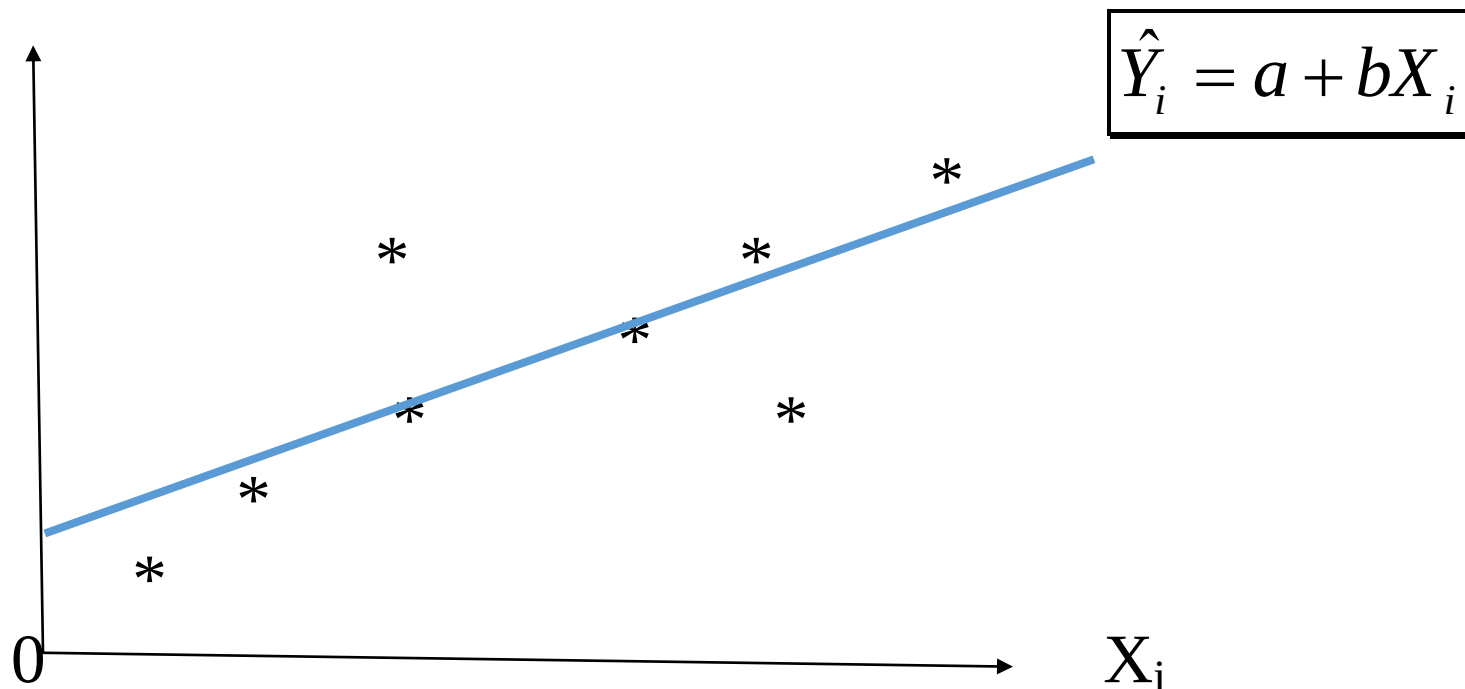
Least Squares:

Suppose we have a sample of size n , with (X_i, Y_i) pair value.

Plot the data points with the Y_i 's on the vertical axis and X_i 's on the horizontal axis.



We can derive a fitted line that passes through these values by measuring α and β by values a and b , such that the fitted line passes close to the observed data.

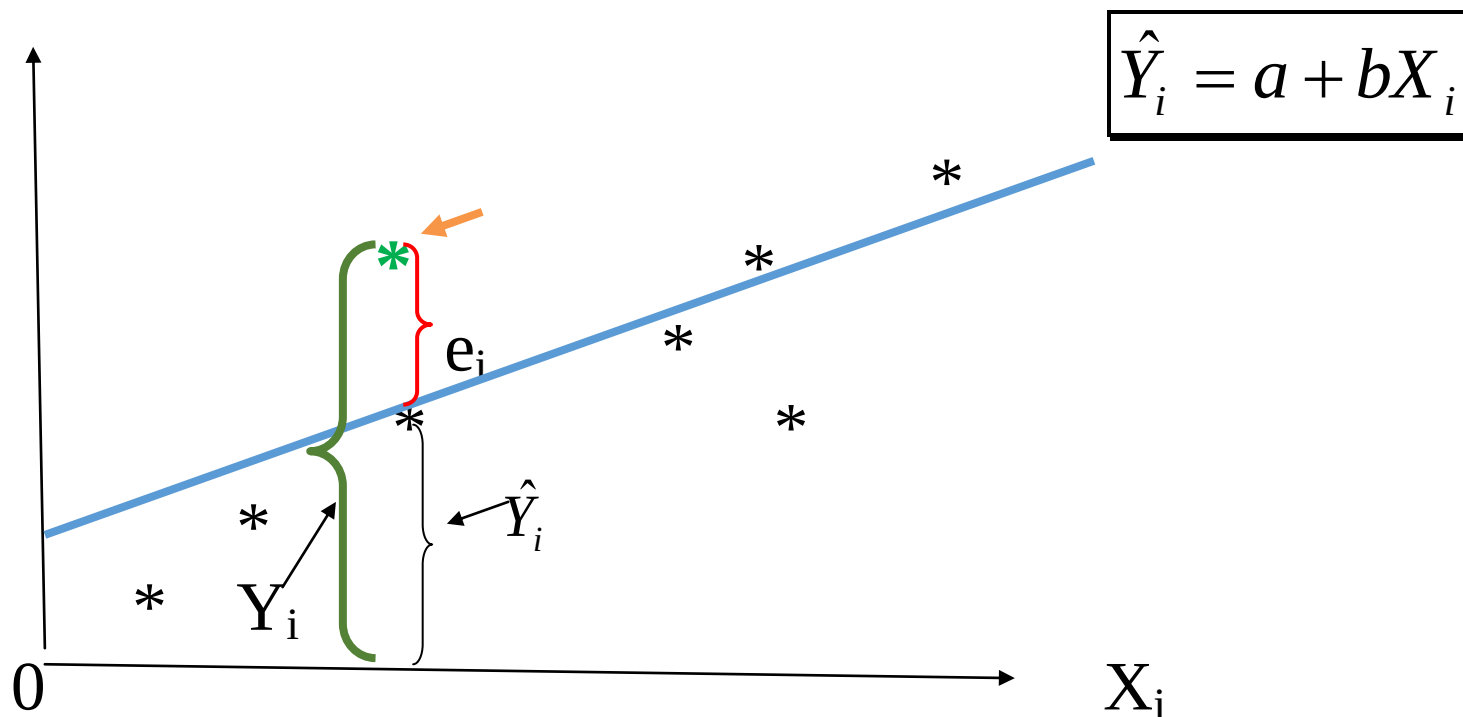
 Y_i 

How should we choose a and b ?

By minimizing the distance between the actual observation and the fitted line:

To illustrate, isolate one point:

Y_i



We want the sum of the residuals to be small as possible.

The problem with choosing a and b so as to minimize $\sum_{i=1}^n e_i$, is offsetting signs.

We could:

(i) Minimize $\sum_{i=1}^n |e_i|$

Or

(ii) Minimize $\sum_{i=1}^n e_i^2 \Leftarrow \text{Least Squares Approach}$

Why Use The Least Squares Approach?

- (I) All observations are given equal weight.
- (II) Our estimators of α and β , $\mathbf{a}$ and $\mathbf{b}$ respectively, have good statistical properties:
 - (i) Unbiased: $E(\mathbf{a}) = \alpha$ and $E(\mathbf{b}) = \beta$.
 - (ii) Efficient: minimum variance estimators out of all linear unbiased estimators (BEST)
- (III) Computationally simple (closed form analytic solution to minimization problem.)
- (IV) Extends trivially to multivariate and non-linear relationships.

Thus, least squares (LS) estimators of α and β are those estimators of a and b which minimize the sum of the squared residuals.

Mathematically, least squares problem is:

$$\text{Min}_{(a,b)} \quad G = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

$$\text{where } \hat{Y}_i = a + bX_i$$

so substituting the above expression into G :

$$G = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - a - bX_i)^2$$

Solve this minimization problem using calculus.

Partially differentiate G with respect to a and b and equate these first order conditions (FOC) to zero. This results in two equations with two unknowns and we solve for a and b .

$$\text{Min}_{(a,b)} \quad G = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - a - bX_i)^2$$

$$(i) \quad \frac{\partial G}{\partial a} = \sum_{i=1}^n \frac{\partial (Y_i - a - bX_i)^2}{\partial a} = \sum_{i=1}^n 2(Y_i - a - bX_i)(-1) = 0 \quad (FOC)$$

$$= 2 \sum_{i=1}^n (Y_i - a - bX_i) = 0$$

which yields the normal equation : $\sum_{i=1}^n (Y_i - a - bX_i) = 0 \Leftrightarrow \sum_{i=1}^n e_i = 0$

Taking the summation sign through:

$$\sum_{i=1}^n (Y_i) - \sum_{i=1}^n a - \sum_{i=1}^n bX_i = 0$$

$$\sum_{i=1}^n (Y_i) - na - b \sum_{i=1}^n X_i = 0$$

rearranging :

$$na = \sum_{i=1}^n (Y_i) - b \sum_{i=1}^n X_i$$

dividing by n :

$$a = \frac{1}{n} \sum_{i=1}^n (Y_i) - b \frac{1}{n} \sum_{i=1}^n X_i$$

$$a = \bar{Y} - b\bar{X} \quad \leftarrow L.S. estimator of \alpha.$$

$$\begin{aligned}
 \text{(ii)} \quad \frac{\partial G}{\partial b} &= \sum_{i=1}^n \frac{\partial (Y_i - a - bX_i)^2}{\partial b} = \sum_{i=1}^n 2(Y_i - a - bX_i)(-X_i) = 0 \quad (FOC) \\
 &= 2 \sum_{i=1}^n (X_i Y_i - aX_i - bX_i^2) = 0
 \end{aligned}$$

which yields the normal equation : $\sum_{i=1}^n (X_i Y_i - aX_i - bX_i^2) = 0$

$$\Rightarrow \sum_{i=1}^n X_i Y_i - a \sum_{i=1}^n X_i - b \sum_{i=1}^n X_i^2 = 0$$

$$\Rightarrow \sum_{i=1}^n X_i Y_i - an\bar{X} - b \sum_{i=1}^n X_i^2 = 0 \quad (*)$$

Since $a = \bar{Y} - b\bar{X}$, we substitute that into (*)

$$\Rightarrow \sum_{i=1}^n X_i Y_i - (\bar{Y} - b\bar{X})n\bar{X} - b \sum_{i=1}^n X_i^2 = 0$$

$$\Rightarrow \sum_{i=1}^n X_i Y_i - (\bar{Y} - b\bar{X})n\bar{X} - b \sum_{i=1}^n X_i^2 = 0$$

$$\Rightarrow \sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y} + nb\bar{X}^2 - b \sum_{i=1}^n X_i^2 = 0$$

$$\Rightarrow \sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y} + b(n\bar{X}^2 - \sum_{i=1}^n X_i^2) = 0$$

Rearranging such that "b" is on the LHS:

$$\Rightarrow b(\sum_{i=1}^n X_i^2 - n\bar{X}^2) = \sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}$$

$$\Rightarrow b = \frac{\sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^n X_i^2 - n\bar{X}^2} \quad \{\text{Least Squares Estimator of } \beta.$$

So, the least squares formulae for a and b for fitting our linear regression model to the data are:

$$a = \bar{Y} - b\bar{X}$$

$$b = \frac{\sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^n X_i^2 - n\bar{X}^2}.$$

Can also write b as:

$$b = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

or

$$b = \frac{\frac{1}{(n-1)} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{(n-1)} \sum_{i=1}^n (X_i - \bar{X})^2}$$

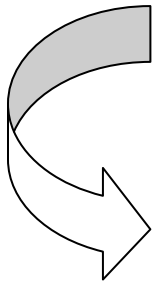
Note:

- (i) The denominator is the sample variance of X_i 's:

$$s_x^2 = \frac{1}{(n-1)} \sum_{i=1}^n (X_i - \bar{X})^2$$

- (ii) The numerator is the sample covariance between X_i and Y_i :

$$s_{XY} = \frac{1}{(n-1)} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$



Reflects how X and Y are related to each other.

Thus, we can also write:

$$b = \frac{S_{XY}}{S_X^2} = \frac{\text{Sample COV between indep. \& dep. variable}}{\text{Sample variance of indep. variable}}$$

- (iii) The sample regression line always passes through the point: $(\bar{X}, \bar{Y})$.

Recall, the fitted regression model is:

$$\hat{Y}_i = a + bX_i = a + b\bar{X}$$

but,

$$a = \bar{Y} - b\bar{X}$$

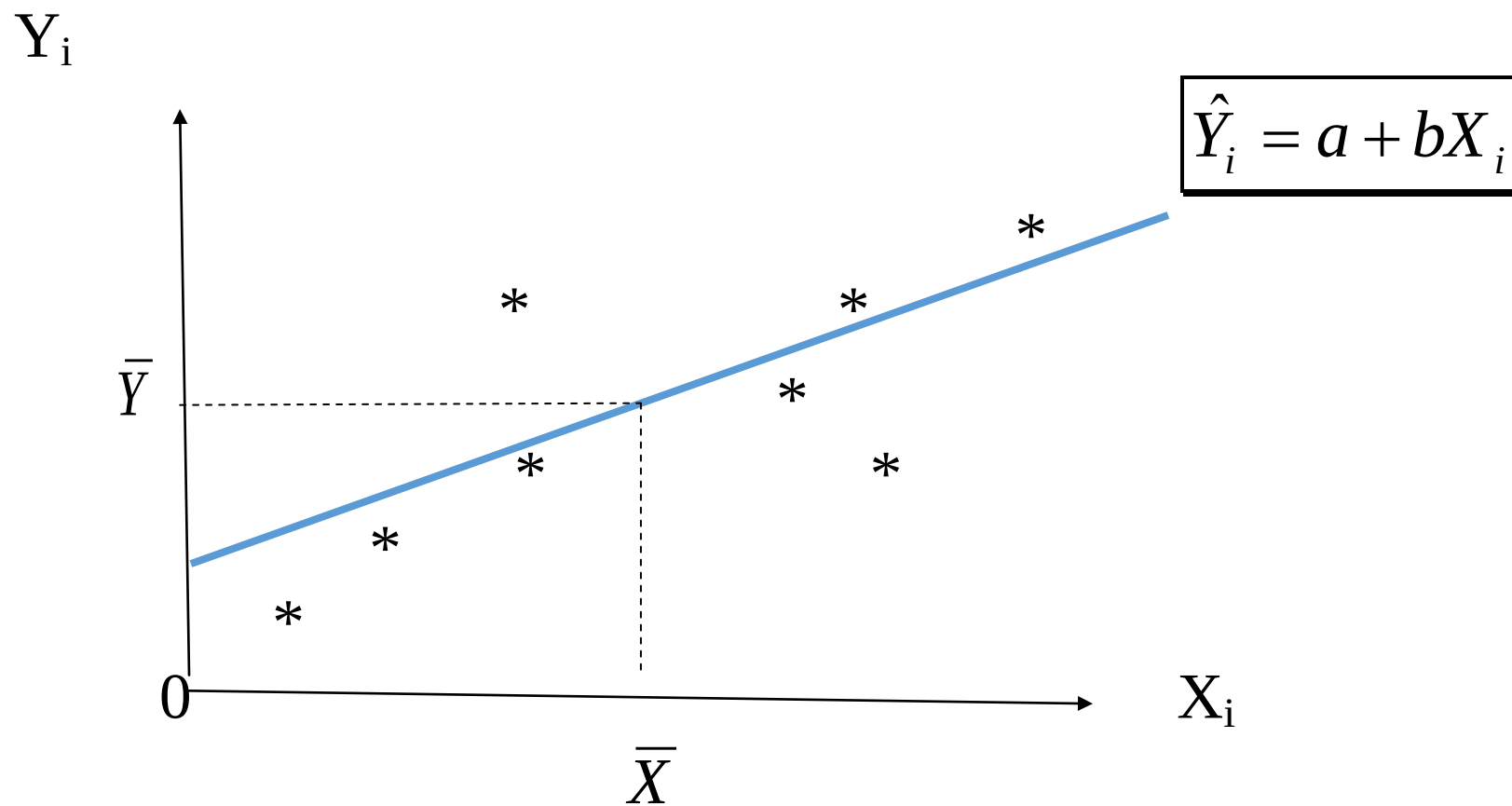
so,

$$\hat{Y}_i = (\bar{Y} - b\bar{X}) + b\bar{X} = \bar{Y}$$

when

$$X_i = \bar{X}.$$

The sample regression line always passes through the point $\bar{X}, \bar{Y}$.



- (iv) With the least squares procedure, the sum of the errors (residuals) equals zero:

$$\sum_{i=1}^n e_i = \sum_{i=1}^n (Y_i - a - bX_i) = 0$$

Recall the normal equation for determining a:

$$\sum_{i=1}^n (Y_i) = na + b \sum_{i=1}^n X_i$$

Equate to zero:

$$\sum_{i=1}^n (Y_i) - \sum_{i=1}^n a - b \sum_{i=1}^n X_i = 0$$

$$\sum_{i=1}^n (Y_i - a - bX_i) = 0 \quad \leftarrow \sum_{i=1}^n e_i = 0$$

$$\sum_{i=1}^n (Y_i - \hat{Y}_i) = 0$$

The sum of the least squares residuals is always zero.

The least squares line is derived such that the line will be situated amongst the data values such that the positive residuals (under-estimates of actual points) always **exactly** cancel out the negative residuals (over-estimates of actual points).

Example: Gas prices (X) and Bike sales (Y)

Sales of Bikes (Y) (millions)	Gas prices (X)
6	0.75
6.5	0.85
8	0.95
10.3	1.05
12	1.15
9	1.25

Sample size, $n=6$.

Estimate the population intercept and slope by least squares estimators a and b to determine the sample

regression model.

$$\hat{Y}_i = a + bX_i$$

Sales of Bikes (Y) (millions)	Gas prices (X) (\$/litre)	X_i^2	$X_i Y_i$
6	0.75	0.5625	4.5
6.5	0.85	0.7225	5.525
8	0.95	0.9025	7.6
10.3	1.05	1.1025	10.815
12	1.15	1.3225	13.8
9	1.25	1.5625	11.25
$\sum Y_i = 51.8$	$\sum X_i = 6$	$\sum X_i^2 = 6.175$	$\sum X_i Y_i = 53.49$

$$\bar{Y} = 51.8 / 6 = 8.633$$

$$\bar{X} = 6 / 6 = 1$$

$$\sum X_i^2 = 6.175$$

$$\sum X_i Y_i = 53.49$$

$$\begin{aligned} b &= \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}}{\sum_{i=1}^n X_i^2 - n \bar{X}^2} \\ &= \frac{53.49 - (6)(1)(8.633)}{6.175 - (6)(1)^2} \\ &= \frac{53.49 - 51.798}{0.175} = \frac{1.692}{0.175} = 9.669 \end{aligned}$$

$$a = \bar{Y} - b \bar{X} = 8.633 - (9.669)(1) = -1.0356$$

$$\hat{Y}_i = -1.0356 + 9.669 X_i$$

For each \$1 increase in the price of gas, bike sales will increase by 9.7 million.

If you put a “hat” on the dependent variable, $(\hat{Y}_i)$, then you do not have to put the residual(e_i) into the regression line. If you do not put the “hat” on the Y_i then you must include the e_i .

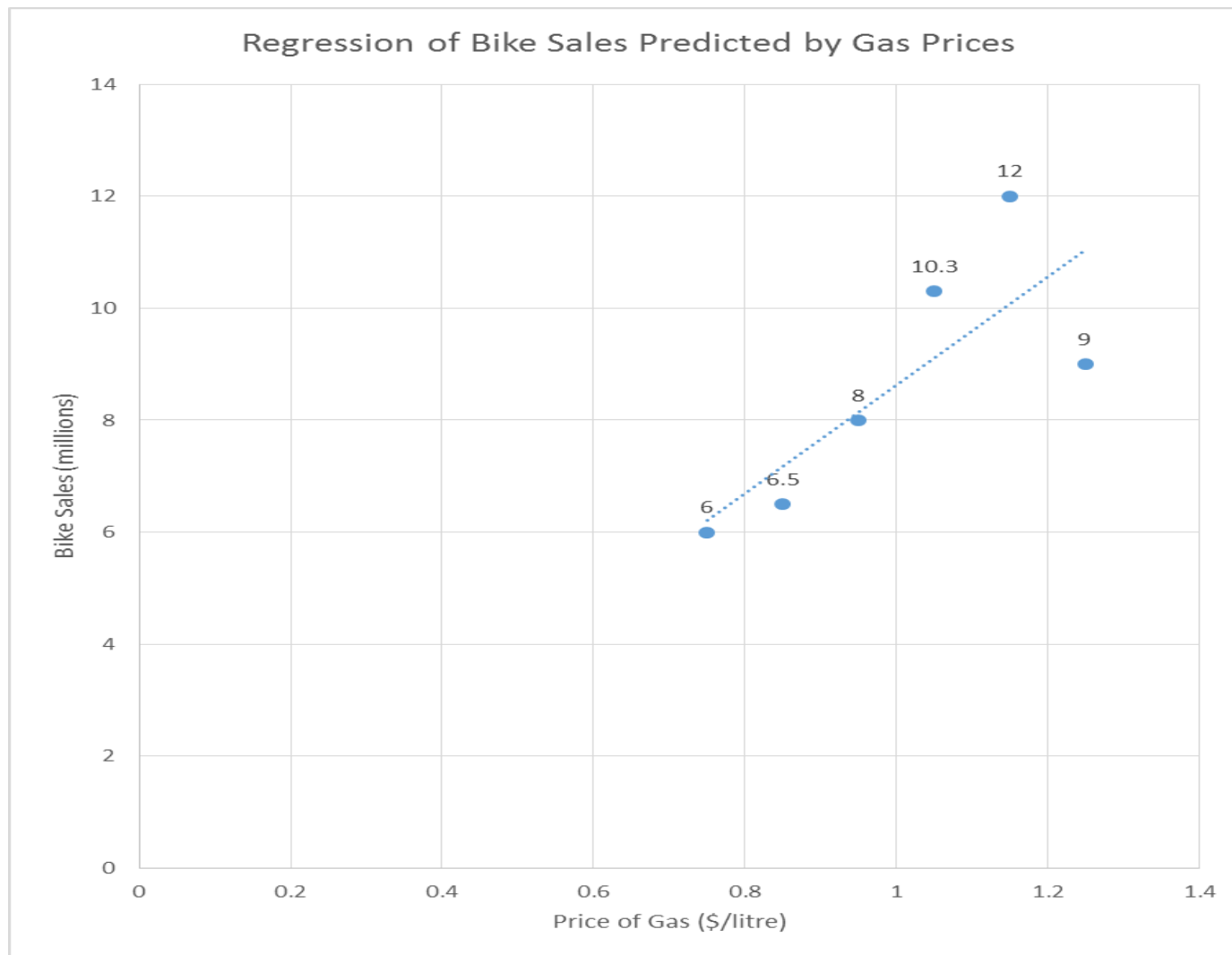
Forecasting:

If price is \$1.50, then:

$$\hat{Y}_i = 1.0356 + 9.669X_i$$

$$\hat{Y}_i = -1.0356 + 9.669(1.50) = 13.4679$$

The best estimate of bike sales is 13.5 million.

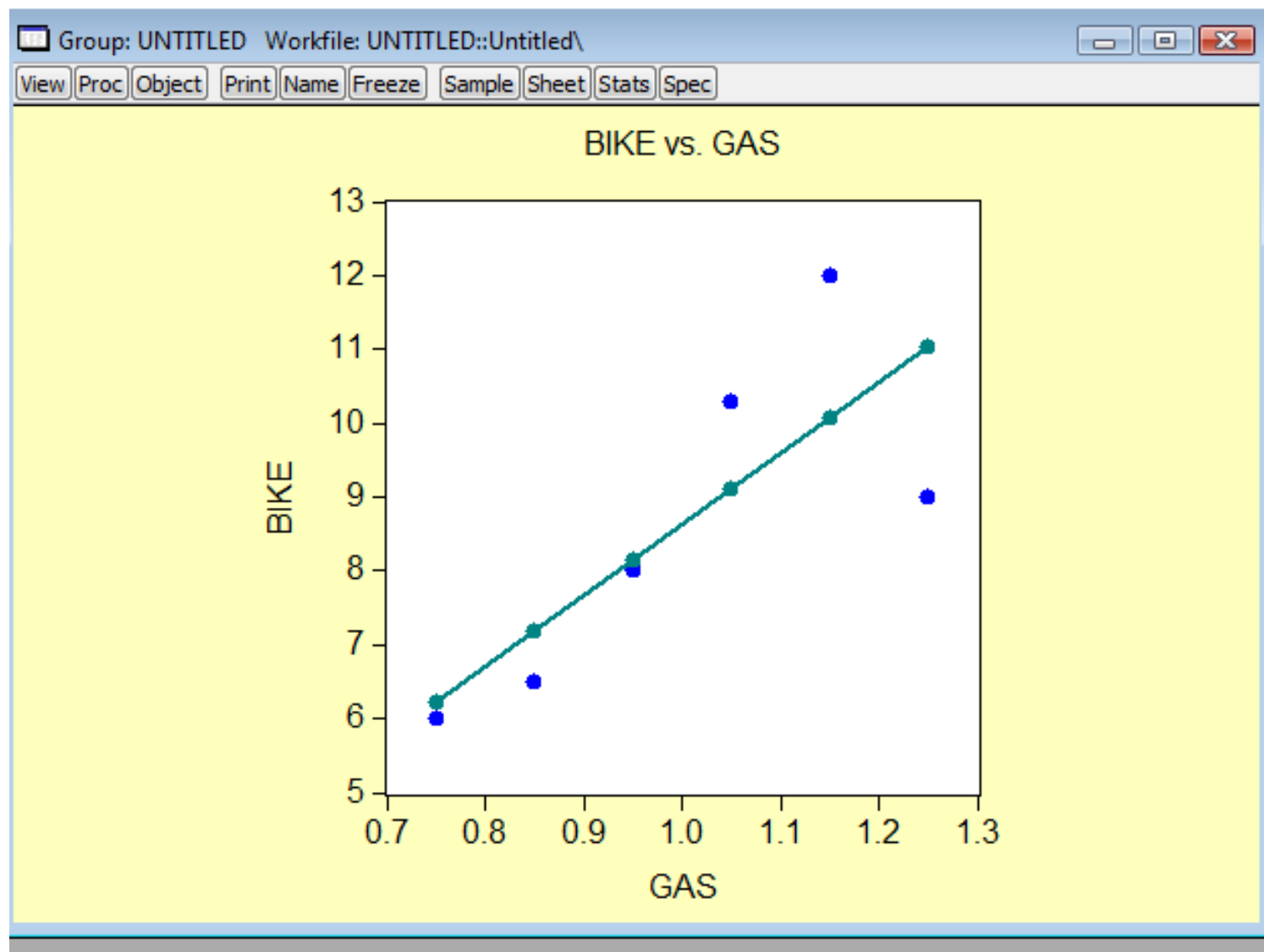


Equation: UNTITLED Workfile: UNTITLED::Untitled\				
View	Proc	Object	Print	Name
Freeze	Estimate	Forecast	Stats	Resids
Dependent Variable: BIKE				
Method: Least Squares				
Date: 08/09/15 Time: 19:02				
Sample: 1 6				
Included observations: 6				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1.023810	3.798302	-0.269544	0.8008
GAS	9.657143	3.744093	2.579301	0.0614
R-squared	0.624512	Mean dependent var	8.633333	
Adjusted R-squared	0.530640	S.D. dependent var	2.286190	
S.E. of regression	1.566266	Akaike info criterion	3.996468	
Sum squared resid	9.812762	Schwarz criterion	3.927055	
Log likelihood	-9.989404	F-statistic	6.652794	
Durbin-Watson stat	1.890266	Prob(F-statistic)	0.061373	

Group: UNTITLED Workfile: UNTITLED::Untitled\

View Proc Object Print Name Freeze Sample Sheet Stats Spec

	BIKE	GAS	
Mean	8.633333	1.000000	
Median	8.500000	1.000000	
Maximum	12.00000	1.250000	
Minimum	6.000000	0.750000	
Std. Dev.	2.286190	0.187083	
Skewness	0.267957	9.29E-17	
Kurtosis	1.802446	1.731429	
Jarque-Bera	0.430335	0.402318	
Probability	0.806406	0.817782	
Sum	51.80000	6.000000	
Sum Sq. Dev.	26.13333	0.175000	
Observations	6	6	



Summary: Regression Model -- Terms and Symbols

Term	Population Symbol	Sample Symbol
Model	$Y_i = \alpha + \beta X_i + \varepsilon$	$Y_i = a + bX_i + e_i$
Error	ε_i	e_i
Slope	β	b
Intercept	α	a
Equation of the line:	$E(Y_i) = \alpha + \beta X_i$	$\hat{Y}_i = a + bX_i$

Concluding Remarks:

- 1) The least squares procedure is simply a curve-fitting technique. It makes no assumptions about the independent variable(s), X , the dependent variable, Y , or the error term.
- 2) We will need to make assumptions about the independent variables, the dependent variable and the error term if we wish to consider how well 'a' and 'b' estimate α and β .
- 3) We will also need to make assumptions if we want to form interval estimates for predicted values of Y or interval estimates for α and β . (Impose some restrictions on how X behaves, or distributed.)

- 4) We will need assumptions if we want to test any hypotheses about the population parameters.

Example: $\begin{matrix} H_0 : \beta = \beta_0 \\ H_a : \beta > \beta_0 \end{matrix}$, where β_0 is some known value.

Special test: $\begin{matrix} H_0 : \beta = 0 \\ H_a : \beta \neq 0 \end{matrix}$. This tests to see if X is significant in explaining Y.

- 5) We can extend least squares principle to non-linear models.

6) We use least squares because it can be shown that under certain assumptions, $\hat{\alpha}$ and $\hat{\beta}$ are unbiased estimators of α and β , respectively. They are also consistent and efficient (i.e. minimum variance).

In other words, under certain assumptions, $\hat{\alpha}$ and $\hat{\beta}$ are the best unbiased estimators of α and β , respectively.



A. A. Mason (1886).

Gauss-Markov Theorem

- With Assumptions 1-7 OLS is:
 1. Unbiased: $E(\hat{\beta}) = \beta$
 2. Minimum Variance – the sampling distribution is as small as possible
 3. Consistent – as $n \rightarrow \infty$, the estimators converge to the true parameters
 - As n increases, variance gets smaller, so each estimate approaches the true value of β .
 4. Normally Distributed. You can apply statistical tests to them.

Assumptions and Estimator Properties

The population model is: $Y_i = \alpha + \beta X_i + \varepsilon$,

where $i=1, 2, 3 \dots N$.

The Five Assumptions:

Gauss-Markov Theorem: Given the five basic assumptions, (below,) the least squares estimators in the linear regression model are best linear unbiased estimators: “BLUE”

Assumption #1:

The random variable ε is statistically independent of X .
That is, $E(\varepsilon, X) = 0$.

Always holds if X is non-stochastic (fixed in repeated sampling.)

Assumption #2:

ε_i is normally distributed for all i .

Assumption #3:

$E(\varepsilon) = 0$. That is, on average the disturbance term is zero for all i .

Assumption #4:

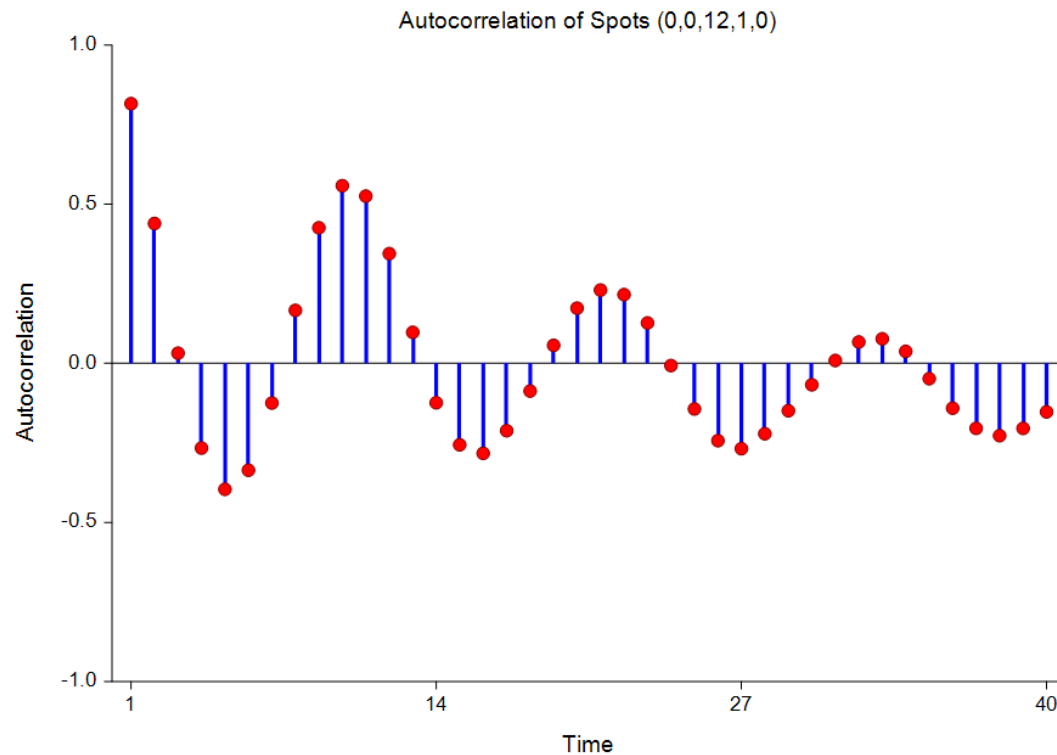
$$\text{Var}(\varepsilon_i) = \sigma_\varepsilon^2 \quad \forall i \text{ and } \forall X_i.$$

This is known as the assumption of homoscedasticity or constant variance, across all observations. Otherwise, heteroscedastic.

Assumption #5:

Any two errors ε_i and ε_j are statistically independent of each other for $i \neq j$. That is, zero covariance (disturbances are uncorrelated). For example, disturbance in observation 1 does not affect observation 2, etc.

If disturbances are correlated across observations, then there exists autocorrelation or serial correlation.



Remarks:

1) From assumptions 1 through 4, we have $\varepsilon_i \sim N(0, \sigma_\varepsilon^2) \forall i$.

2) Y_i is random only because ε_i is random if X is non-random or non-stochastic.

$$3) \quad E(Y_i) = E(\alpha + \beta X_i + \varepsilon_i) = \alpha + \beta X_i$$

Since X_i is non-stochastic and $E(\varepsilon_i)=0$.

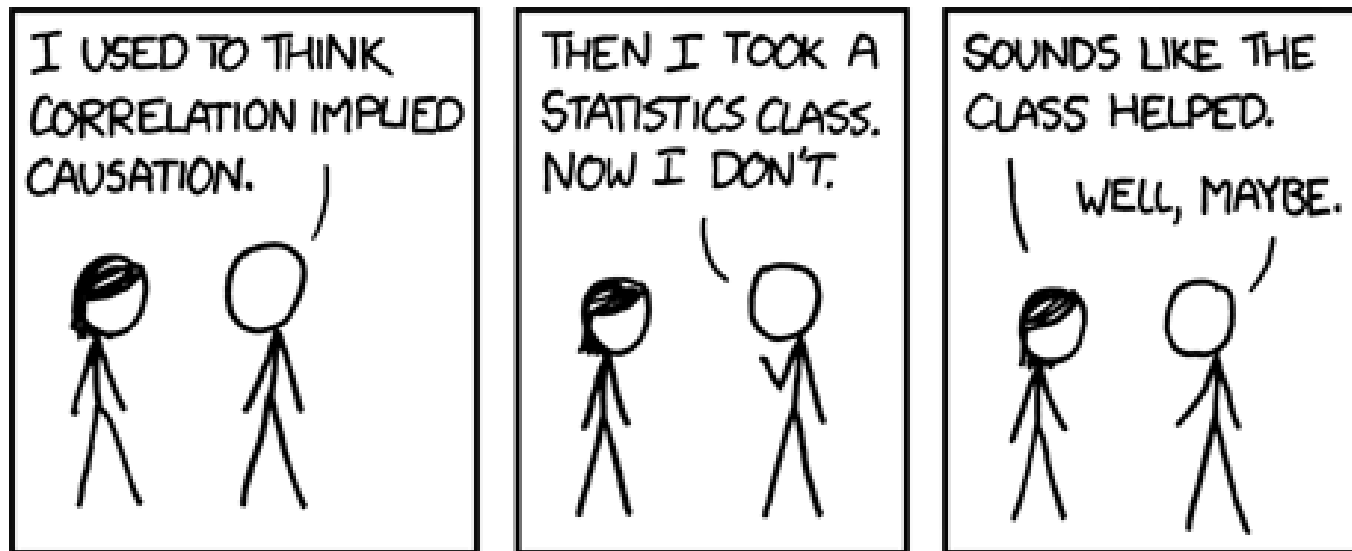
$$Var(Y_i) = Var(\alpha + \beta X_i + \varepsilon_i) = Var(\varepsilon_i) = \sigma_\varepsilon^2.$$

Since α , β , and X are non-stochastic (No variability).

Y_i is a linear function of ε_i and is normal.

That is, from $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, it follows that $Y_i \sim N(\alpha + \beta X_i, \sigma_\varepsilon^2)$.

- 4) As ε_i and ε_j are statistically independent of each other for $i \neq j$, so to are Y_i and Y_j for $i \neq j$.



Measures of Goodness of Fit

This section will explore two measures of goodness of fit:

- 1) The standard error of the estimate: which is the measure of the absolute fit of the sample points to the sample regression line.
- 2) The **coefficient of determination** R^2 : which is a measure of the relative goodness of fit of a sample regression line.

Notation:

Total deviation of Y: the difference between the observed value of Y_i and the mean of the Y-values, $\bar{Y}$.

$(Y_i - \bar{Y})$ (By adding and subtracting $\hat{Y}$, we can separate the deviation into two parts:)

The total deviation can be expressed as the sum of two other deviations:

$$(Y_i - \hat{Y}_i) \text{ and } (\hat{Y}_i - \bar{Y}).$$

The first deviation expression is the residual e_i : $(Y_i - \hat{Y}_i) = e_i$.

Since the e_i 's are random, the term $(Y_i - \hat{Y}_i) = e_i$ is referred to as the unexplained deviation.

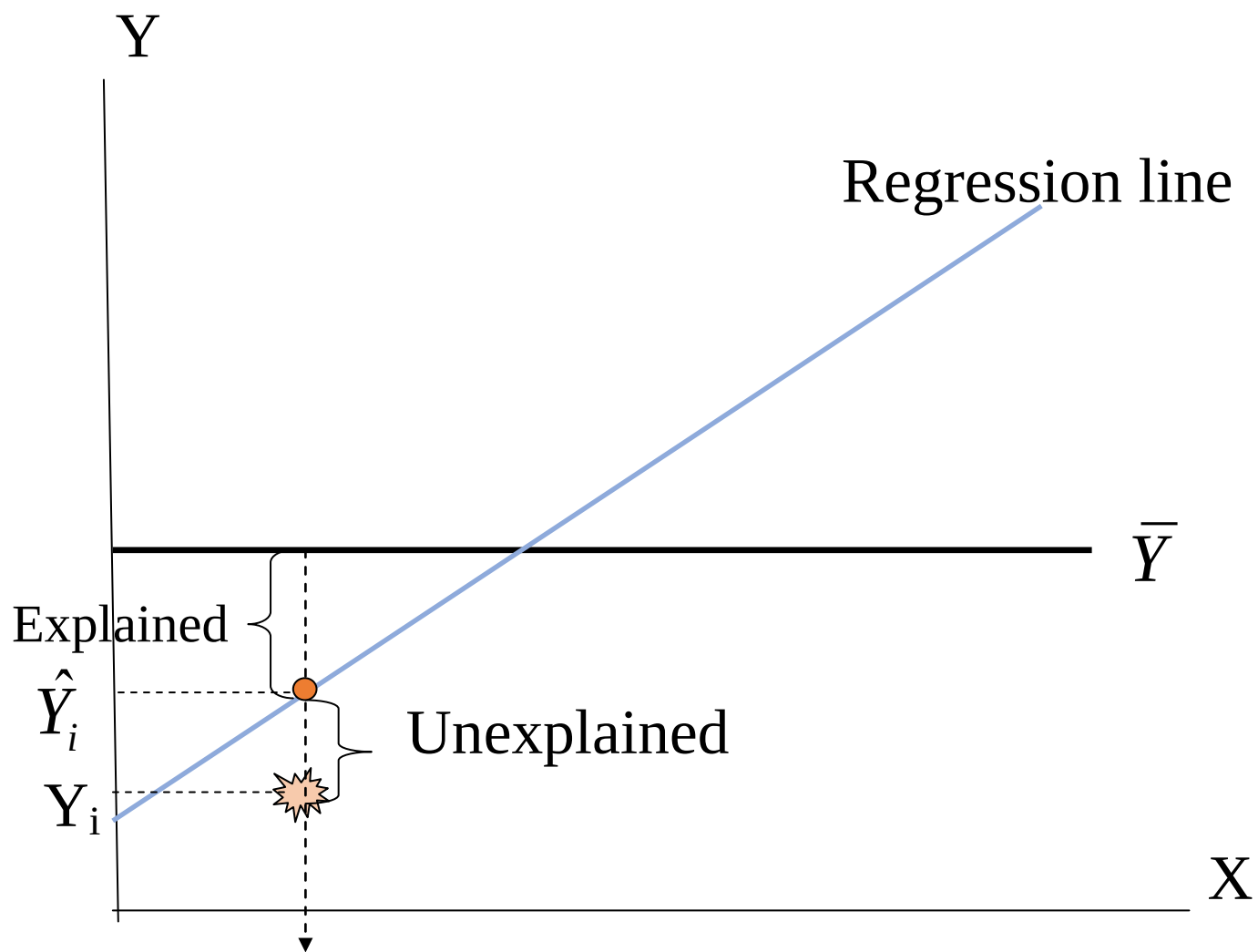
(We cannot determine why $(Y_i \neq \hat{Y}_i)$).

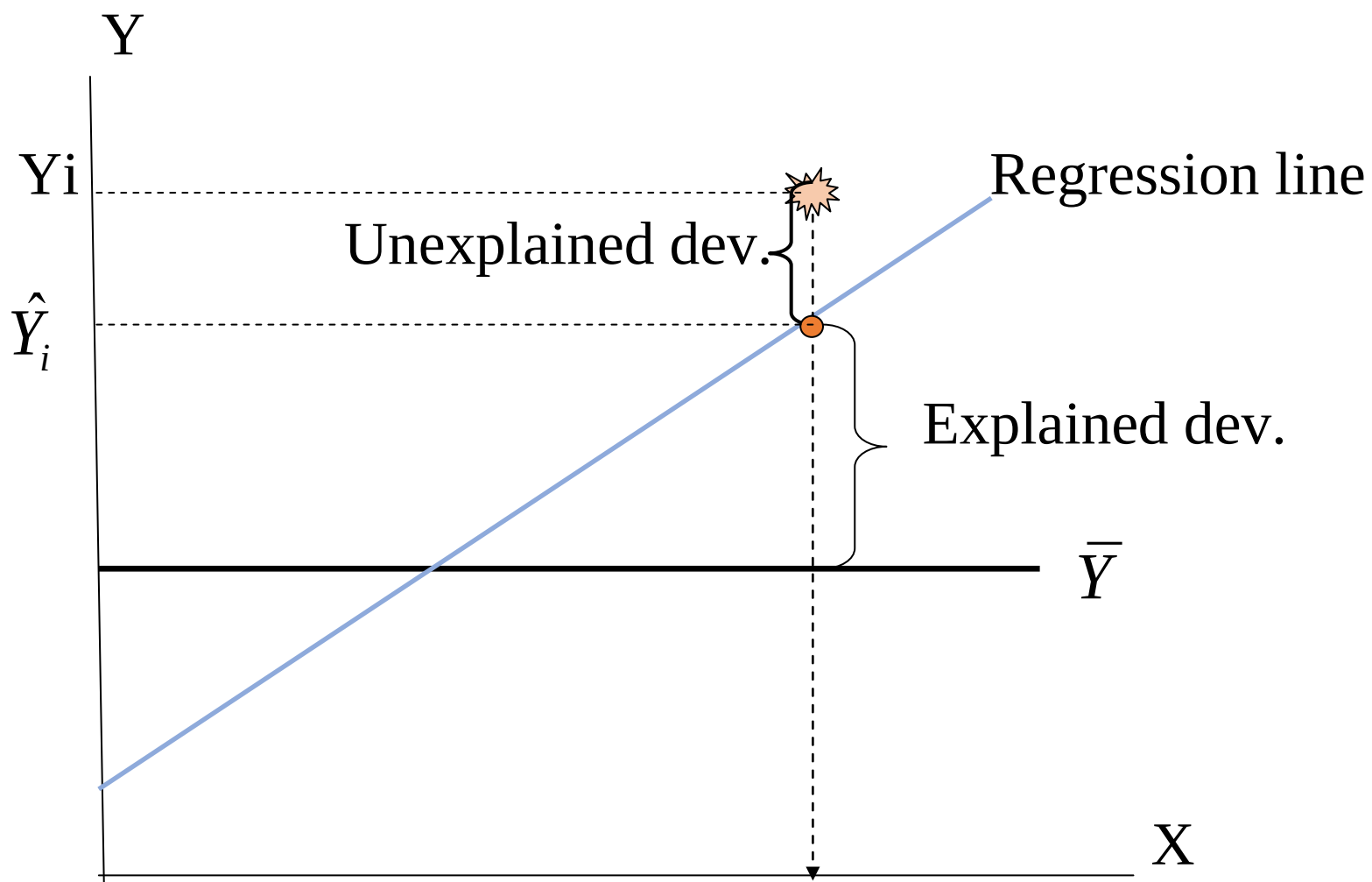
The second deviation can be explained by the regression line.

$(\hat{Y}_i - \bar{Y})$ is the explained deviation, because it is possible to explain $\hat{Y}_i$ differs from $\bar{Y}$ because X_i differs from $\bar{X}$.

Total Deviation = Unexplained Deviation+ Explained Deviation

$$(Y_i - \bar{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$





The two parts of the total deviation are independent.

Hence, we can square each deviation and sum over all 'n' observations as follows:

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

Breaking into three parts:

$$\sum_{i=1}^n (Y_i - \bar{Y})^2$$

⇐ The total variation

⇐ The total sum of squares

⇐ SST

$$\sum_{i=1}^n (Y_i - \hat{Y})^2$$

⇐ The unexplained variation

⇐ The sum of squares error

⇐ SSE = $\sum_{i=1}^n e_i^2$

$$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \Leftarrow \text{The explained variation due to the regression.}$$

$$\Leftarrow \text{The sum of squares regression}$$

$$\Leftarrow \text{SSR}$$

Total deviation = Unexplained deviation + Explained deviation

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

$$\text{SST} = \text{SSE} + \text{SSR}$$

Why are we dissecting the total variation into two components?

We can now determine the goodness of fit of a regression in terms of the size of SSE.

If the fit is perfect, $SSE = 0$.

$$\boxed{(Y_i = \hat{Y}_i) \quad \forall i}$$

If the fit is not perfect $SSE \neq 0$

Calculation of SST, SSR and SSE

The best way to illustrate the calculation of these measures is through the example from earlier:

Sales of Bikes (Y) (millions)	Gas prices (X) (\$/litre)	X_i^2	$X_i Y_i$
6	0.75	0.5625	4.5
6.5	0.85	0.7225	5.525
8	0.95	0.9025	7.6
10.3	1.05	1.1025	10.815
12	1.15	1.3225	13.8
9	1.25	1.5625	11.25
$\sum Y_i = 51.8$	$\sum X_i = 6$	$\sum X_i^2 = 6.175$	$\sum X_i Y_i = 53.49$

$$\bar{Y} = 51.8 / 6 = 8.633$$

$$\bar{X} = 6 / 6 = 1$$

$$\sum X_i^2 = 6.175$$

$$\sum X_i Y_i = 53.49$$

$$b = \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}}{\sum_{i=1}^n X_i^2 - n \bar{X}^2}$$

$$= \frac{53.49 - (6)(1)(8.633)}{6.175 - (6)(1)^2}$$

$$= \frac{53.49 - 51.798}{0.175} = \frac{1.692}{0.175} = 9.669$$

$$a = \bar{Y} - b \bar{X} = 8.633 - (9.669)(1) = -1.0356$$

$$\hat{Y}_i = -1.0356 + 9.669 X_i$$

Sales of Bikes (Y) (millions)	Gas prices (X) (\$/litre)	$(Y_i - \bar{Y})^2$ Total deviation	$e_i^2 = (Y_i - \hat{Y}_i)^2$ Unexplained deviation	$(\hat{Y}_i - \bar{Y})^2$ Explained deviation	$\hat{Y}_i$
6	0.75	6.932689	0.046720823	5.841163922	6.21615
6.5	0.85	4.549689	0.466557303	2.102355003	7.18305
8	0.95	0.400689	0.022485002	0.233337303	8.14995
10.3	1.05	2.778889	1.399843923	0.234110823	9.11685
12	1.15	11.336689	3.672014063	2.104675563	10.08375
9	1.25	0.134689	4.205165423	5.845031523	11.05065
$\sum Y_i = 51.8$	$\sum X_i = 6$	SST=26.1333	SSE=9.81279	SSR=16.3607	

We should find that $SST = SSR + SSE$

A more convenient way to determine SST is the following:

$$SST = \sum_{i=1}^n Y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n Y_i \right)^2$$

The amount of variation **explained** by the regression equation can be determined without first calculation SSE. The method eliminates the calculation of each estimated value ($\hat{Y}_i$).

Explained variation:

$$SSR = b \left(\sum (X_i - \bar{X})(Y_i - \bar{Y}) \right) \\ = b(\text{covariation of } X \text{ and } Y)$$

Once SST and SSR are found, the unexplained variation, SSE, is found by subtraction:

$$SSE = SST - SSR.$$

How Good is the Fit In Regression Analysis?

(I) Standard Error of the Estimate

$$S_e = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2} = \sqrt{\frac{SSE}{n-2}}$$

The number of degrees of freedom in this case is $n-2$ because two sample statistics (a and b) must be calculated before the value of $\hat{Y}_i$ can be computed.

The information from this statistic provides an indication of the fit of the regression model. It also indicates the size of the residuals that result.

The units of S_e are always the same as those for the variable Y since ' e ' is a leftover (residual) component of Y .

Problem: (units) Is S_e large or small?

To determine whether S_e is large or small, we compare it to the average size of Y .

Using a coefficient of variation type measure, we can define the typical percentage error in predicting Y using the sample data and the regression line.

Coefficient of Variation For the Regression Residuals:

First compute the S_e :

$$S_e = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{9.81279}{6-2}} = \sqrt{2.453197} = 1.5663$$

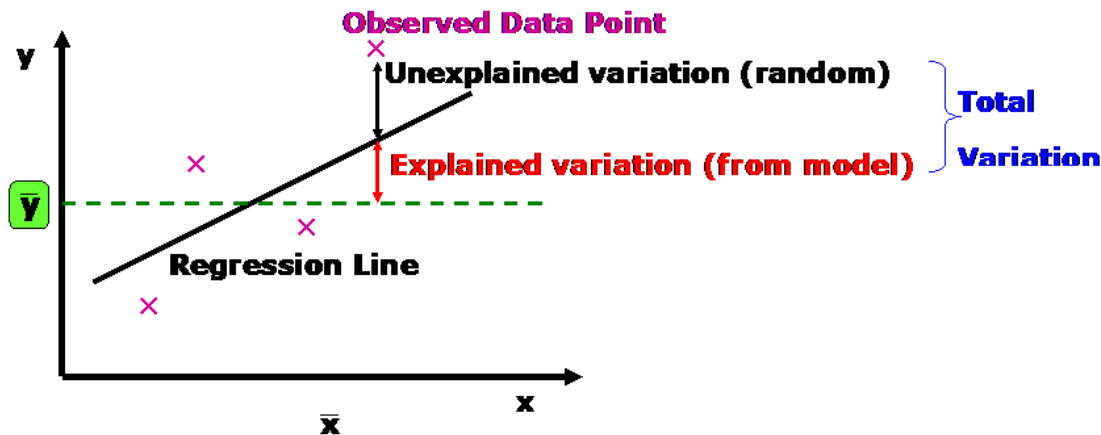
$$C.V. \text{ residuals} := \left(\frac{S_e}{\bar{Y}} \right) \times 100 = \left(\frac{1.5663}{8.633} \right) \times 100 = 18.1429$$

For this example, the C.V. residuals = 18.14%. In predicting the number of bike sales, the value will be incorrect by 18.14% on the average.

(II) Coefficient of Determination

The second measure of goodness of fit is the coefficient of determination. (R^2)

It facilitates in the interpretation of the relative amount of variation that has been **explained** by the sample regression line.



Deriving R^2 :

Recall: $SST = SSE + SSR$

$$\frac{SST}{SST} = \frac{SSE}{SST} + \frac{SSR}{SST}$$

If we divide this expression by SST: $1 = \frac{SSE}{SST} + \frac{SSR}{SST}$.

These two ratios must sum to one and are mutually exclusive.

Recall, SSE is the unexplained variation in Y.

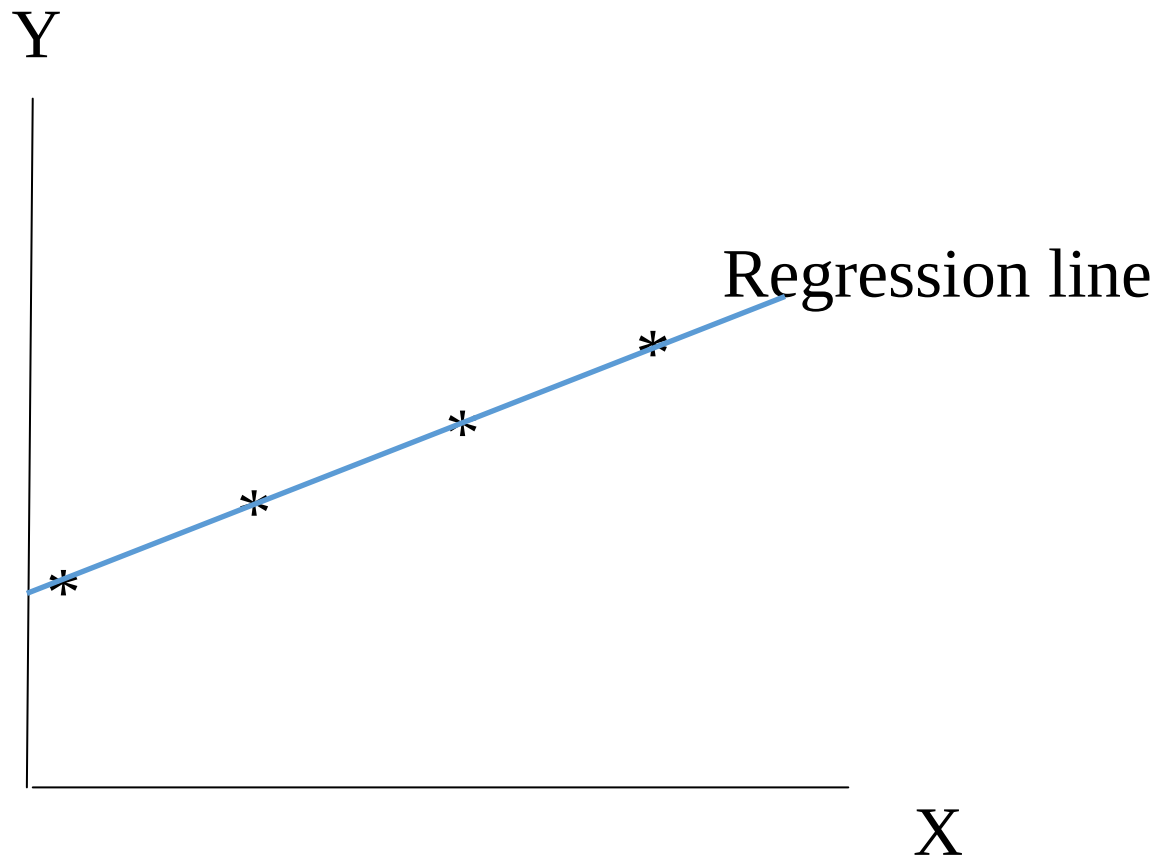
The ratio $\frac{SSE}{SST}$ represents the proportion of total variation that is unexplained by the regression relation.

The ratio $\frac{SSR}{SST}$ represents the relative measure of goodness of fit and is called the coefficient of determination.

It measures the *proportion* of total variation that is explained by the regression line.

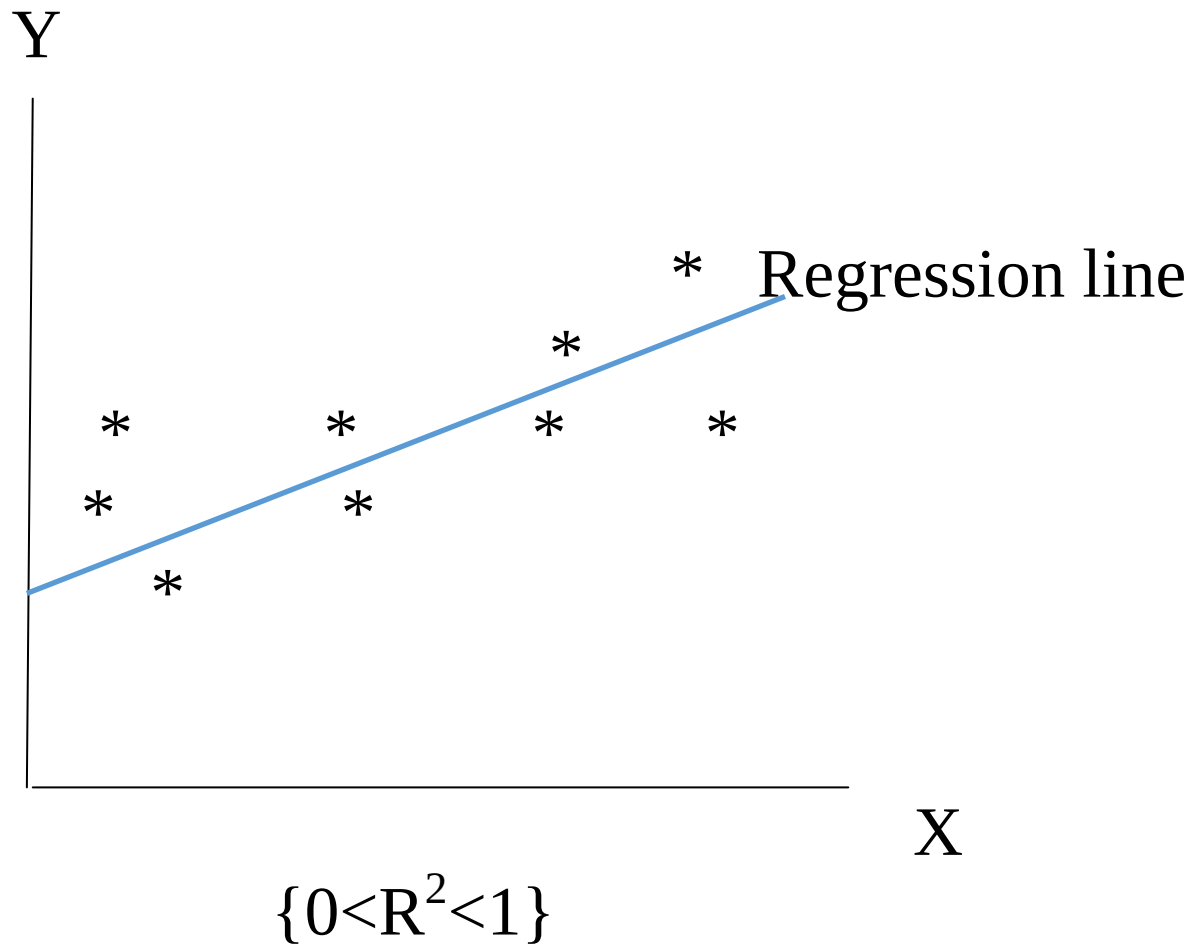
$$R^2 = \frac{SSR}{SST} = \frac{\text{Variation explained}}{\text{Total variation}}.$$

If the regression line perfectly fits all the sample points, $SSR=SST$ and the coefficient of determination would achieve its maximum value of 1: $R^2=1$.

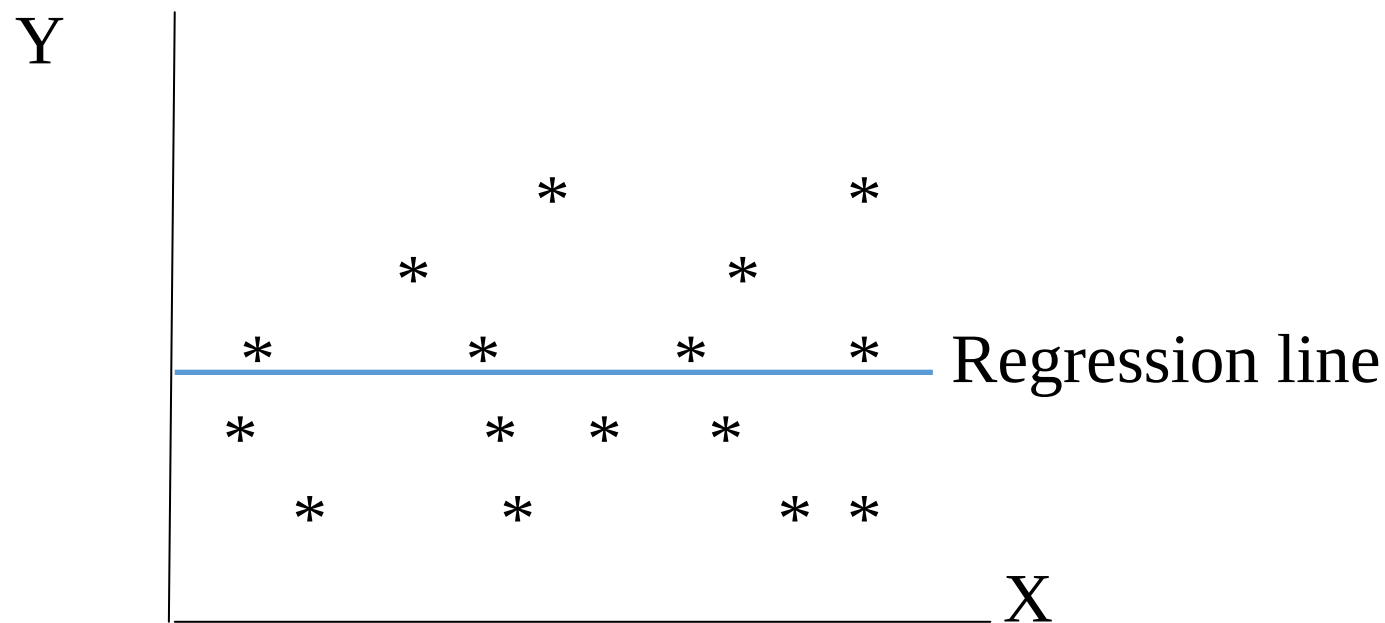


Perfect fit: $R^2 = 1$; $\hat{Y}_i = Y_i$

As the degree of fit becomes less accurate and less of the variation in Y is explained by their relation with X , R^2 decreases.



The lowest value of R^2 is zero, which will occur whenever $SSR=0$ and $SSE=SST$. This occurs when $\hat{Y}_i = \bar{Y}$ for all observations.



There is no systematic relation between Y and X:

$$R^2=0; \text{ $\hat{Y}_i = \bar{Y}$ } ; b=0.$$

The regression line is horizontal with a slope =0.
Hence, X has no effect in explaining Y.

Example: Determine the coefficient of determination for the bike/gas example:

$$SSR=16.3607$$

$$SST=26.1333$$

Hence,

$$R^2 = \frac{SSR}{SST} = \frac{16.3607}{26.1333} = 0.6260.$$

The interpretation of this result is that 62.5% of the total sample variation in bike sales is explained by the linear relation of gas prices in \$.

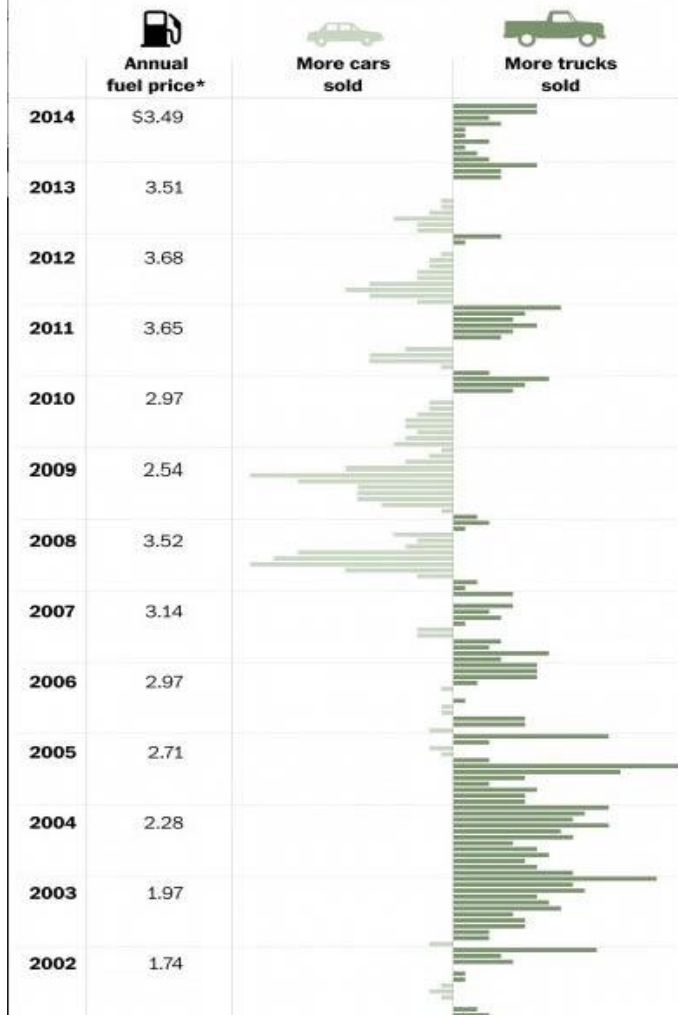
The remaining 37.4% of the variation in bike sales is still unexplained.

Most likely some other variable has been omitted from the regression model that could explain some additional portion of the variation. (“Multiple regression model” is required.)

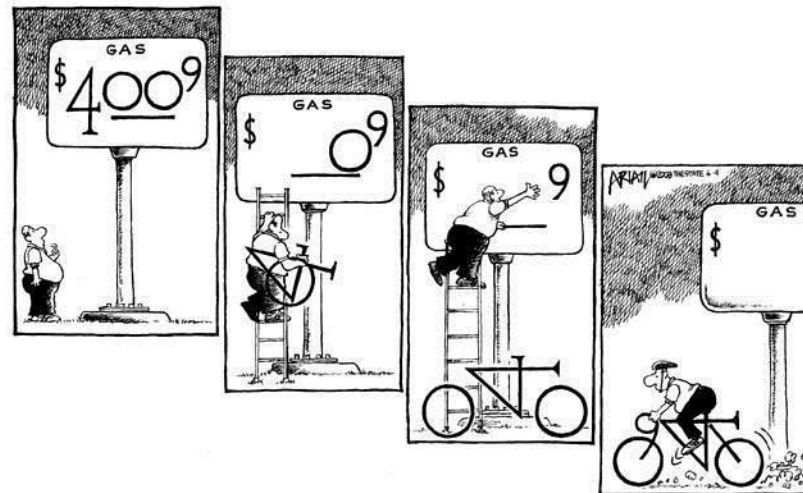


Fuel prices' impact on truck sales

U.S. fuel prices fell below \$3 per gallon in October for the first time in four years. High fuel prices' effect on truck sales mostly wore off by 2013, though.



* Inflation adjusted.
Sources: Edmunds, AAA
Darla Cameron/THE WASHINGTON POST



Correlation Analysis

The measurement of how well two or more variables vary together is called **correlation analysis**.

One measure of the population relationship between two random variables is the population covariance:

$$\text{Cov}(X, Y) = E[(X_i - \bar{X})(Y_i - \bar{Y})]$$

Usually this measure is not a good indicator of the relative strength of the relationship between two variables because its magnitude depends on the **units** used to measure the variables.

Hence, it is necessary to standardize the covariance of two variables in order to have a good measure of fit. (Unit-less measure.)

Standardization is carried out by dividing the covariance of X and Y by the standard deviation of X times the standard deviation of Y:

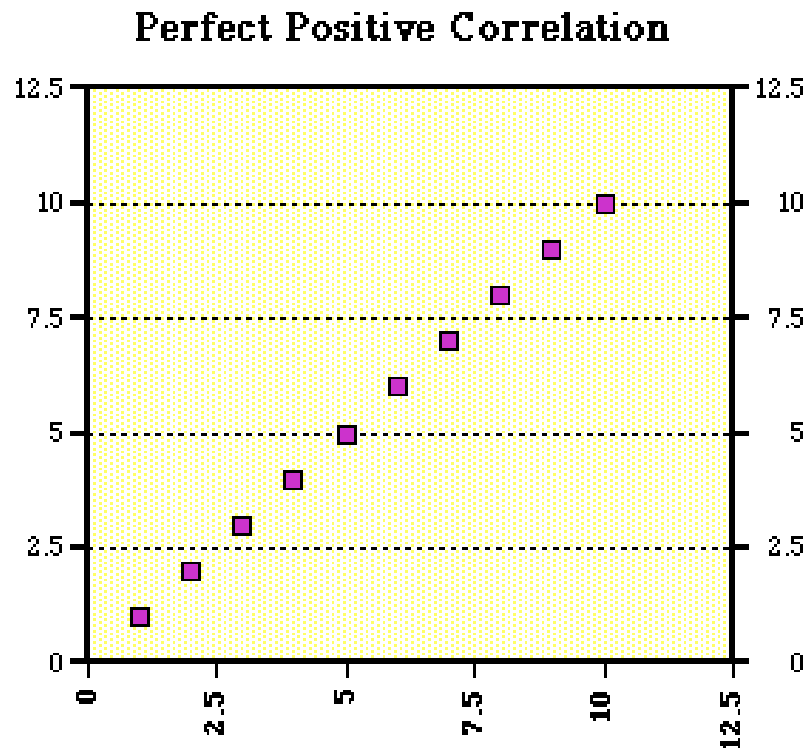
This is referred to as:

Population Correlation Coefficient :

$$\rho = \frac{\text{Covariance of X and Y}}{(\text{STD.dev. of X}) \times (\text{STD.dev. of Y})} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

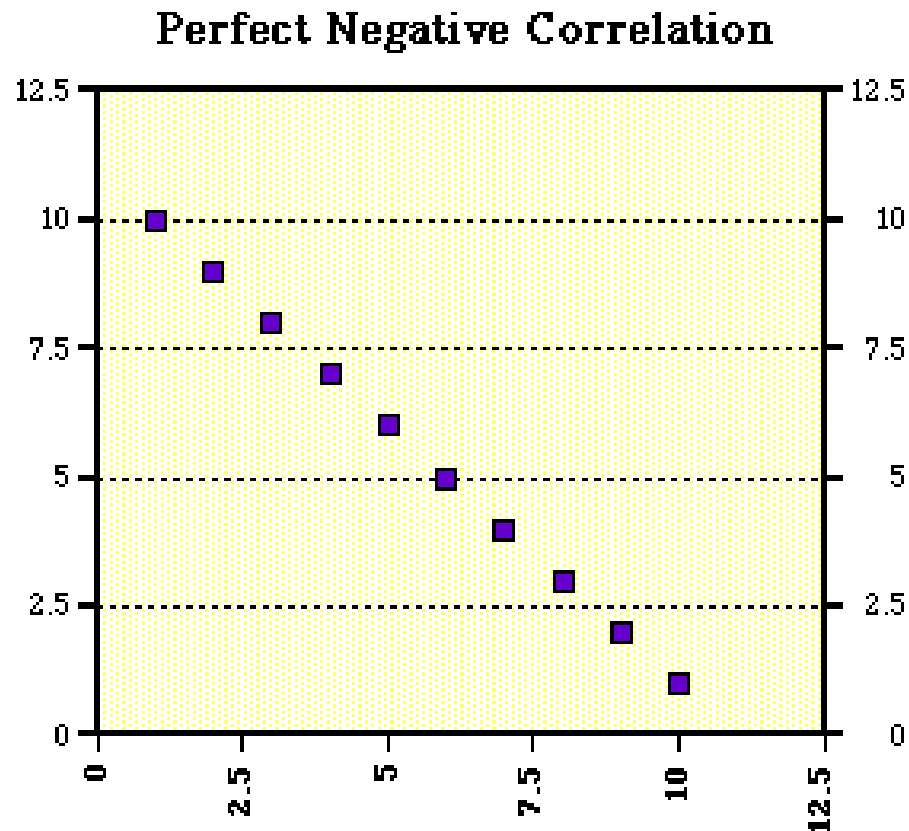
Three Special Cases:

Case #1: Perfect Positive Correlation: $\rho = 1$.



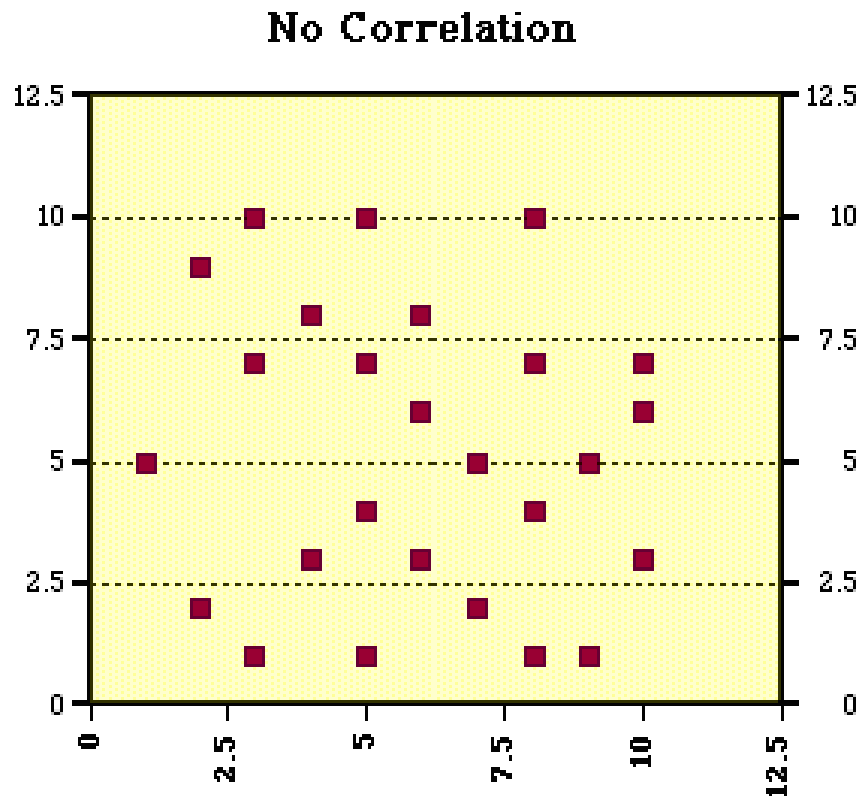
All values of X and Y fall on a positively sloped straight line:

Case #2: Perfect Negative Correlation: $\rho = -1$.



All values of X and Y fall on a negatively sloped straight line.

Case #3: Zero Correlation: $\rho=0$.



If X and Y are not linearly related, they are independent random variables. Then, the value of the correlation coefficient will be zero, since $\text{Cov}(X, Y) = 0$.

Thus, ρ measures the strength of the linear association between X and Y .

Values of ρ close to zero indicate a weak relation;

Values close to $+1$ indicate a strong positive relation;

Values close to -1 indicate a strong negative correlation.

Sample Correlation Coefficient

We employ sample data to estimate the population parameter ρ . The sample correlation coefficient is denoted by the letter ' r '.

We determine the value of ' r ' in the same manner as ρ , except that we substitute for each population parameter its best estimate based on the sample data.

Sample Covariance : S_{XY} :

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

The Sample Correlation Coefficient:

$$\begin{aligned} r &= \frac{S_{XY}}{S_X S_Y} = \frac{\text{Cov.sample values of X and Y}}{\left(\begin{array}{c} \text{Sample standard} \\ \text{deviation of X} \end{array} \right) \left(\begin{array}{c} \text{Sample standard} \\ \text{deviation of Y} \end{array} \right)} \\ &= \frac{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2} \times \sqrt{\frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2}} \end{aligned}$$

divide out the degrees of freedom :

$$r = \frac{SC_{XY}}{\sqrt{SS_X} \sqrt{SS_Y}} = \frac{\text{Sample covariation}}{(\text{Sample Variations})}$$

The Connection Between Correlation and Regression

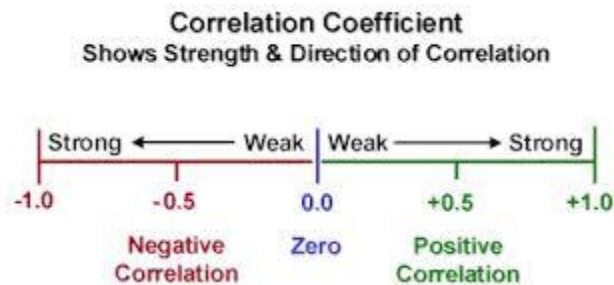
- 1) The correlation coefficient and the coefficient of determination measure the strength of the association between X and Y.

For a model with only one explanatory variable, the measure of R^2 is the square of the correlation measure 'r', as they measure the same relationship.

When there is a weak relationship, both measures are close to zero.

A **strong** relation is indicated between the two variables, X and Y, when r approaches -1 or $+1$.

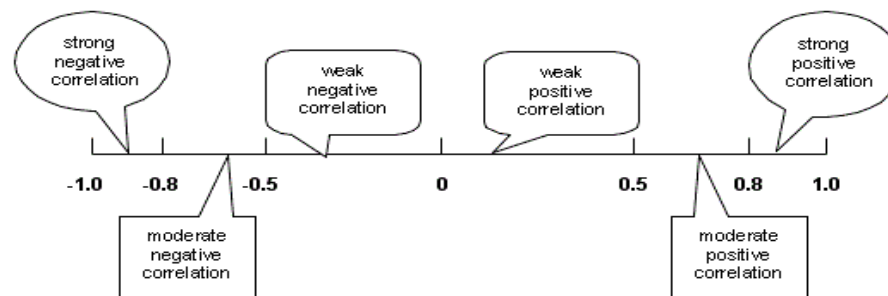
And, since the square of ' r ' is always positive, R^2 approaches $+1$ when there is a strong relation.



Interpreting correlation

- Coefficient of determination
 - Correlation squared
 - Indicates % of shared variance
- | Strength | r | r^2 |
|-----------|---------|----------|
| Weak: | .1 - .3 | 1 - 10% |
| Moderate: | .3 - .5 | 10 - 25% |
| Strong: | > .5 | > 25% |

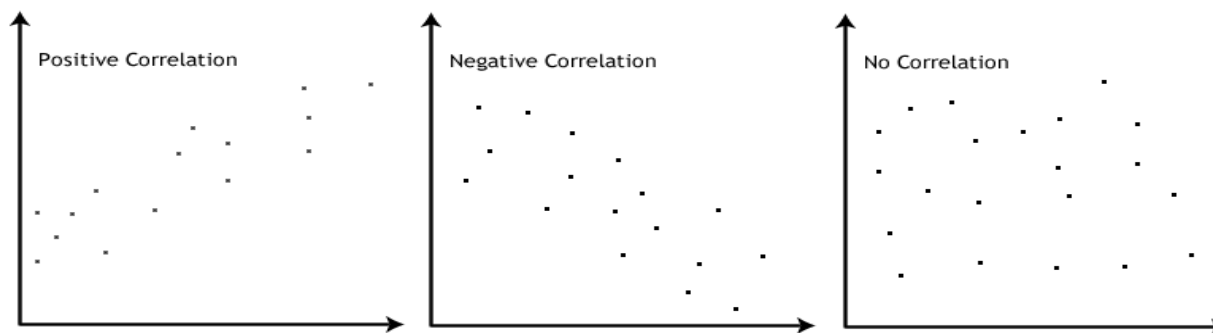
27



- 2) The correlation coefficient provides information about the **direction** of the association between X and Y. Either positively or negatively related.

A positive correlation must always correspond to a regression line with a positive slope (b), indicating a **direct** relationship.

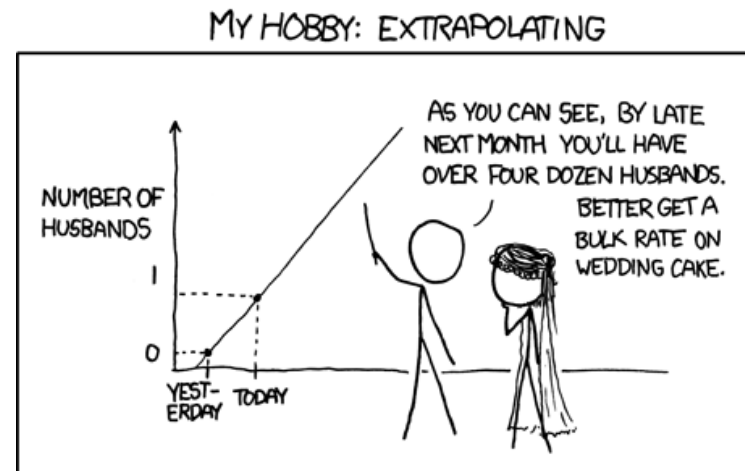
A negative correlation corresponds to an **inverse** relationship.

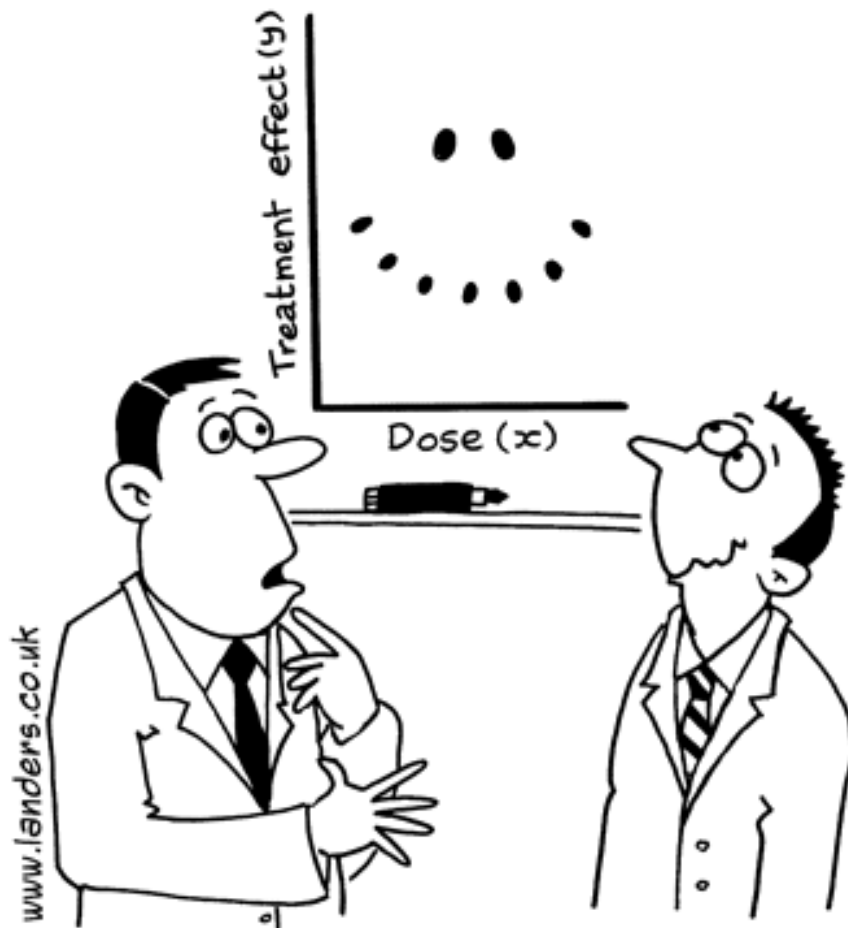


- 3) In correlation analysis the X and Y variables are assumed to have probability **distribution**.

In regression analysis, the independent variable X can be considered as a set of given values.

When X is an independent random variable, correlation between X and Y is zero.





"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."

4) In correlation analysis there is no specified “dependent” and “independent” variable designation.

⇒ In regression analysis we impose a model, such that Y is being explained by X.

⇒ If the dependent and independent variables are switched, the regression model will not be the same, although the sign on the slope parameter will remain the same.

