Simple Linear Regression

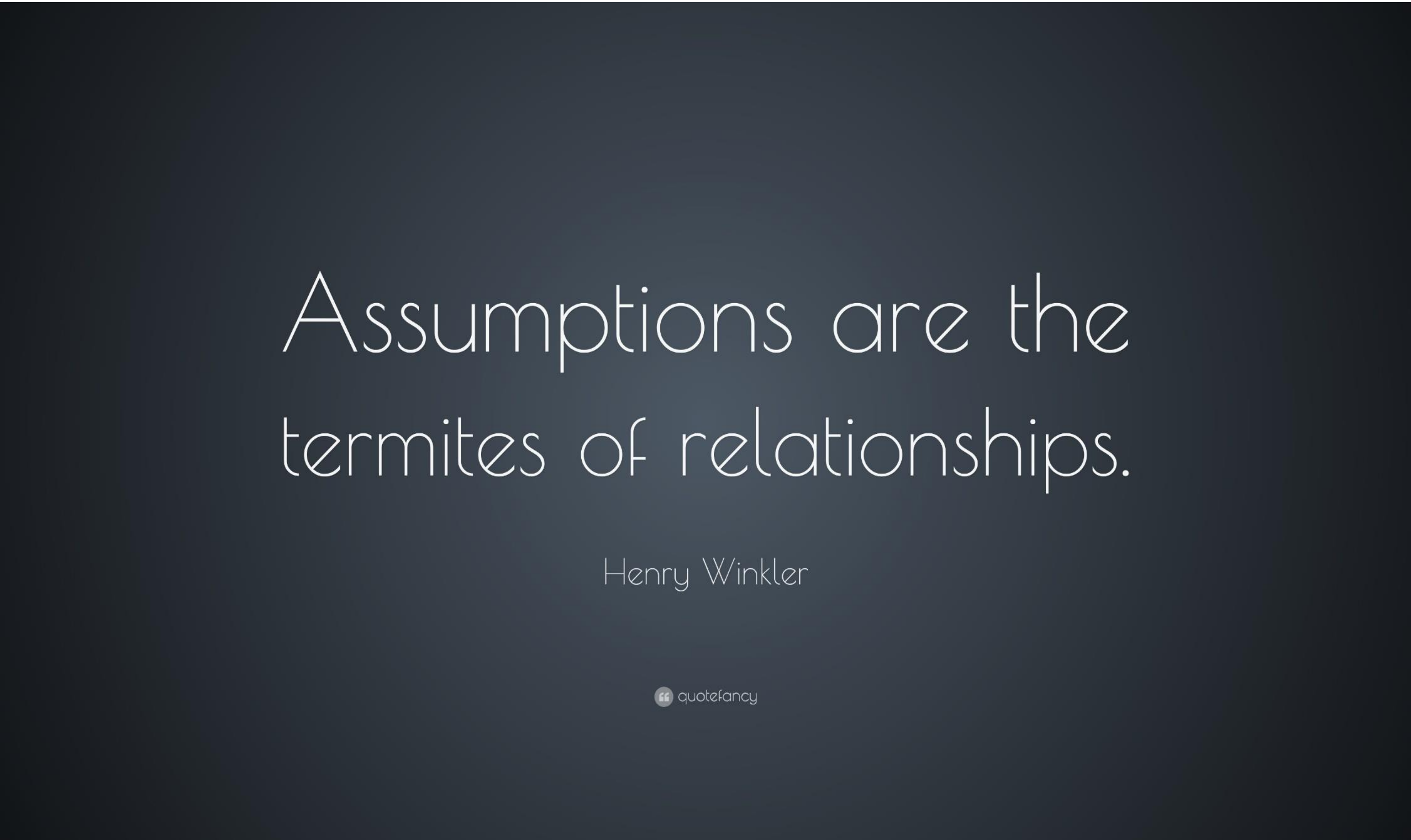
Course Taught at SUAD

Dr. Tanujit Chakraborty

@ Sorbonne

tanujitisi@gmail.com

Quote of the day..



Assumptions are the termites of relationships.

Henry Winkler

quotefancy

This presentation includes…

- Introduction to Relationship Analysis
- Correlation Analyses
- Simple Linear Regression

Correlation

- Correlation coefficient between two R.V.s X and Y, usually denoted by $r(X,Y)\ or\ r_{XY}$ is a numerical measure of linear relationship between them and is defined as:

$$r_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

- r_{XY} provided a measure of linear relationship between X and Y.
- It is a measure of degree of relationship.

- Remember that variance is:

$$Var_X = \frac{\Sigma(X-\overline{X})^2}{N-1} = \frac{\Sigma(X-\overline{X})(X-\overline{X})}{N-1}$$

- The formula for covariance is:

$$Cov_{XY} = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{N-1}$$

Important Results (Proof!)

- Result 1: Limits for $r_{XY}\ is\ -1\ and\ 1$ i.e., $-1 \leq r_{XY} \leq 1$.

- Result 2: Correlation coefficient is independent of change of origin and scale.

- Result 3: Two independent variables are uncorrelated.

- Result 4: Correlation coefficient can't measure nonlinear relationship between two variables.

- Result 5: $r_{XY} = r_{YX}$

Properties of Correlation coefficient

- The correlation coefficient is not only invariant under changes of unit of measurements but also unaffected by changes of origin for both variables. I.e., if all x-values or y-values are added or subtracted by the same constant then the value of the correlation coefficient remains unchanged.

- The above properties (invariance under changes of origin and scale) can be summarized by saying that the correlation coefficient is invariant under linear transformation of x and y (except for the sign).

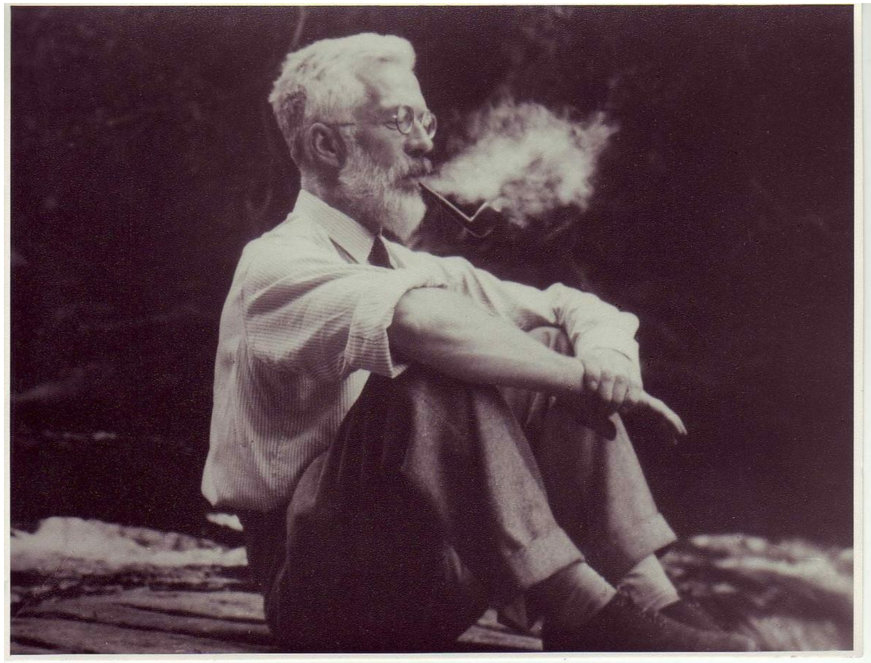
- We have already seen that if x and y are positively/negatively related then the value of r will be positive/negative. r has other good properties. The value of r always lies in-between –1 and +1. r takes the value +1 when all the values are on the positively sloped straight line and the value –1 when all the points are on the negatively sloped straight line.

- As the scatter points move closer to the (hypothetical) straight line, the value of $|r|$ moves to 1. As the points move away from the straight line, the value of r approaches zero. Thus the value of r of diagram 16 is higher than that of diagram 15 and the value of ρ of diagram 19 is higher in absolute value than that of diagram 18.

Heart Disease and Cigarettes

- Data on coronary heart disease and cigarette smoking in 21 developed countries (Landwehr and Watkins, 1987)

- Data have been rounded for computational convenience.



Do you know who is he?

Data

Surprisingly, the U.S. is the first country on the list--the country with the highest consumption and highest mortality.

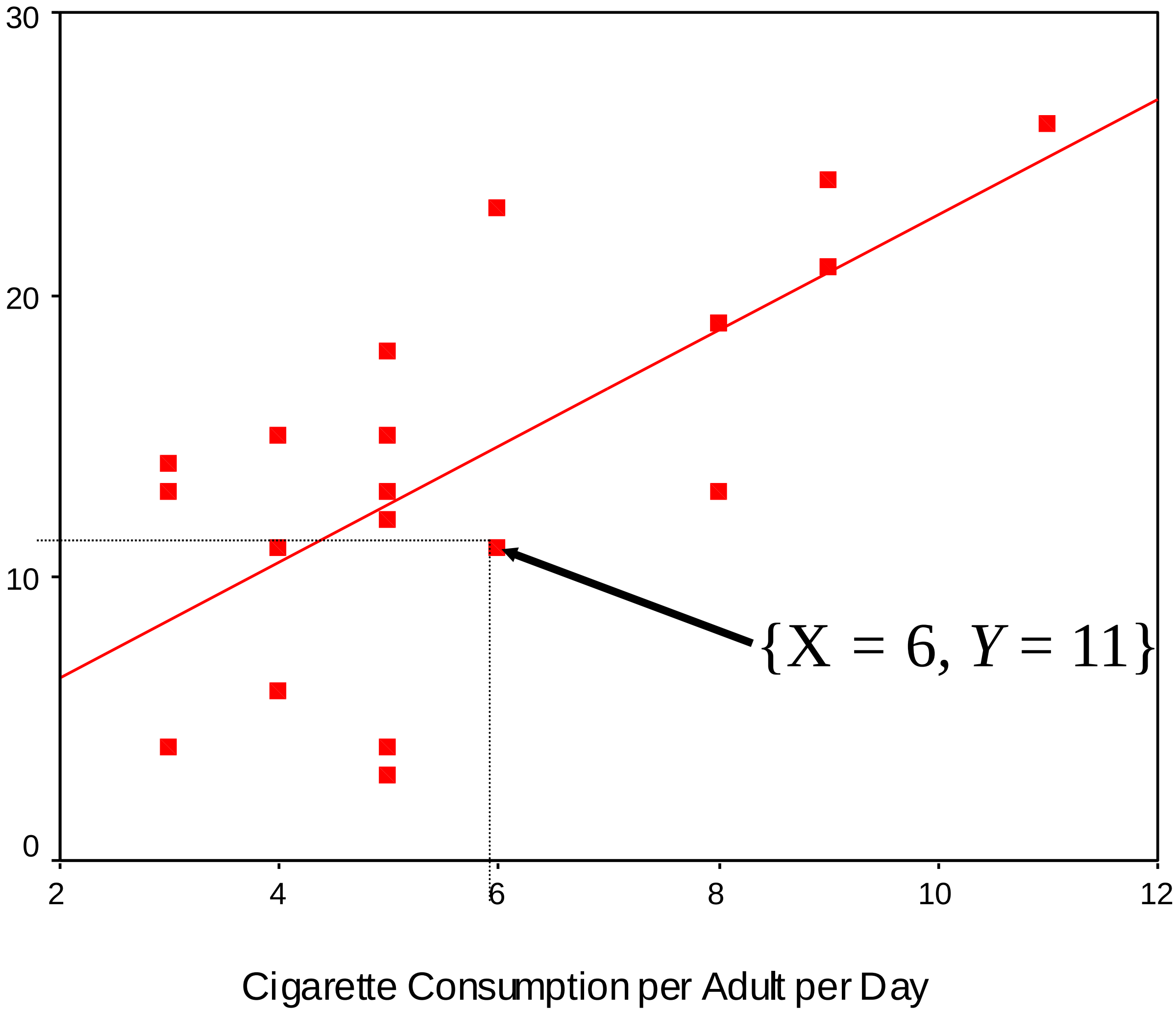
Scatterplot of Heart Disease

- Chronic Heart Disease (CHD) Mortality goes on ordinate (Y axis)
- Cigarette consumption on abscissa (X axis)
- Best fitting line included for clarity

Country	Cigarettes	CHD
1	11	26
2	9	21
3	9	24
4	9	21
5	8	19
6	8	13
7	8	19
8	6	11
9	6	23
10	5	15
11	5	13
12	5	4
13	5	18
14	5	12
15	5	3
16	4	11
17	4	15
18	4	6
19	3	13
20	3	4
21	3	14

What Does the Scatterplot Show?

- As smoking increases, so does coronary heart disease mortality.
- Relationship looks strong
- Not all data points on line.
- This gives us residuals” or “errors of prediction”



$\{X = 6, Y = 11\}$

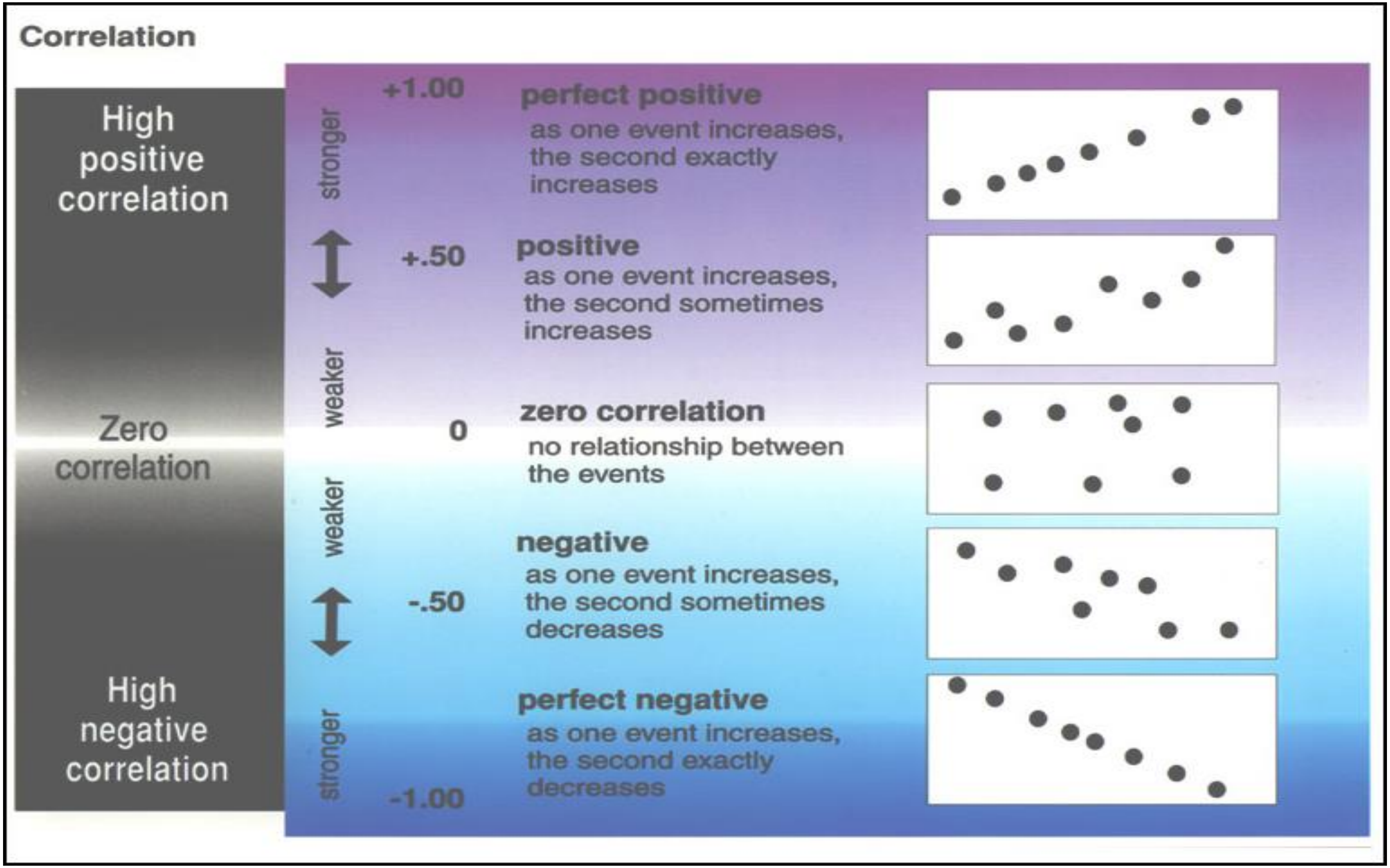
Example: Heart Disease and Cigarettes

Country	X (Cig.)	Y (CHD)	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})*(Y-\bar{Y})$
1	11	26	5.05	11.48	57.97
2	9	21	3.05	6.48	19.76
3	9	24	3.05	9.48	28.91
4	9	21	3.05	6.48	19.76
5	8	19	2.05	4.48	9.18
6	8	13	2.05	-1.52	-3.12
7	8	19	2.05	4.48	9.18
8	6	11	0.05	-3.52	-0.18
9	6	23	0.05	8.48	0.42
10	5	15	-0.95	0.48	-0.46
11	5	13	-0.95	-1.52	1.44
12	5	4	-0.95	-10.52	9.99
13	5	18	-0.95	3.48	-3.31
14	5	12	-0.95	-2.52	2.39
15	5	3	-0.95	-11.52	10.94
16	4	11	-1.95	-3.52	6.86
17	4	15	-1.95	0.48	-0.94
18	4	6	-1.95	-8.52	16.61
19	3	13	-2.95	-1.52	4.48
20	3	4	-2.95	-10.52	31.03
21	3	14	-2.95	-0.52	1.53
Mean	5.95	14.52			
SD	2.33	6.69			
Sum					222.44

$$Cov_{cig.\&CHD} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{N-1} = \frac{222.44}{21-1} = 11.12$$

$$r = \frac{\mathrm{cov}_{XY}}{s_X s_Y} = \frac{11.12}{(2.33)(6.69)} = \frac{11.12}{15.59} = .713$$

Remarks from Scatter Plots & Correlation Coefficient



Correlation
High positive correlation
Zero correlation
High negative correlation
stronger
weaker
weaker
stronger
+1.00
+.50
0
-.50
-1.00
perfect positive
as one event increases, the second exactly increases
positive
as one event increases, the second sometimes increases
zero correlation
no relationship between the events
negative
as one event increases, the second sometimes decreases
perfect negative
as one event increases, the second exactly decreases

Partial Correlation Coefficients

- We begin with the following example.

- Suppose data on IQ (x_1), result in the final examination (x_0) and number of times visited cinema hall (x_2) were taken from a group of students and the simple correlations were calculated as

  $$r_{01} = 0.8,\ r_{02} = 0.3 \text{ and } r_{12} = 0.6.$$

  $r_{02} = 0.3$ has been found to be significantly different from zero. But it is an unexpected result. We do not expect x_0 and x_2 to have a positive correlation. It means that as the students increase there visit to cinema hall their results are likely to be better. There must be something wrong.

- After scrutiny the investigator discovered that the intelligent students mostly visited the cinema hall. To find the true correlation between x_0 and x_2 we should thus eliminate the effect of IQ. This can be done by separately regressing x_0 and x_2 on x_1 and finding the simple correlation of the two residuals.

- The correlation coefficient between two variables after eliminating the effect of a third variable is known as partial correlation coefficient. It is also possible to eliminate the effect of as many variables as we want.

Partial Correlation Coefficients

The formula for partial correlation coefficient of x_0 and x_2 after eliminating the effect of x_1 is

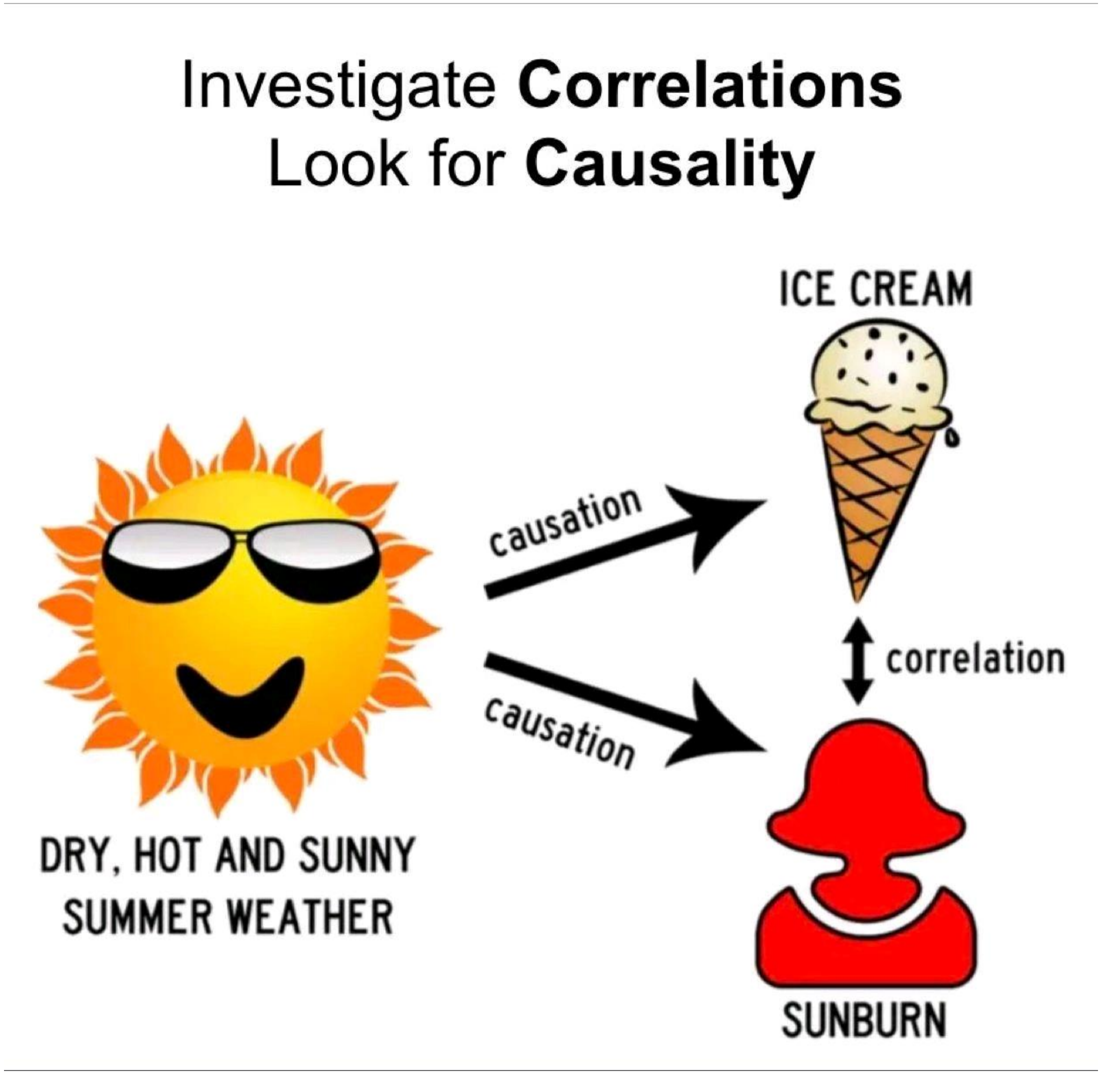
$$r_{02.1} = (r_{02} - r_{01}\, r_{21})/(\sqrt{}(1 - r_{01}^2)\sqrt{}(1 - r_{21}^2))$$

In the above example the value of $r_{02.1}$ is

$$(0.3\text{-}0.8\times 0.6)/(\sqrt{}(1\text{-}.8^2)\sqrt{}(1\text{-}.6^2)) = -\ 0.375.$$

This has a negative sign as expected.

Correlation Vs. Causation

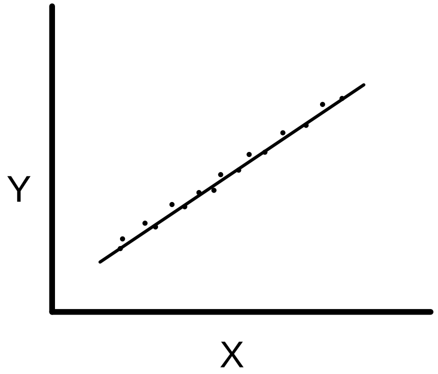


Investigate **Correlations**
Look for **Causality**

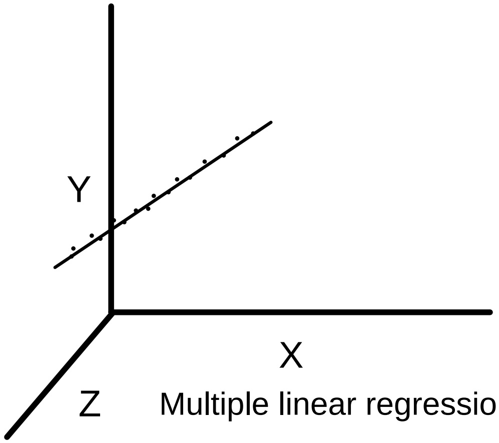
Regression Analysis

- The regression analysis is a statistical method to deal with the formulation of mathematical model depicting relationship amongst variables, which can be used for the purpose of prediction of the values of dependent variable, given the values of independent variables.

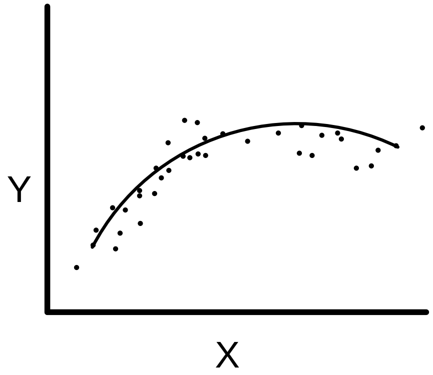
- **Classification of Regression Analysis Models**
  - Linear regression models
    1. Simple linear regression
    2. Multiple linear regression
  - Non-linear regression models



Simple linear regression

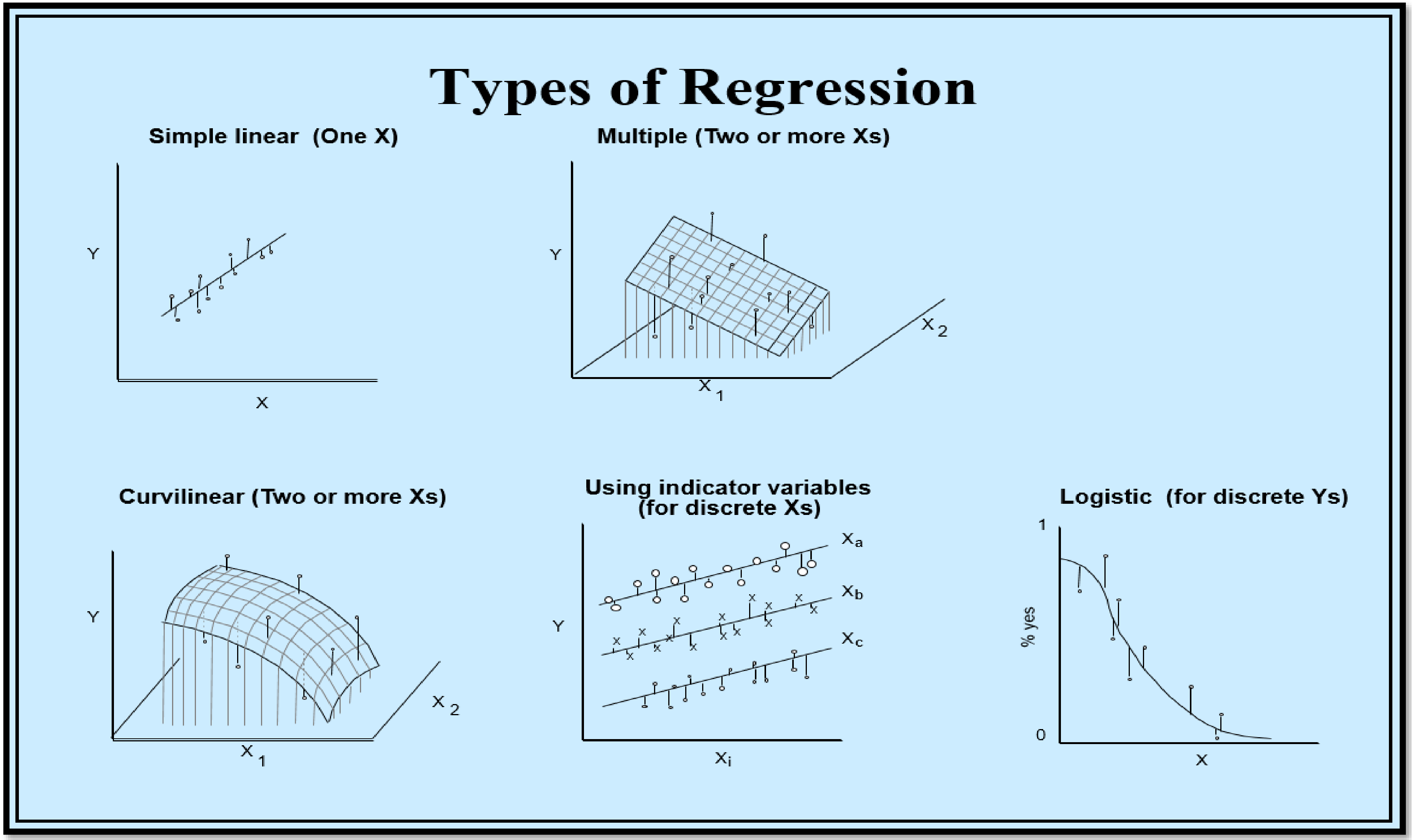


Multiple linear regression



Non-linear regression

Regression Analysis



Types of Regression

Simple linear (One X)

Multiple (Two or more Xs)

Curvilinear (Two or more Xs)

Using indicator variables (for discrete Xs)

Logistic (for discrete Ys)

Earlier Developments of Regression

- The earliest form of regression was the method of least squares, which was published by Legendre in 1805 and by Gauss in 1809. Legendre and Gauss both applied the method to the problem of determining, from astronomical observations; the orbits of bodies about the Sun. Gauss published a further development of the theory of least squares in 1821, including a version of the Gauss–Markov theorem.

- The term "regression" was coined by Francis Galton in the nineteenth century to describe a biological phenomenon. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as regression toward the mean).

- Galton’s work was later extended by Udny Yule and Karl Pearson to a more general statistical context. In the work of Yule and Pearson, the joint distribution of the response and explanatory variables is assumed to be Gaussian. This assumption was weakened by R.A. Fisher in his works of 1922 and 1925. Fisher assumed that the conditional distribution of the response variable is Gaussian, but the joint distribution need not be.

Introduction

- Regression Analysis is a Statistical tool for investigating the relationship between a dependent variable and one or more independent variables.

- Scatter plots are used to investigate the possible relationship between the variables.

- The simple linear regression model is
$$Y = \alpha + \beta x + \epsilon$$

- For a given x, the corresponding observation Y consists of the value $\alpha + \beta x$ plus an amount ϵ.

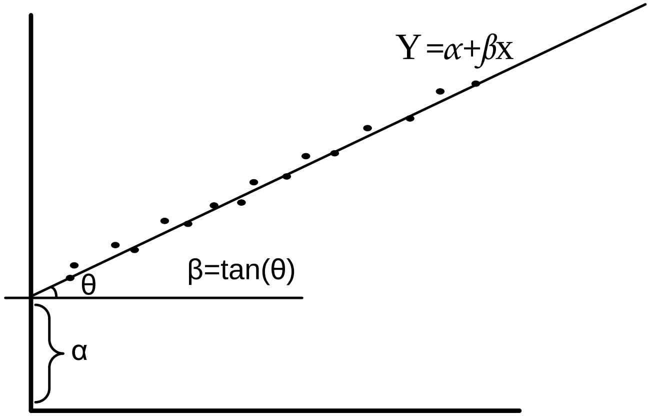
Linear Regression

- The line of regression is the line which gives the best estimate to the value of one variable for any specific value of the other variable.
- That means we want to understand the relationship.
- The line of regression is the line of "best fit" and is obtained by **the principle of least squares**.
- Y - the variables you are predicting
  - i.e. dependent variable
- X - the variables you are using to predict
  - i.e. independent variable
- $\hat{Y}$ - your predictions (also known as Y')

Simple Linear Regression Model

In simple linear regression, we have only two variables:

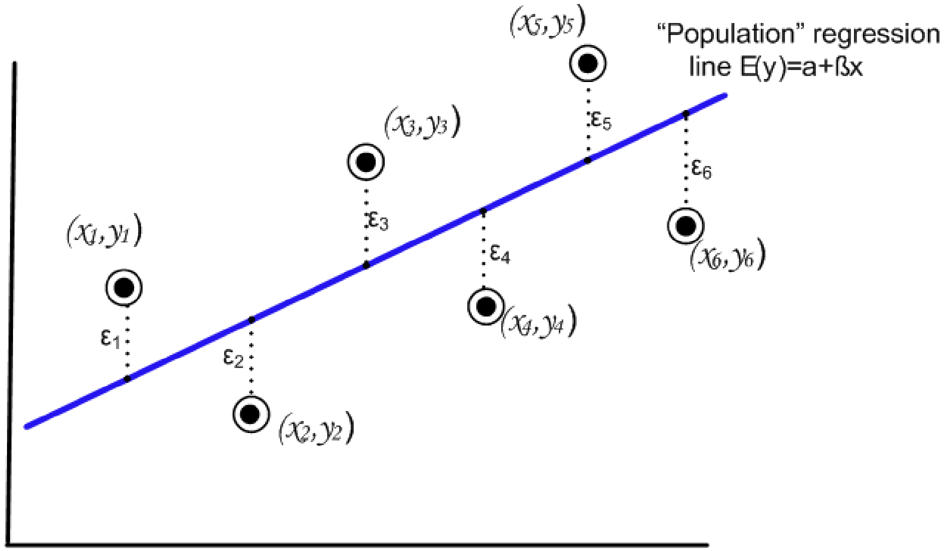
- Dependent variable (also called Response), usually denoted as Y.
- Independent variable (alternatively called Regressor), usually denoted as x.
- A reasonable form of a relationship between the Response Y and the Regressor x is the linear relationship, that is in the form $Y = \alpha + \beta x$



Note:

- There are infinite number of lines (and hence α_s and β_s)
- The concept of regression analysis deal with finding the best relationship between Y and x (and hence best fitted values of α and β) quantifying the strength of that relationship.

Regression Analysis



Given the set $[(x_i, y_i), i = 1,2, \dots \dots, n]$ of data involving n pairs of (x, y) values, our objective is to find "true" or population regression line such that $Y = \alpha + \beta x + \in$

Here, $\in$ is a random variable with $E(\in) = 0$ and $var(\in) = \sigma^2$. The quantity σ^2 is often called the **error variance**.

Note:

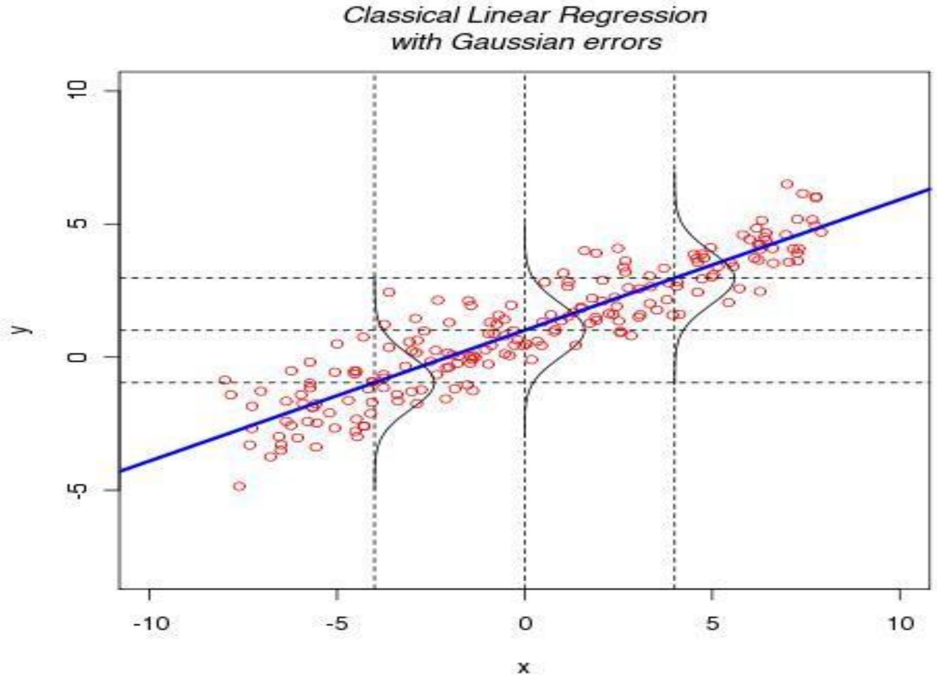
- $E(\in) = 0$ implies that at a specific x, the y values are distributed around the "true" regression line $Y = \alpha + \beta x$ (i.e., the positive and negative errors around the true line is reasonable).
- α and β are called **regression coefficients**.
- α and β values are to be estimated from the data.

Assumption on the model

- The Linear Regression Model for $i\ th$ observation

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, 2, \ldots, n.$$

- Assumptions:
  - $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$
  - $Cov(\epsilon_i, \epsilon_j) = 0, i \neq j$
  - $\epsilon_i\ (iid) \sim N(0, \sigma^2)$



Assumption on the model

Consequent assumptions on Y:

- $E(Y_i) = \alpha + \beta x_i$

- $Var(Y_i) = \sigma^2$

- $Y_i's\ are\ independent\ and\ normally\ distributed.$

Basic Assumptions of the Linear Regression Model

- The slope β of the population regression line is the *average* change in Y associated with a 1-unit increase in x.

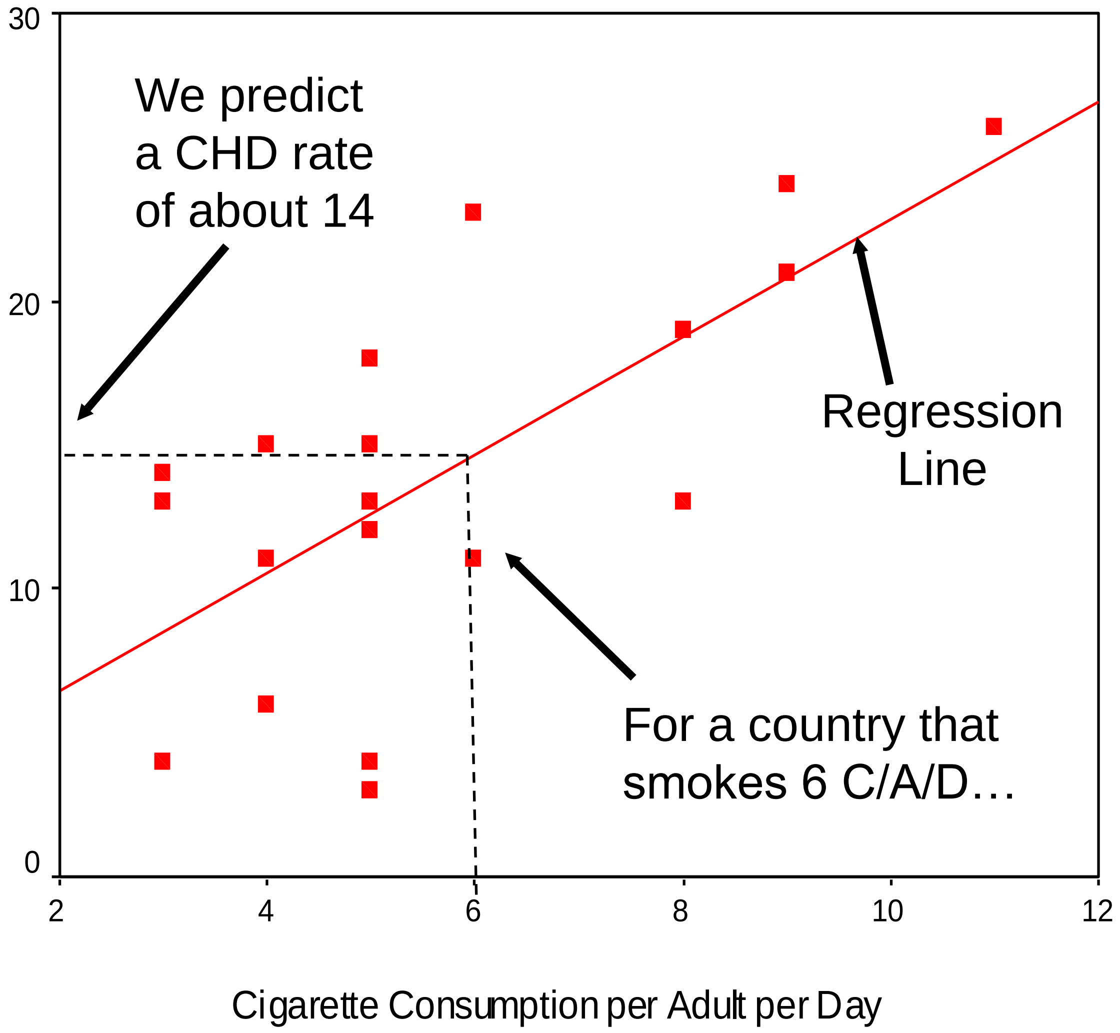
- The Y intercept α is the height of the population line when $x = 0$.

- The value of σ determines the extent to which (x, y) observations deviate from the regression line.

- When σ is small, most observations will be quite close to the line, but when σ is large, there are likely to be some large deviations.

Example: Heart Disease and Cigarettes

Country	Cigarettes	CHD
1	11	26
2	9	21
3	9	24
4	9	21
5	8	19
6	8	13
7	8	19
8	6	11
9	6	23
10	5	15
11	5	13
12	5	4
13	5	18
14	5	12
15	5	3
16	4	11
17	4	15
18	4	6
19	3	13
20	3	4
21	3	14



We predict a CHD rate of about 14

Regression Line

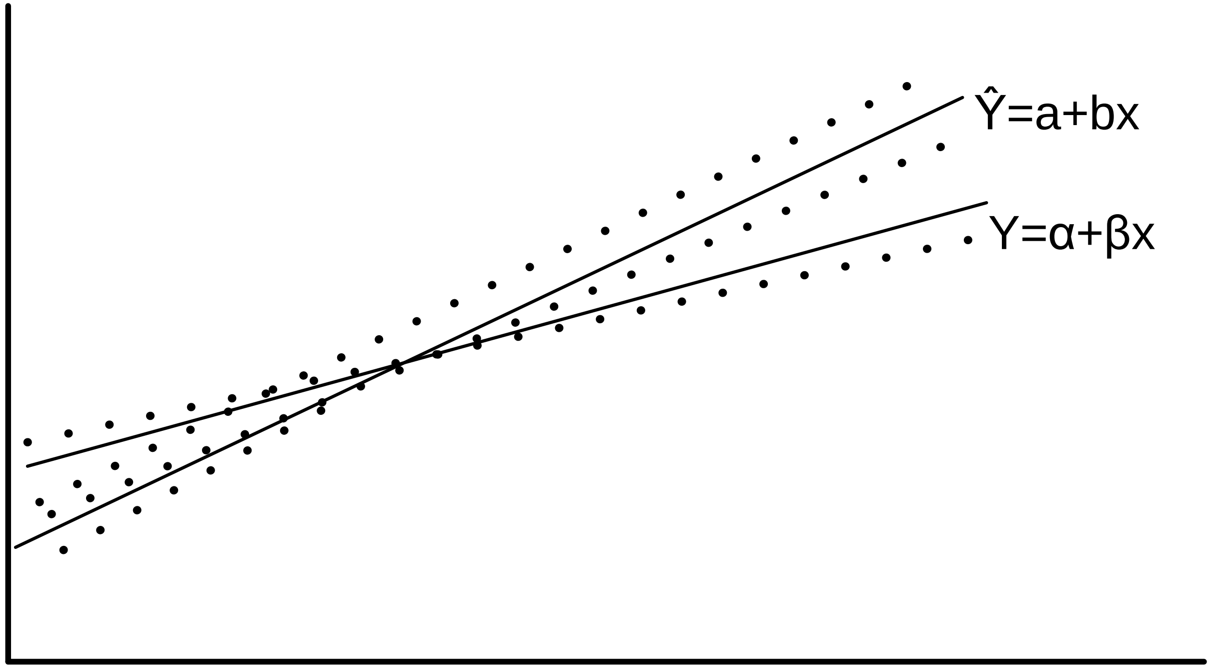
For a country that smokes 6 C/A/D…

True versus Fitted Regression Line

- The task in regression analysis is to estimate the regression coefficients α and β.
- Suppose, we denote the estimates a for α and b for β. Then the fitted regression line is

$$\hat{Y} = a + bx$$

where $\hat{Y}$ is the predicted or fitted value.



Regression line

Formula:

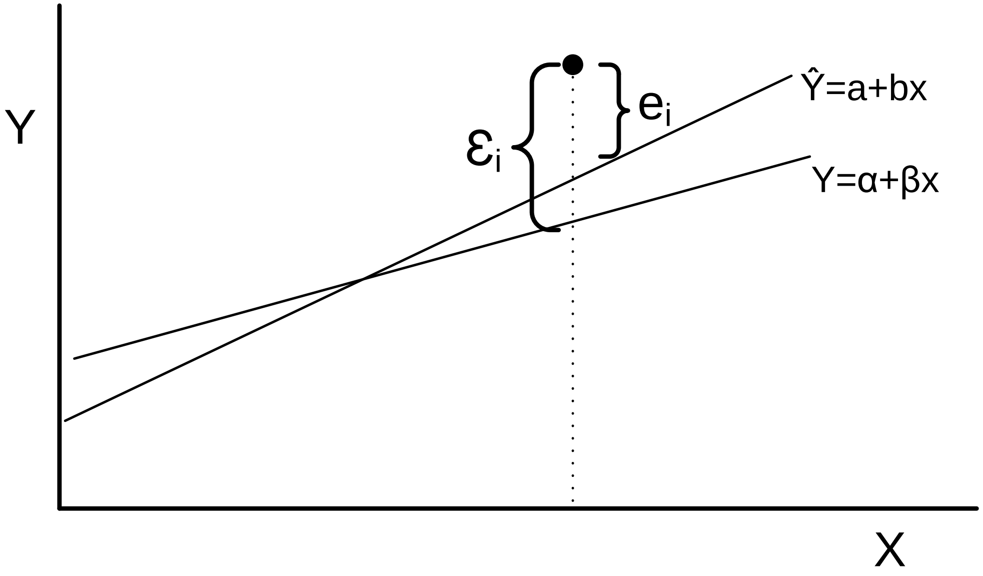
$$\hat{Y} = a + bX$$

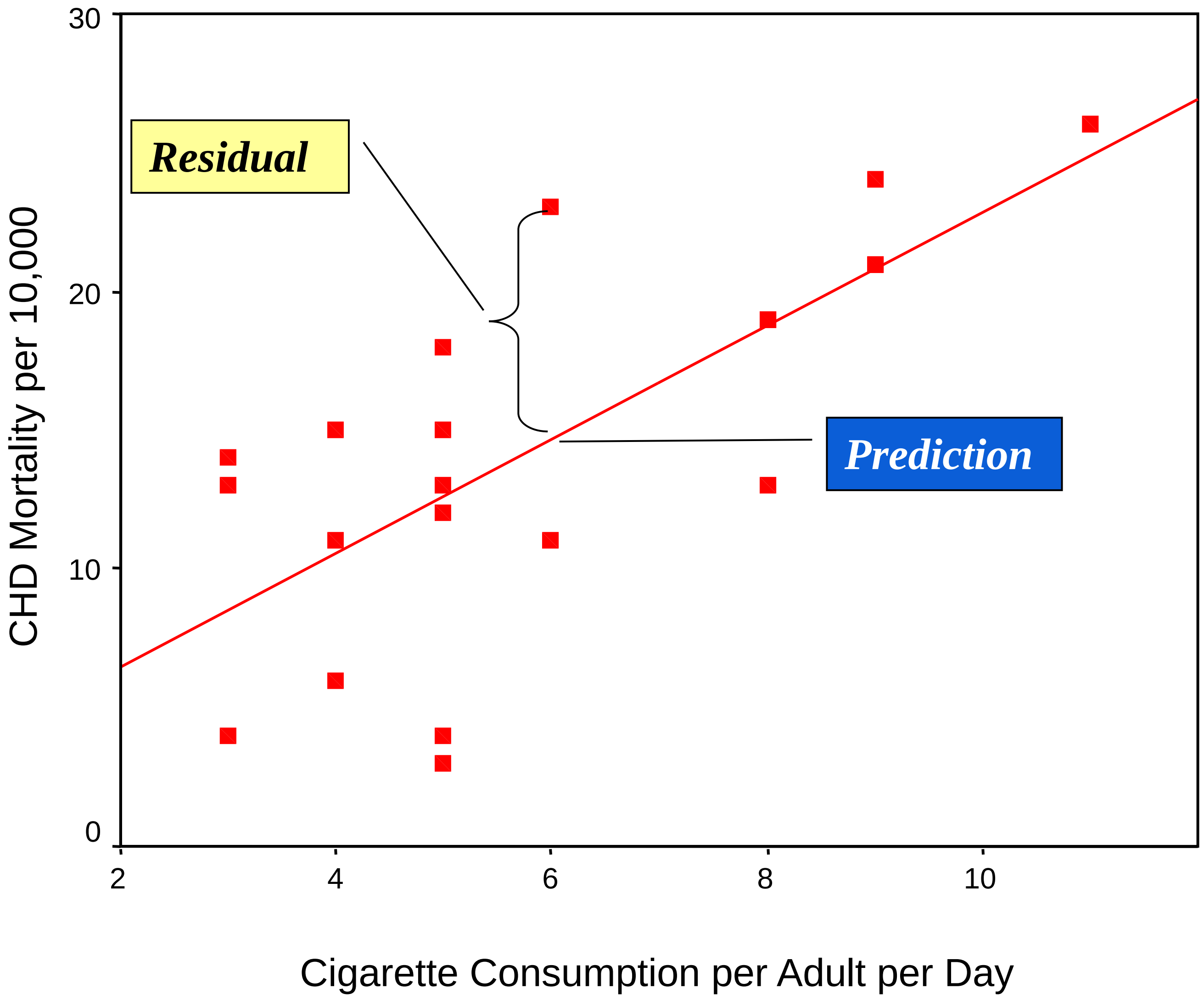
- $\hat{Y}$ = the predicted value of Y (e.g. CHD mortality)
- $\boldsymbol{X}$ = the predictor variable (e.g. average cig./adult/country)
- a and b are “Coefficients”
- b = slope ,i.e. change in predicted Y for one unit change in X
- a = intercept , i.e. value of when $X = 0$

Least Square Method to estimate α and β

This method uses the concept of residual. A residual is essentially an error in the fit of the model $\widehat{\mathrm{Y}} = a + bx$. Thus, i^{th} residual is

$$e_i = Y_i - \hat{Y}_i,\ i = 1,2,3, \dots \dots, n$$




Residual
Prediction
30
20
10
0
2
4
6
8
10
CHD Mortality per 10,000
Cigarette Consumption per Adult per Day

Least Square method

- The residual sum of squares is often called **the sum of squares of the errors** about the fitted line and is denoted as SSE

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

- We are to minimize the value of SSE and hence to determine the parameters of *a* and *b*.
- Differentiating SSE with respect to *a* and *b*, we have

$$\frac{\partial(SSE)}{\partial a} = -2\sum_{i=1}^{n} (y_i - a - bx_i)$$

$$\frac{\partial(SSE)}{\partial b} = -2\sum_{i=1}^{n} (y_i - a - bx_i).x_i$$

For minimum value of SSE, $\frac{\partial(SSE)}{\partial a} = 0$

$$\frac{\partial(SSE)}{\partial b} = 0$$

Least Square method to estimate α and β

- Thus, we set

$$na+b\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}y_i$$

$$a\sum_{i=1}^{n}x_i+\mathrm{b}\sum_{i=1}^{n}{x_i}^2=\sum_{i=1}^{n}x_iy_i$$

- These two equations can be solved to determine the values of a and b, and it can be calculated that

$$\boxed{\begin{gathered}b=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}=\frac{Cov(X,Y)}{\sigma_X^2}\\[2em] a=\overline{y}-b\overline{x}\end{gathered}}$$

- Line of regression of Y on X passes through the point $(\bar{x},\bar{y})$:

$$Y-\bar{y}=\frac{\sigma_Y}{\sigma_X}r_{XY}(X-\bar{x})$$

Properties of Least square fit:

- Prop 1: $\sum e_i = \sum(y_i - \hat{y}_i) = 0.$

- Prop 2: $\sum y_i = \sum \hat{y}_i$.

- Prop 3: $\sum x_i e_i = 0.$

- Prop 4: $\sum \hat{y}_i e_i = 0.$

Linear Regression

- Regression Equation of Y on X:

$$Y - \bar{y} = \frac{\sigma_Y}{\sigma_X} r_{XY} (X - \bar{x})$$

- Regression Equation of X on Y:

$$X - \bar{x} = \frac{\sigma_X}{\sigma_Y} r_{XY} (Y - \bar{y})$$

- Regression Coefficient of Y on X:

$$b_{YX} = \frac{\sigma_Y}{\sigma_X} r_{XY}$$

- Regression Coefficient of X on Y:

$$b_{XY} = \frac{\sigma_X}{\sigma_Y} r_{XY}$$

For the Data: Heart Disease and Cigarettes

- $Cov(X,Y) \ = 11.12619$
- $\sigma^2{}_X \ = \ 2.334014^2 \ = \ 5.447619$
- $b \ = \ 11.12619/5.447619 \ = \ 2.042395$
- $a \ = \ 14.52381 \ - \ 2.042395 * 5.952381 \ = \ 2.366696$
- the equation of the least-squares line:

$$\boxed{\hat{y} = 2.367 + 2.042\, x}$$

Properties of Regression Coefficient

- Property 1: Correlation coefficient is the geometric mean between the regression coefficient.

- Property 2: If one of the regression coefficient is greater than unity, the other must be less than unity.

- Property 3: Regression coefficients are independent of the change of origin but not of scale.

Exercises:

- Studies have shown that people who suffer sudden cardiac arrest have a better chance of survival if a defibrillator shock is administered very soon after cardiac arrest. How is survival rate related to the time between when cardiac arrest occurs and when the defibrillator shock is delivered? The accompanying data give *y* = survival rate (percent) and *x*=5 mean call-to shock time (minutes) for a cardiac rehabilitation center (in which cardiac arrests occurred while victims were hospitalized and so the call-to-shock time tended to be short) and for four communities of different sizes:

Mean call-to-shock time, *x* :	2	6	7	9	12
Survival rate, *y* :	90	45	30	5	2

a) Construct a scatterplot for these data. How would you describe the relationship between mean call-to shock time and survival rate?

b) Find the equation of the least-squares line.

c) Use the least-squares line to predict survival rate for a community with a mean call-to-shock time of 10 minutes.

Exercises:

Let *x* be the size of a house (in square feet) and *y* be the amount of natural gas used (therms) during a specified period. Suppose that for a particular community, *x* and *y* are related according to the simple linear regression model with

β = slope of population regression line = 0.017

α = *y* intercept of population regression line = -5.0

Houses in this community range in size from 1000 to 3000 square feet.

a. What is the equation of the population regression line?

b. Graph the population regression line by first finding the point on the line corresponding to x = 1000 and then the point corresponding to x =2000, and drawing a line through these points.

c. What is the mean value of gas usage for houses with 2100 sq. ft. of space?

d. What is the average change in usage associated with a 1 sq. ft. increase in size?

e. What is the average change in usage associated with a 100 sq. ft. increase in size?

Assumptions Revisited

In regression analysis:

- All assumptions are made about the residuals

- No assumptions are made for X or Y

- **Residuals** need to be:
  - Bell-shaped (normal)
  - Stable (over time)*
  - Random
  - Unrelated (we think X *is* related to Y) Plot your data before doing regression analysis

- Residuals need to show certain properties for regression to work properly

- The regression equation can be used to predict (or possibly manage) output data from input/process data

Predicted Values and Residuals

- The **predicted** or **fitted values** result from substituting each sample *x* value in turn into the equation for the least-squares line. This gives

$$\hat{y}_n = nth\ predicted\ value = a + bx_n$$

- The **residuals** from the least-squares line are the *n* quantities

$$e_n = y_n - \hat{y}_n = \text{nth residual}$$

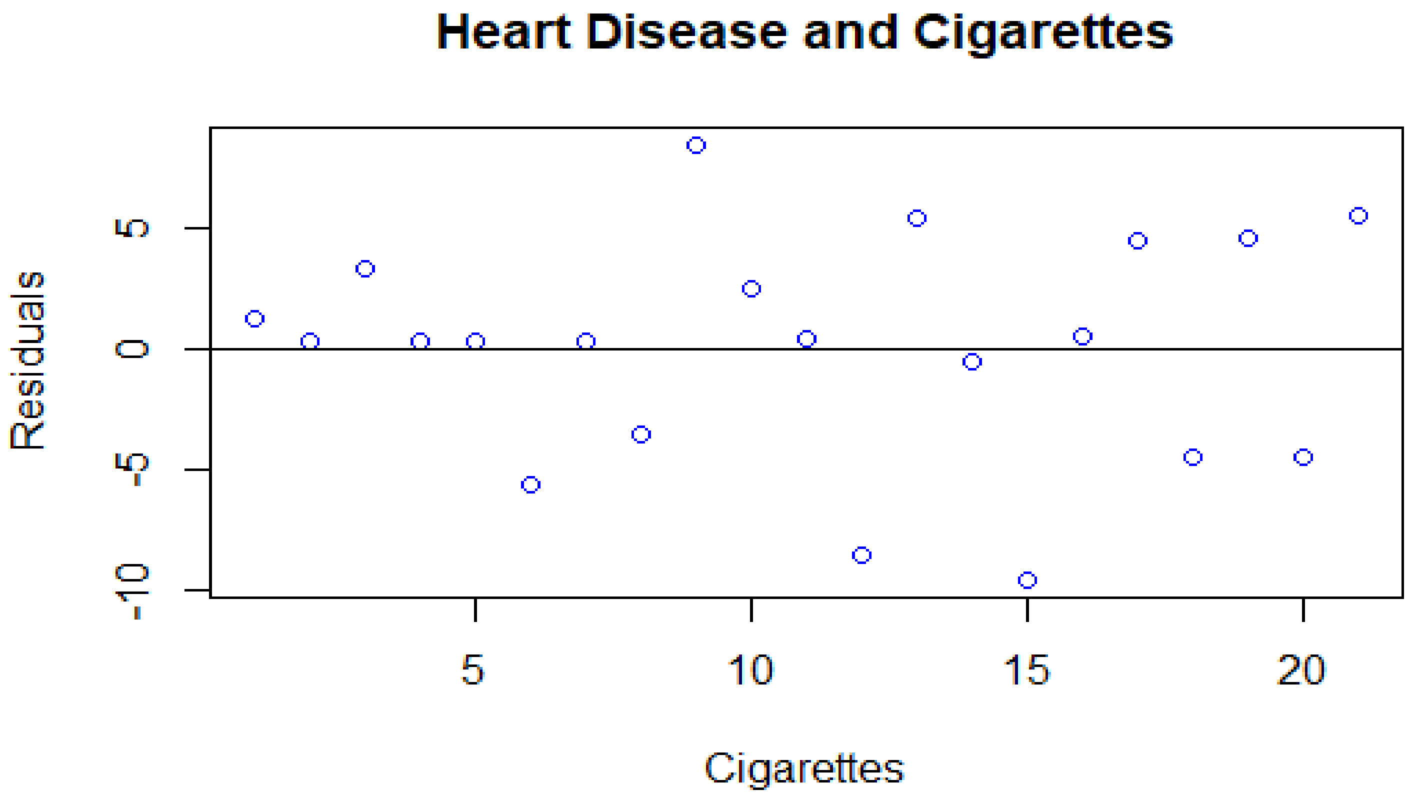
- Each residual is the difference between an observed *y* value and the corresponding predicted *y* value.
- **Plotting the residual:** A **residual plot** is a scatterplot of the (*x*, residual) pairs. Isolated points or a pattern of points in the residual plot indicate potential problems.
- Residuals are highly useful for studying whether a given regression model is appropriate for the data at hand.

Residual : Heart Disease and Cigarettes

The equation of the least-squares line: $\hat{y} = 2.367 + 2.042\,x$

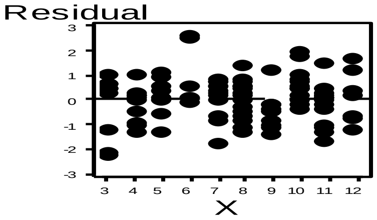
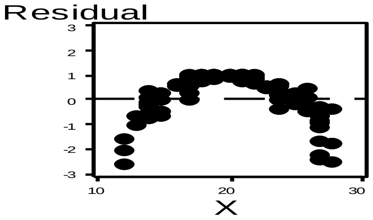
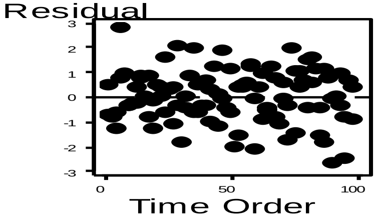
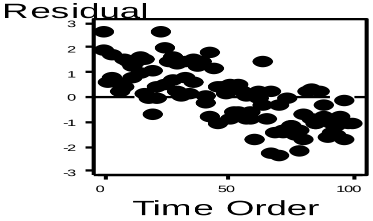
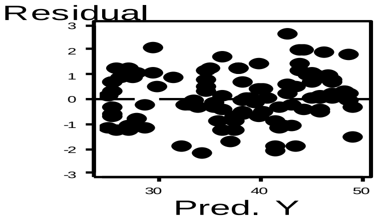
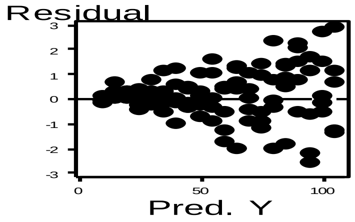
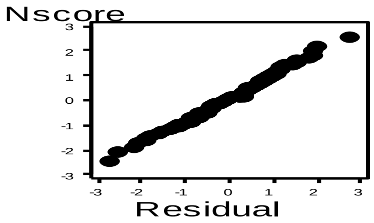
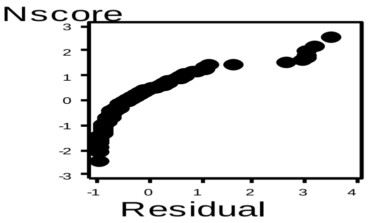
	x	y	$\hat{y}$	$y - \hat{y}$
1	11	26	24.829	1.1669580
2	9	21	20.745	0.2517483
3	9	24	20.745	3.2517483
⋮	⋮	⋮	⋮	⋮
7	8	19	18.703	0.2941434
8	6	11	14.619	−3.6210664
9	6	23	14.619	8.3789336
⋮	⋮	⋮	⋮	⋮
13	5	18	12.577	5.4213287
14	5	12	12.577	−0.5786713
15	5	3	12.577	−9.5786713
⋮	⋮	⋮	⋮	⋮
18	4	6	10.535	−4.5362762
19	3	13	8.493	4.5061189
20	3	4	8.493	−4.4938811
21	3	14	8.493	5.5061189

Residual Plot: Heart Disease and Cigarettes



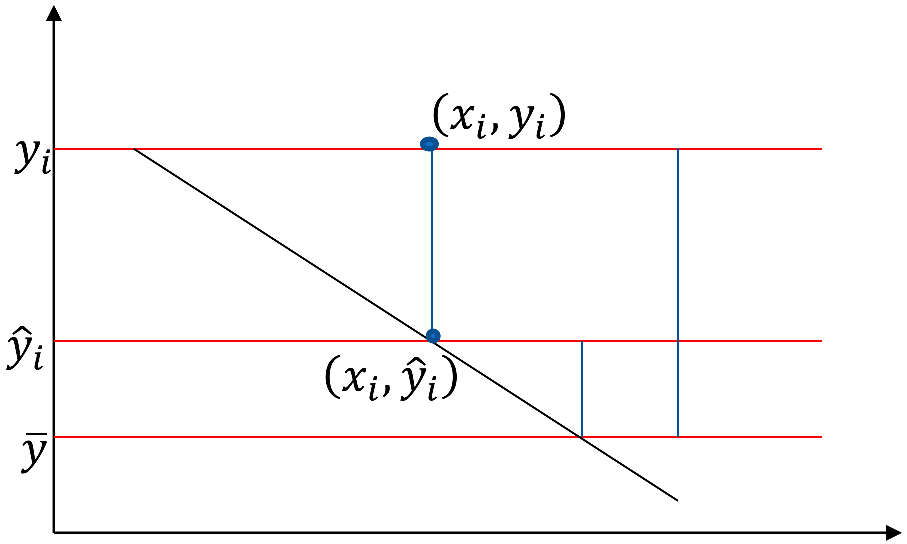
Residuals plots must be checked to ensure the assumptions hold; otherwise, the regression equation may be incorrect or misleading.

Checking Assumptions About Residuals

Residuals Plot	Good	Bad	Meaning / Actions
1. Residuals vs Each X *Used to check that the residuals are not related to the Xs*			The relationship between X & Y is not a straight line, but a curve. Try a transformation on X, Y, or both. Or use X^2 in a multiple regression.
2. Time Plot of Residuals *Used to check for stability over time*			*Any* pattern visible over time means another factor, related to time, influences Y. Try to discover it and include it in a multiple regression.
3. Residuals vs Predicted Y (Fits) *Used to check that they are constant over the range of Ys*			This fan shape means the variation increases as Y gets larger (it's not constant). Try a square root, log, or inverse transformation on Y.
4. Normal Probability Plot of Residuals *Used to check that residuals are Normal*			The residuals are not Normal. Try a transformation on X or Y or both.

Variation

- A set of data : (x_i, y_i).
- $\hat{Y}: predicted\ value\ of\ Y$
- $\bar{Y}: mean\ of\ Y - value$



Deviation of the i^{th} observation from the mean= Deviation of the i^{th} observation from the predicted value + Deviation of the i^{th} predicted value from the mean

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

- Total variation in data = $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

- The sum of the squared of the differences between each predicted y-value

  and $\bar{Y}$ = $\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$
- The sum of the squared of the differences between the y-value of each ordered pair and each corresponding predicted y-value= $\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$.

R^2 : Measure of Quality of Fit

- A quantity R^2, is called **coefficient of determination** is used to measure the proportion of variability of the fitted model.
- We have $SSE = \sum_{i=1}^{n}(y_i - \hat{y})^2$
- It signifies the **variability due to error**.
- Now, let us define the total corrected sum of squares, defined as

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

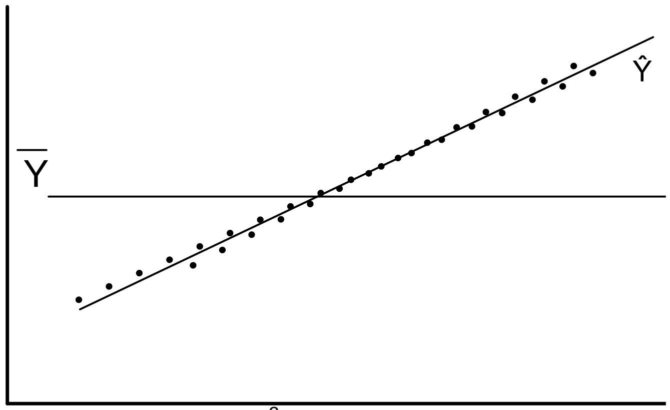
- SST represents the variation in the response values. The R^2 is

$$R^2 = 1 - \frac{SSE}{SST}$$

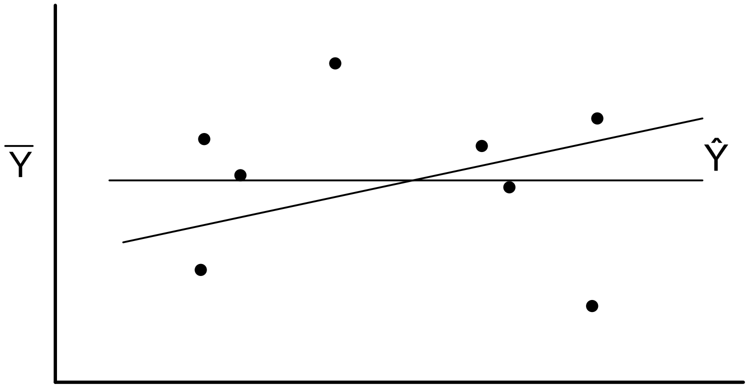
Note:

- If fit is perfect, all residuals are zero and thus $R^2 = 1.0$ (very good fit)
- If SSE is only slightly smaller than SST, then $R^2 \approx 0$ (very poor fit)
- The value of R^2 is often converted to a percentage (by multiplying by 100) and interpreted as the percentage of variation in *y* that can be explained by an approximate linear relationship between *x* and *y*.

R^2 : Measure of Quality of Fit



R^2≈ 1.0 (Very good fit)

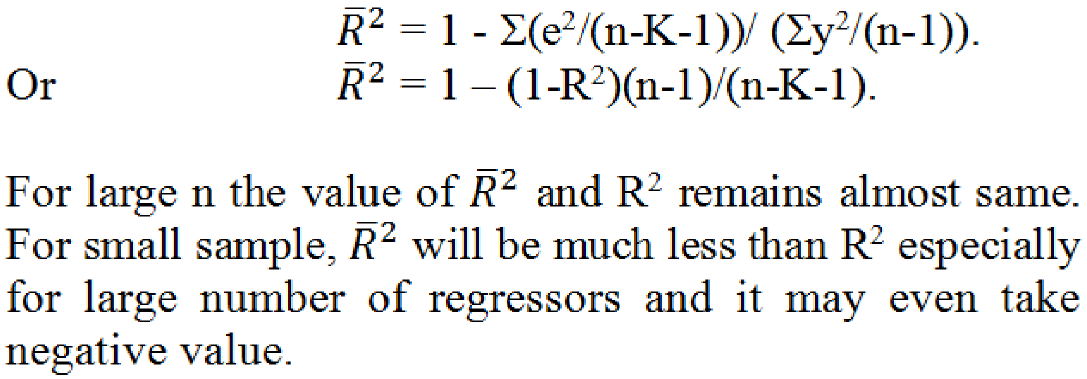


$R^2 \approx 0$ (Very poor fit)

Adjusted R^2

- The above formula for R^2 does not take into account the loss of degrees of freedom from the introduction of the additional explanatory variables in the function. The inclusion of additional explanatory variables in the function can never reduce the coefficient of multiple determination and will usually raise it.

- We introduce adjusted R^2 to compare the goodness of fit of two regression equations with different degrees of freedom. The formula for adjusted R^2 is



$$\bar{R}^2 = 1 - \Sigma(e^2/(n\text{-}K\text{-}1))/\ (\Sigma y^2/(n\text{-}1)).$$

Or

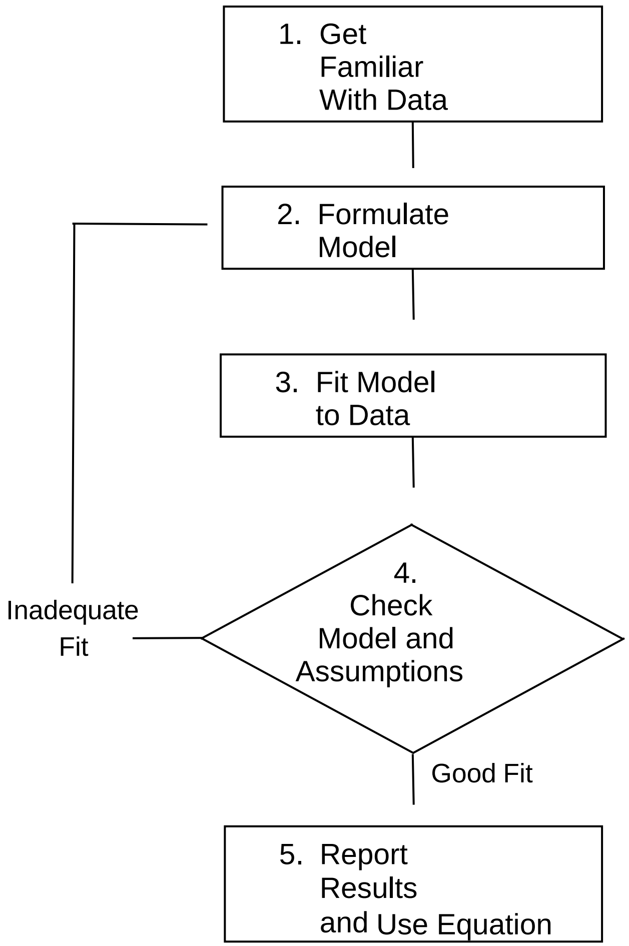
$$\bar{R}^2 = 1 - (1\text{-}R^2)(n\text{-}1)/(n\text{-}K\text{-}1).$$

For large n the value of $\bar{R}^2$ and R^2 remains almost same. For small sample, $\bar{R}^2$ will be much less than R^2 especially for large number of regressors and it may even take negative value.

Review: Interpreting Output for Regression

Name	Definition	Range	Meaning
P-value for slope	Probability that the slope is significant (different from zero)	0 to 1	If less than .05, the slope is significant (different from zero) and X is linearly related to Y.
r	Correlation coefficient	-1 to +1	Indicates the strength of a linear relationship. Numbers near zero indicate no linear relationship.
R-Square (R-sq)	Percent of explained variation $= r^2$	0 to 100%	% of variation in the Y-values explained by the linear relationship with X.
s	Standard deviation of the residuals (unexplained variation)	0 to ∞	Indicates how much the typical observed value differs from the fitted value, in units of the original data.
Residual	= Observed Y – Predicted Y	$-\infty$ to $+\infty$	Residuals are assumed to be random, and Normal with a mean of zero (represent common cause variation).
Standardized Residual	$= \frac{\text{residual}}{\text{standard deviation}}$	About –3 to about +3	If the absolute value of a standardized residual is > 3, then it's an unusual observation. Investigate it.
Influential Observation	An observation whose X-value has a large influence on the values of the coefficients (the regression line)	$-\infty$ to $+\infty$	View them on a plot to decide whether you will keep them or drop them from the regression analysis.

Five Step Regression Procedure: Overview



1. Get Familiar With Data

- Look at plots
- Look at descriptive statistics

2. Formulate Model

- Linear or curvilinear?
- One X or more Xs?
- Transform?
- Discrete X, discrete Y?

3. Fit Model to Data

- Do the regression

4. Check Model and Assumptions

Inadequate Fit

Good Fit

- Look at residuals plots
- Look at unusual observations
- Look at R-Sq
- Look at P-values for b

5. Report Results and Use Equation

- Make predictions for X-values of interest

Exercise:

The data in the accompanying table is from the paper **“Six-Minute Walk Test in Children and Adolescents” (*The Journal of Pediatrics* [2007]: 395–399)**. Two hundred and eighty boys completed a test that measures the distance that the subject can walk on a flat, hard surface in 6 minutes. For each age group shown in the table, the median distance walked by the boys in that age group is also given.

Age Group	Representative Age (Mid point of Age Group)	Median Six-minute Walk Distance (Meters)
3-5	4	544.3
6-8	7	584.0
9-11	10	667.3
12-15	13.5	701.1
16-18	17	727.6

Exercise:

a) With *x* = representative age and *y* = median distance walked in 6 minutes, construct a scatterplot. Does the pattern in the scatterplot look linear?

b) Find the equation of the least-squares regression line that describes the relationship between median distance walked in 6 minutes and representative age.

c) Compute the five residuals and construct a residual plot. Are there any unusual features in the plot?

Inferential Properties of Least Squares Estimators

Prop 1: Both a and b are unbiased estimator of α and β,

$i.e., E(a) = \alpha$ and $E(b) = \beta.$

Proof: $b = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \sum_{i=1}^{n} c_i y_i,$

$where\ c_i = \frac{x_i-\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$. Therefore, b is a linear combination of y_i.

Similarly, $a = \bar{y} - b\bar{x}$. Therefore, a is a linear combination of y_i's.

The Linear Regression Model for i^{th} observation

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, 2, \ldots, n.$$

Now, $\bar{Y} = \alpha + \beta\bar{x} + \bar{\epsilon}$. Therefore, $E[Y_i - \bar{Y}] = \beta(x_i - \bar{x})$.

Now, $E[b] = E\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right] = \beta\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \beta.$

Again, $E[a] = E[\bar{Y} - b\bar{x}] = E[\bar{Y}] - \bar{x}E[b] = \alpha + \beta\bar{x} - \beta\bar{x} = \alpha.$

Hence, both a and b are unbiased estimator of α and β.

Inferential Properties of Least Squares Estimators

Prop 2: $Var(a) = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]$, $Var(b) = \frac{\sigma^2}{S_{xx}}$.

Proof:

$$Var(b) = Var\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] = \sum_{i=1}^{n} Var\left[\frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} Y_i\right]$$

$$= \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} Var[Y_i] = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sigma^2 = \frac{\sigma^2}{S_{xx}}.$$

$$Var(a) = Var[\bar{Y} - b\bar{x}] = Var[\bar{Y}] + \bar{x}^2\frac{\sigma^2}{S_{xx}} - 2\bar{x}Cov[\bar{Y}, b].$$

Now, $Cov[\bar{Y}, b] = Cov\left[\frac{\sum Yi}{n}, \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] = Cov\left[\frac{\sum Yi}{n}, \sum d_i Y_i\right]$, $d_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

$$Cov[\bar{Y}, b] = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Cov[Y_i, Y_i]}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Var[Y_i]}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}\left[\sum_{i=1}^{n} x_i - n\bar{x}\right] = 0.$$

Hence, $Var(a) = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]$.

Estimating σ^2

- The value of σ determines the extent to which observed points (x, y) tend to fall close to or far away from the population regression line.

- A point estimate of σ is based on $SSResid = \sum(y - \hat{y})^2$, where $\hat{y}_1 = a + bx_1, \hat{y}_2 = a + bx_2, \dots, \hat{y}_n = a + bx_n$ are the fitted or predicted y values and the residuals are $y_1 - \hat{y}_1, \dots, y_n - \hat{y}_n$.

- The unbiased estimator of σ^2 is $s_e^2 = \frac{SSResid}{n-2}$, i.e. $\mathrm{E}[s_e^2] = \sigma^2$

  where $SSResid = \sum e_i^2 = \sum(y - \hat{y})^2 = S_{YY} - b^2 S_{XX}$.

- The number of degrees of freedom associated with estimating σ^2 or σ in simple linear regression is $n - 2$.

- Residual Mean Square: $MSResid = \frac{SSResid}{n-2}$.

- $\frac{SSResid}{\sigma^2} \sim \chi^2_{n-2}$

Calculating R^2

- The coefficient of determination is defined as
$$R^2 = 1 - \frac{SSResid}{SSTo},$$
where $SSTo = \sum(y-\bar{y})^2 = S_{yy}$.
- the value of σ represents the magnitude of a typical deviation of a point (x, y) in the population from the population regression line.
- Similarly, s_e is the magnitude of a typical sample deviation (residual) from the least-squares line.
- The smaller the value of s_e, the closer the points in the sample fall to the line and the better the line does in predicting y from x.

Example

A simple linear regression model was used to describe the relationship between *y* = hardness of molded plastic and *x* = amount of time elapsed since the end of the molding process. Summary quantities included *n* =15, SSResid = 1235.470, and SSTo = 25,321.368.

a. Calculate a point estimate of σ. On how many degrees of freedom is the estimate based?

b. What percentage of observed variation in hardness can be explained by the simple linear regression model relationship between hardness and elapsed time?

Inferences About the Slope of the Population Regression Line

- Properties of the Sampling Distribution of b

  1. $\mu_b = E(b) = \beta$, so the sampling distribution of b is always centered at the value of $\beta. i. e\ b$ is an unbiased statistic for estimating β.
  2. The S.D of the statistic b is $\sigma_b = \frac{\sigma}{\sqrt{S_{xx}}}$.
  3. The statistic b is normally distributed.

- The normality of the sampling distribution of b implies that the standardized variable $z = \frac{b-\beta}{\sigma_b}$ has a standard normal distribution.
- However, inferential methods cannot be based on this variable, because the value of σ_b is not known.
- One way to proceed is to estimate σ with s_e, yielding an estimated standard deviation.

Inferences About the Slope of the Population Regression Line

- The estimated standard deviation of the statistic b is

$$s_b = \frac{s_e}{\sqrt{S_{xx}}} = \sqrt{\frac{SSRes}{(n-2)S_{xx}}}$$

- When the four basic assumptions of the simple linear regression model are satisfied, the probability distribution of the standardized variable

$$t = \frac{b-\beta}{s_b}$$

  is the t-distribution with df $n-2$.

- Confidence Interval for β : $b \pm (t_{n-2,\frac{\alpha}{2}})\cdot s_b$

Hypothesis Tests Concerning β

- Null Hypothesis : $H_0\colon\ \beta = 0$
- Alternative Hypothesis: $H_1\colon\ \beta \neq 0$
- Test Statistic : $t = \frac{b - hypothesized\ value}{s_b}$
- df : $n - 2$.
- Reject $H_0\colon\ \beta = 0$ if $|t| > t_{n-2,\alpha/2}$.
- Under H_0 the population regression line is a horizontal line, and the value of y in the simple linear regression model does not depend on x. That is, $y = \alpha + e$.
- In this situation, knowledge of x is of no use in predicting y.
- If $\beta \neq 0$, there is a useful linear relationship between x and y, and knowledge of x is useful for predicting y.
- The test of $H_0\colon \beta = 0$ versus $H_a\colon \beta \neq 0$ is called the *model utility test for simple linear regression.*

Example

An experiment to study the relationship between *x* = time spent exercising (minutes) and *y* = amount of oxygen consumed during the exercise period resulted in the following summary statistics.

$$n = 20, \quad \sum x = 50, \quad \sum y = 16{,}705,$$

$$\sum x^2 = 150, \quad \sum y^2 = 14{,}194231, \quad \sum xy = 44{,}194$$

a) Estimate the slope and *y* intercept of the population regression line.

b) One sample observation on oxygen usage was 757 for a 2-minute exercise period. What amount of oxygen consumption would you predict for this exercise period, and what is the corresponding residual?

c) Compute a 99% confidence interval for the average change in oxygen consumption associated with a 1-minute increase in exercise time.

Example

A simple linear regression model was used to describe the relationship between sales revenue *y* (in thousands of dollars) and advertising expenditure *x* (also in thousands of dollars) for fast-food outlets during a 3-month period. A sample of 15 outlets yielded the accompanying summary quantities.

$$\sum x = 14.10 \quad \sum y = 1438.50 \quad \sum x^2 = 13.92 \quad \sum y^2 = 140{,}354$$

$$\sum xy = 1387.20 \quad \sum (y - \bar{y})^2 = 2401.85 \quad \sum (y - \hat{y})^2 = 561.46$$

a. What proportion of observed variation in sales revenue can be attributed to the linear relationship between revenue and advertising expenditure?

b. Calculate s_e and s_b.

c. Obtain a 90% confidence interval for β, the average change in revenue associated with a $1000 (that is, 1-unit) increase in advertising expenditure.

d. Test the hypothesis H_0: $\beta = 0$ versus H_a: $\beta \neq 0$ using a significance level of .05. What does your conclusion say about the nature of the relationship between *x* and *y*?

Interval Estimation of Mean Response E(Y) for given $x = x_0$

- SLR: $Y = \alpha + \beta x + \epsilon$
- $E[Y|x = x_0] = \alpha + \beta x_0.$
- $a + bx_0$ is an unbiased estimator of expected response at $x = x_0$
- $Var[a + bx_0] = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$
- Sampling Distribution of $a + bx_0$: *When* σ *known*

$$a + bx_0 \sim N\left[\alpha + \beta x_0, \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right]$$

- Sampling Distribution of $a + bx_0$: *When* σ *unknown*

$$\frac{a + bx_0 - \alpha + \beta x_0}{\sqrt{\frac{SSRes}{n-2}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \sim t_{n-2}.$$

- Confidence interval on $E[Y|x = x_0] = \alpha + \beta x_0$:

$$a + bx_0 \pm t_{n-2,\frac{\alpha}{2}} \sqrt{\frac{SSRes}{n-2}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

Prediction Interval for a Single y

- We want to predict y_0 at $x = x_0$.

$$y_0 = \alpha + \beta x_0 + \epsilon$$

- $\hat{y}_0 = \ a + bx_0$.
- Vertical discrepancy between y_0 and $\hat{y}_0$: $U = y_0 - \hat{y}_0$.
- $E[U] = 0$
- $Var[U] = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$
- Sampling Distribution : When σ known

$$U \sim N\left(0, \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\right)$$

- Sampling Distribution : When σ unknown

$$\frac{U - 0}{\sqrt{\frac{SSRes}{n-2}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \sim t_{n-2}$$

- Prediction Interval for y_0:

$$\hat{y}_0 \pm t_{n-2,\frac{\alpha}{2}} \sqrt{\frac{SSRes}{n-2}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

Exercise:

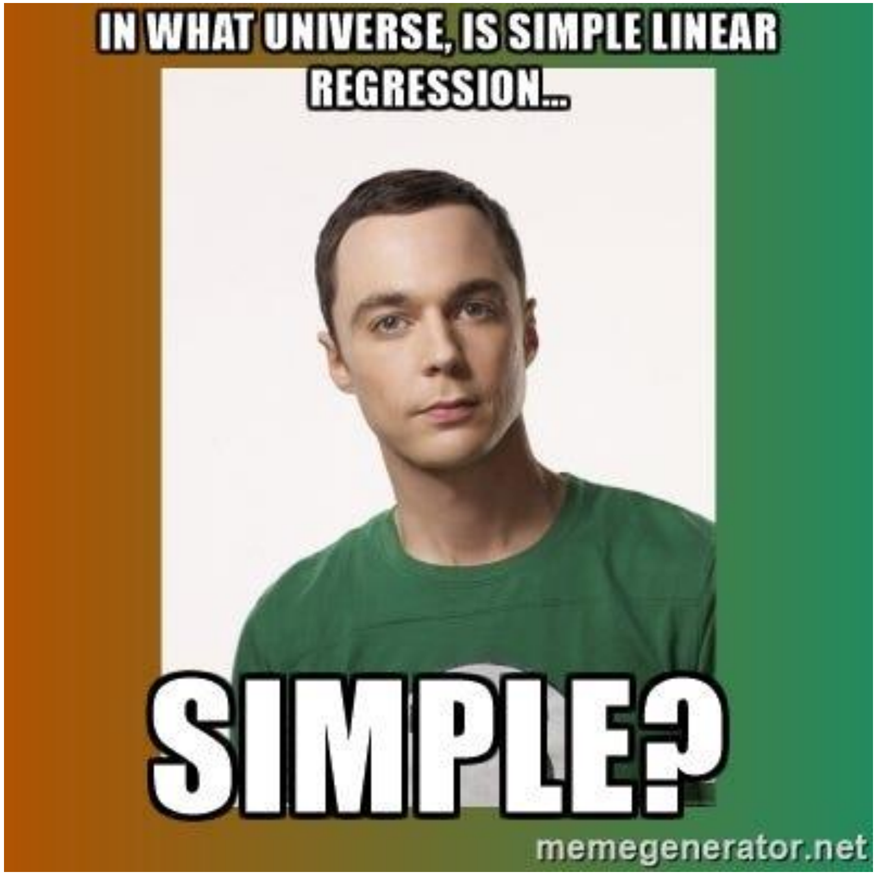
The article **“Effect of Temperature on the pH of Skim Milk” (*Journal of Dairy Research* [1988]: 277– 280)** reported on a study involving x = temperature (°C) under specified experimental conditions and y = milk pH. The accompanying data (read from a graph) are a representative subset of that which appeared in the article:

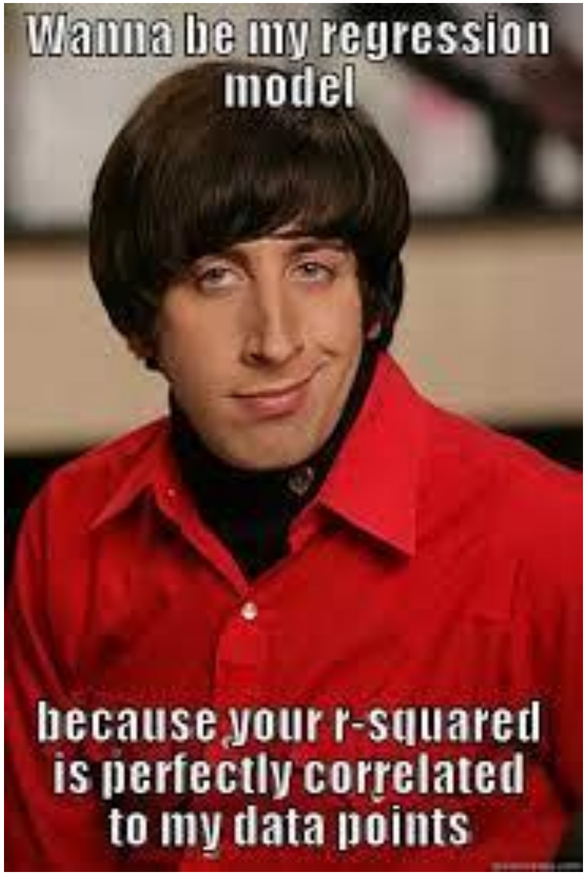
x :	4	4	24	24	25	38	38	40
y:	6.85	6.79	6.63	6.65	6.72	6.62	6.57	6.52
x :	45	50	55	56	60	67	70	78
y:	6.50	6.48	6.42	6.41	6.38	6.34	6.32	6.34

$\sum x = 678 \;\; \sum y = 104.54 \;\; \sum x^2 = 36{,}056 \;\; \sum y^2 = 683.4470 \;\; \sum xy = 4376.36$

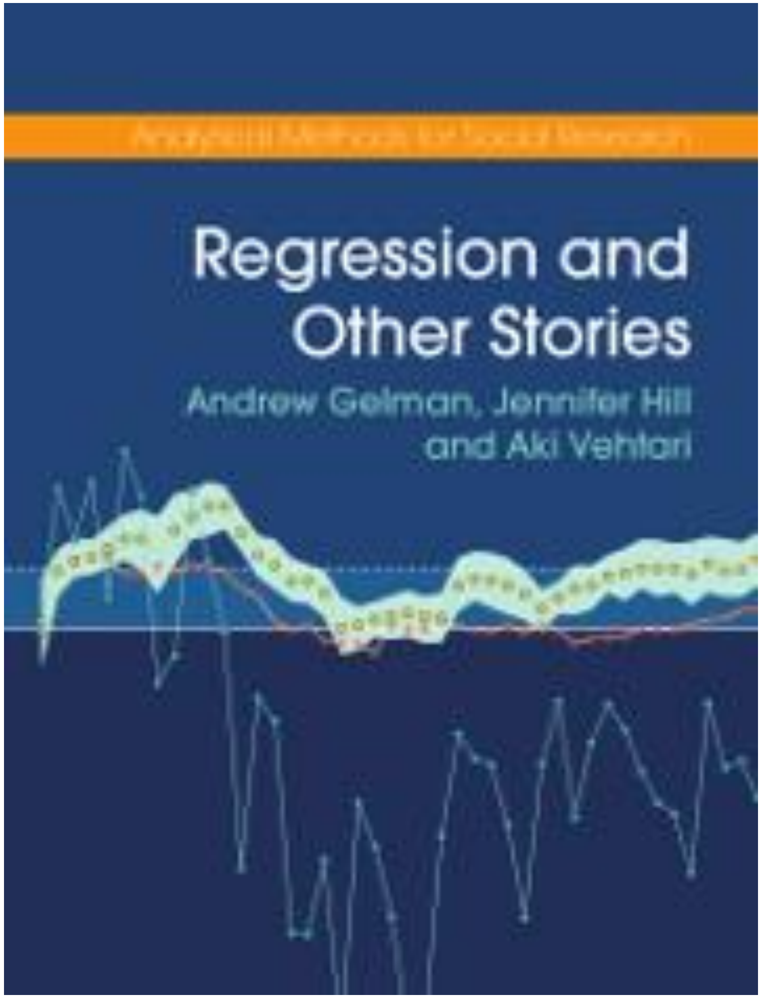
- Do these data strongly suggest that there is a negative linear relationship between temperature and pH? State and test the relevant hypotheses using a significance level of .01.
- Obtain a 95% confidence interval for $\alpha + \beta(40)$ the mean milk pH when the milk temperature is 40°C.
- Obtain a 95% prediction interval for a single pH observation to be made when milk temperature = 40°C.
- Would you recommend using the data to calculate a 95% confidence interval for the mean pH when the temperature is 90°C? Why or why not?

Comments


IN WHAT UNIVERSE, IS SIMPLE LINEAR REGRESSION...
SIMPLE?
memegenerator.net


Wanna be my regression model
because your r-squared is perfectly correlated to my data points

Reference



Analytical Methods for Social Research

Regression and Other Stories

Andrew Gelman, Jennifer Hill and Aki Vehtari