BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Statistical Methods

ISM Team

innovate achieve lead

BITS Pilani
Pilani Campus

Session 11
Regression & Correlation
(15th / 16th February 2025)

IMP Note to Self

Agenda

Regression

- Linear
- Non-linear
- Polynomial Regression
- Multiple Regression
- Logistic Regression

Correlation

Objective of Regression

- To exploit the relationship between two (or more) variables so that we can gain information about one of them through knowing values of the other(s), when the relation between the variable is not in a deterministic form.
- We try to find the deterministic linear relation $\boldsymbol{y = \beta_0 + \beta_1 x}$ from that non deterministic form of data and we call it as *linear probabilistic model.*

Devore, J. L. (2011). Probability and Statistics for Engineering and the Sciences. Brooks/C

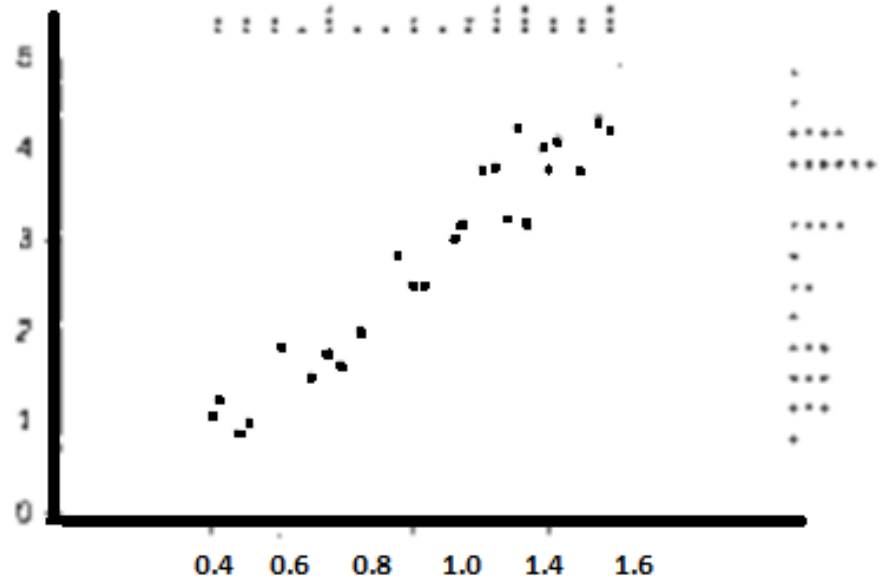
Simple Linear Regression Model

- The data available is in the form of (x_i, y_i), pairs of n tuples.
- The variable whose value is fixed by the experimenter will be denoted by x, called the **independent, predictor,** or **explanatory variable.**
- For fixed x, the second variable will be ***random***; we denote this random variable and its observed value by Y and y, respectively, and refer it as the **dependent** or **response variable.**
- **scatter plot** is used to visualize the nature of any relationship. In such a plot, each (x_i, y_i) is represented as a point plotted on a two-dimensional coordinate system.

Scatter Plot

i	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
x_i	1.15	1.20	1.25	1.25	1.28	1.30	1.34	1.37	1.40	1.43	1.46	1.49	1.55	1.58	1.60
y_i	3.18	3.76	3.68	3.82	3.21	4.27	3.12	3.99	3.75	4.10	4.18	3.77	4.34	4.21	4.92

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_i	.40	.42	.48	.51	.57	.60	.70	.75	.75	.78	.84	.95	.99	1.03	1.12
y_i	1.02	1.21	.88	.98	1.52	1.83	1.50	1.80	1.74	1.63	2.00	2.80	2.48	2.47	3.05



Observations:

- There is a strong tendency for *y* to increase as *x* increases
- there is evidence of a substantial (though not perfect) linear relationship between the two variables
- The value of *y* could be predicted from *x* by finding a line that is reasonably close to the points in the plot

A Linear Probabilistic Model

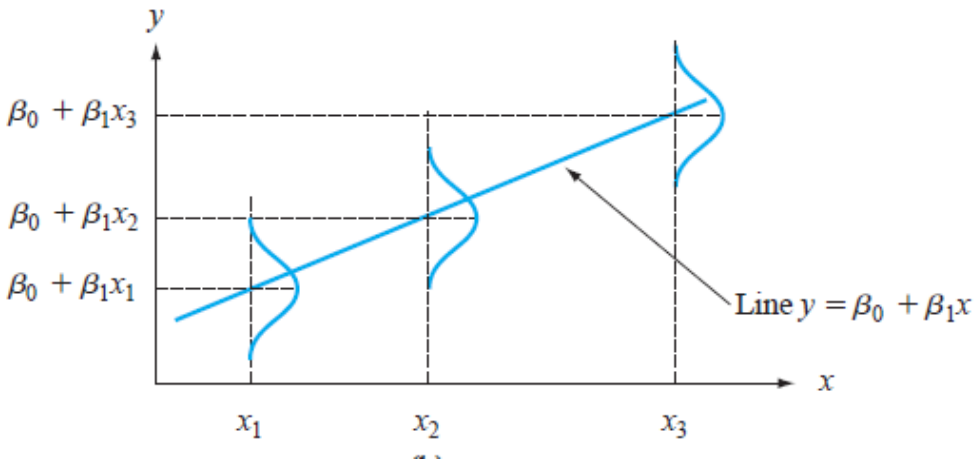
$$Y = \beta_0 + \beta_1 x + \epsilon$$

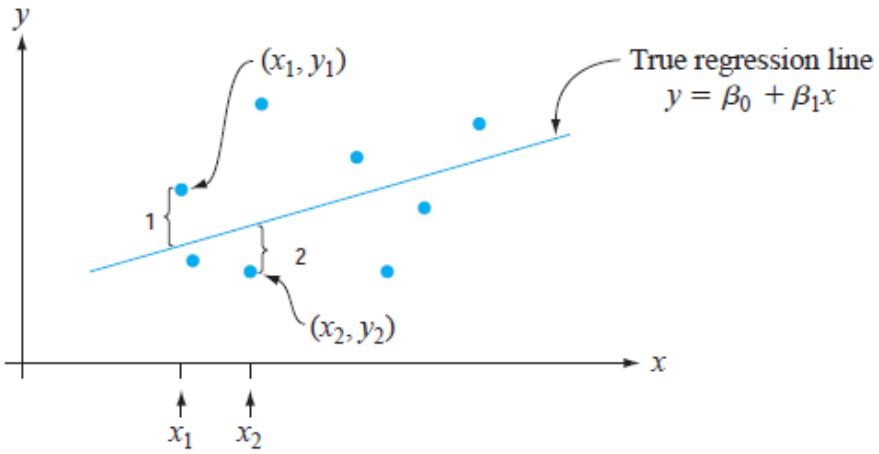
- There are Parameters β_0, $\beta_1$and σ^2, such that for any fixed value of the independent variable *x*, the dependent variable is a random variable related to *x* through this **model equation.**

- The variable ϵ is usually referred to as the **random deviation** or **random error term** in the model. Without ϵ, any observed pair (x_i, y_i) would correspond to a point falling exactly on the line , called the **true** (or **population**) **regression line.**

- The ϵ can either the positive or negative, which makes the (x_i, y_i) to fall on either side of line of regression.

Linear Regression line:





Least Square method

Principle of Least Squares

The vertical deviation of the point (x_i, y_i) from the line $y = b_0 + b_1x$ is

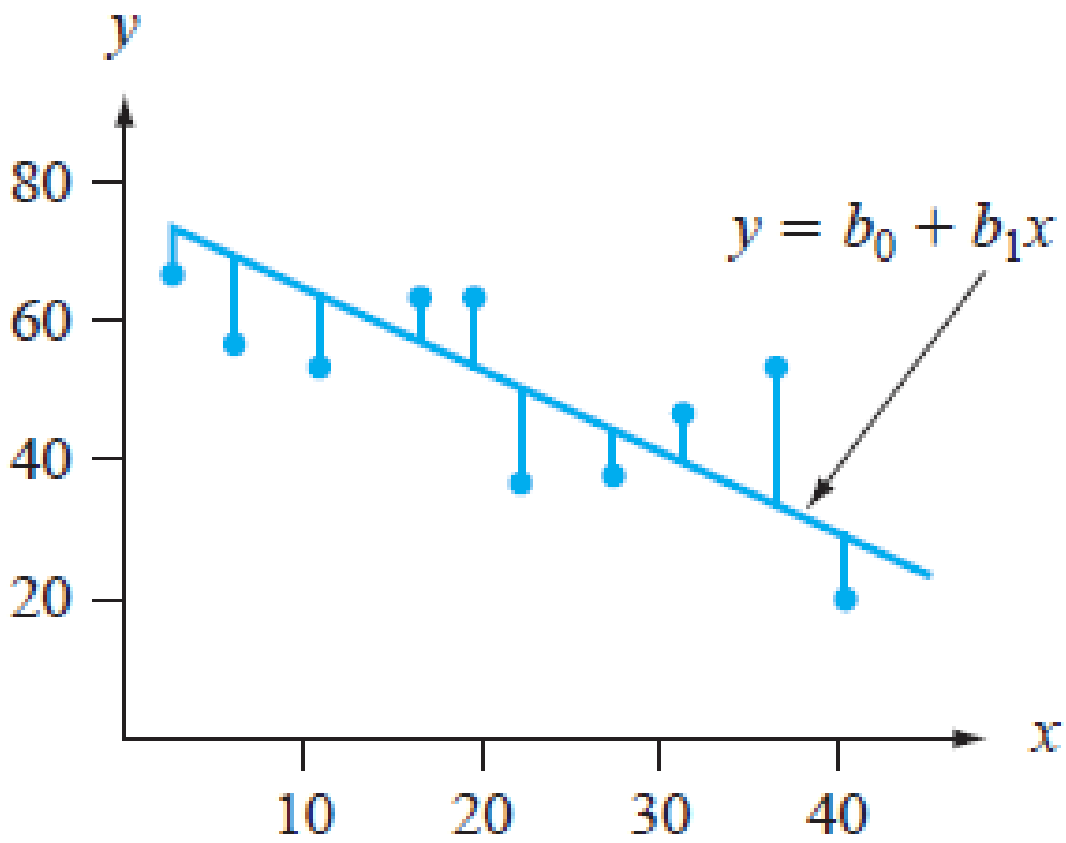
$$\text{height of point} - \text{height of line} = y_i - (b_0 + b_1x_i)$$

The sum of squared vertical deviations from the points $(x_1, y_1), \ldots, (x_n, y_n)$ to the line is then

$$f(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1x_i)]^2$$

The point estimates of β_0 and β_1, denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$ and called the **least squares estimates,** are those values that minimize $f(b_0, b_1)$. That is, $\hat{\beta}_0$ and $\hat{\beta}_1$ are such that $f(\hat{\beta}_0, \hat{\beta}_1) \leq f(b_0, b_1)$ for any b_0 and b_1. The **estimated regression line** or **least squares line** is then the line whose equation is $y = \hat{\beta}_0 + \hat{\beta}_1 x$.

Least Square line:



Least Square method

$$f(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

The minimizing values of b_0 and b_1 are found by taking partial derivatives of $f(b_0, b_1)$ with respect to both b_0 and b_1, equating them both to zero [analogously to $f'(b) = 0$ in univariate calculus], and solving the equations

$$\frac{\partial f(b_0, b_1)}{\partial b_0} = \sum 2(y_i - b_0 - b_1 x_i)(-1) = 0$$
$$\frac{\partial f(b_0, b_1)}{\partial b_1} = \sum 2(y_i - b_0 - b_1 x_i)(-x_i) = 0$$

Cancellation of the -2 factor and rearrangement gives the following system of equations, called the **normal equations:**

$$nb_0 + (\sum x_i) b_1 = \sum y_i$$
$$(\sum x_i) b_0 + (\sum x_i^2) b_1 = \sum x_i y_i$$

These equations are linear in the two unknowns b_0 and b_1. Provided that not all x_i's are identical, the least squares estimates are the unique solution to this system.

Least Square Estimation

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

- The least squares estimate of the slope coefficient β_1 of the true regression line is:

$$b_1 = \widehat{\beta_1} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

Where,

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

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}$$

The least squares estimate of the intercept β_0 of the true regression line is

$$b_0 = \widehat{\beta_0} = \frac{\sum y_i - \widehat{\beta_1}\sum x_i}{n} = \bar{y} - \widehat{\beta_1}\bar{x}$$

Estimated Regression Line $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$

Example

Find the line of regression and coefficient of regression (y on x) for the given data:

x	y
132.0	46.0
129.0	48.0
120.0	51.0
113.2	52.1
105.0	54.0
92.0	52.0
84.0	59.0
83.2	58.7
88.4	61.6
59.0	64.0
80.0	61.4
81.5	54.6
71.0	58.8
69.2	58.0

$$S_{xx} = \sum_{i=1}^{n} x_i{}^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} \qquad S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n} \qquad \widehat{\beta_1} = \frac{S_{xy}}{S_{xx}} \qquad b_0 = \bar{y} - \widehat{\beta_1}\bar{x} \qquad \hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$$

Example

x	y	x^2	y^2	x*y
132.0	46.0	17424	2116	6072
129.0	48.0	16641	2304	6192
120.0	51.0	14400	2601	6120
113.2	52.1	12814.24	2714.41	5897.72
105.0	54.0	11025	2916	5670
92.0	52.0	8464	2704	4784
84.0	59.0	7056	3481	4956
83.2	58.7	6922.24	3445.69	4883.84
88.4	61.6	7814.56	3794.56	5445.44
59.0	64.0	3481	4096	3776
80.0	61.4	6400	3769.96	4912
81.5	54.6	6642.25	2981.16	4449.9
71.0	58.8	5041	3457.44	4174.8
69.2	58.0	4788.64	3364	4013.6
1,307.5	**779.2**	**128,913.9**	**43,745.2**	**71,347.3**

$S_{xx} = 128913.93 - (1307.5)^2/14$

= 6802.7693

$S_{xy} = 71347.3 - (1307.5)(779.2)/14$

= -1424.41429

$\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}} = \frac{-1424.41429}{6802.7693} = -0.20938742$

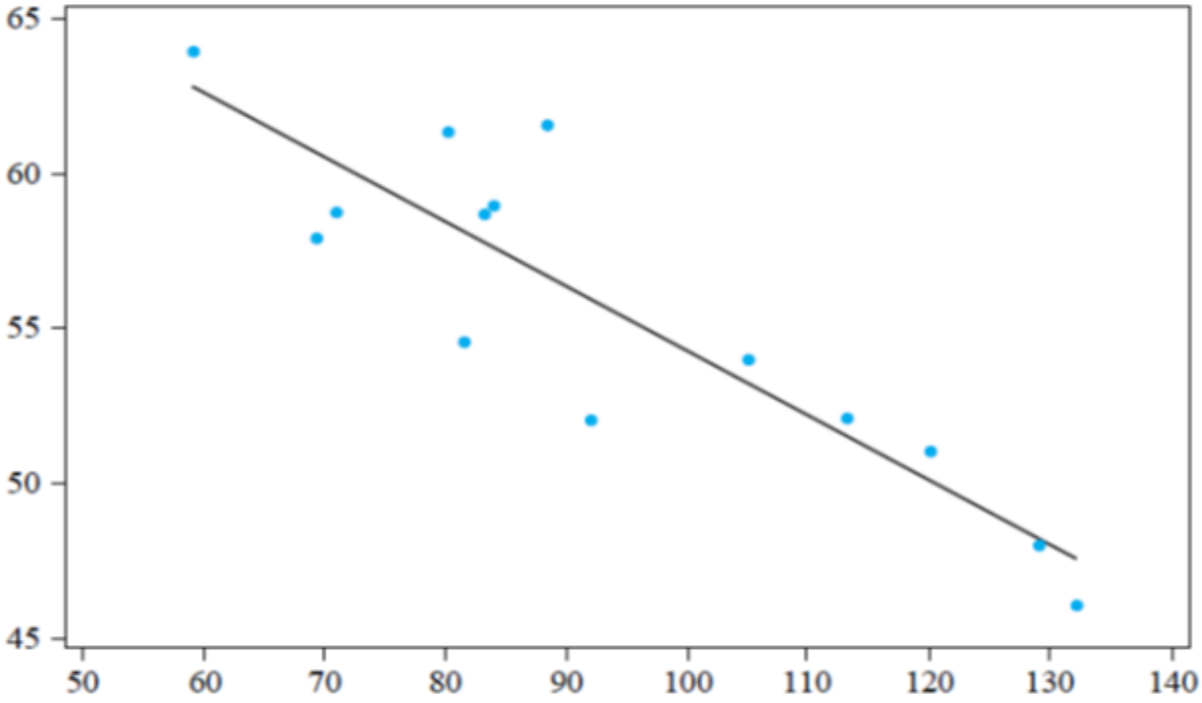
$\widehat{\beta_0} = 55.657143 - (-0.20938742)(93.392857)$
$= 75.212432$

The line of regression will be

$\hat{y} = 75.212432 - 0.20938742x$

Example

- Line of regression $\hat{y} = 75.212432 - 0.20938742x$
- This equation can be used to predict any value within the range, but not beyond the range.

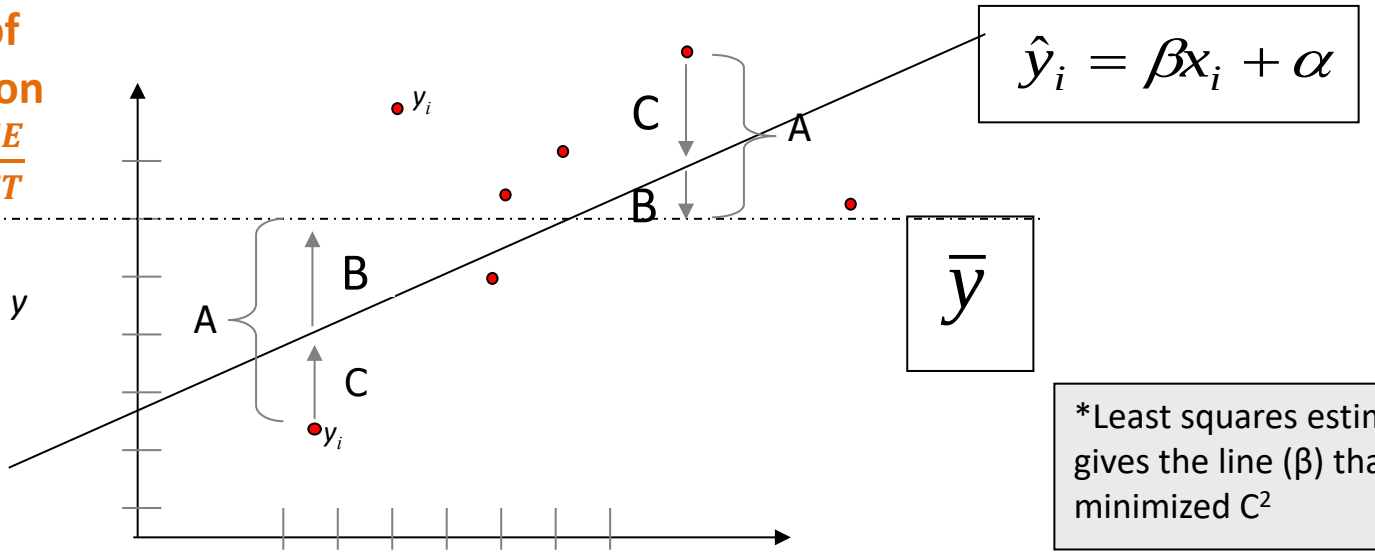


Regression Picture

(SSE_{reg}/SST_{total}) *(Proportion of total variation that cannot be explained by Simple Linear Regression Model)*

Coefficient of Determination

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



*Least squares estimation gives the line (β) that minimized C^2

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

A^2 B^2 C^2

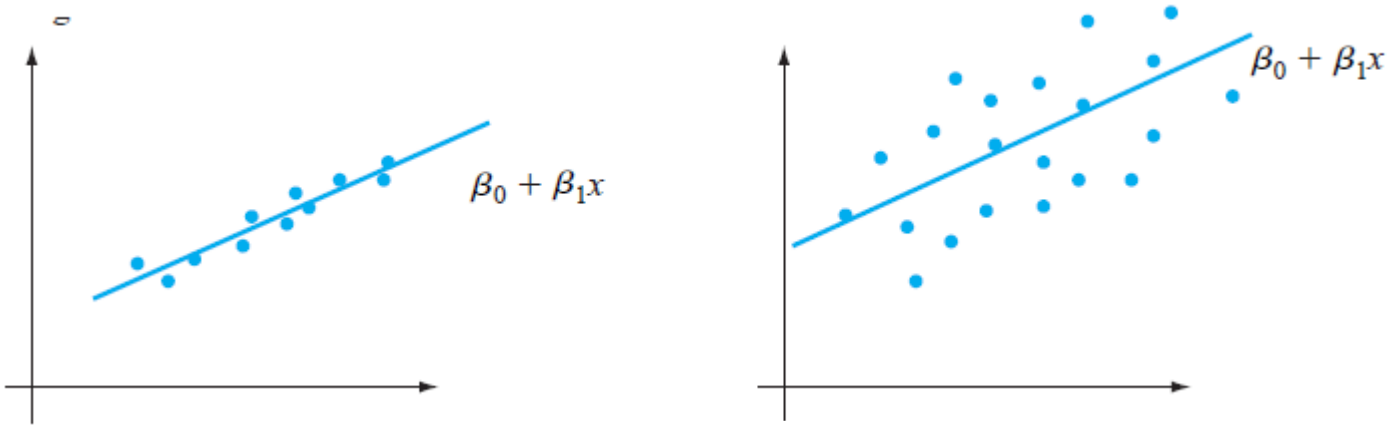
SST_{total} (S_{yy})
Total squared distance of observations from naïve mean of y
Total variation

Ss_{reg}
Distance from regression line to naïve mean of y
Variability due to x (regression)

$SSE_{residual}$
Variance around the regression line **(Error Sum of Square)**
Additional variability not explained by x—what least squares method aims to minimize

Estimating σ^2 and σ

- The parameter σ^2 determines the amount of variability inherent in the regression model.
- A large value of σ^2 will lead the observed points to spread out about the true regression line, whereas when σ^2 is small the observed points will tend to fall very close to the true line.



Typical sample for σ^2: (a) small; (b) large

Devore, J. L. (2011). Probability and Statistics for Engineering and the Sciences. Brooks/C

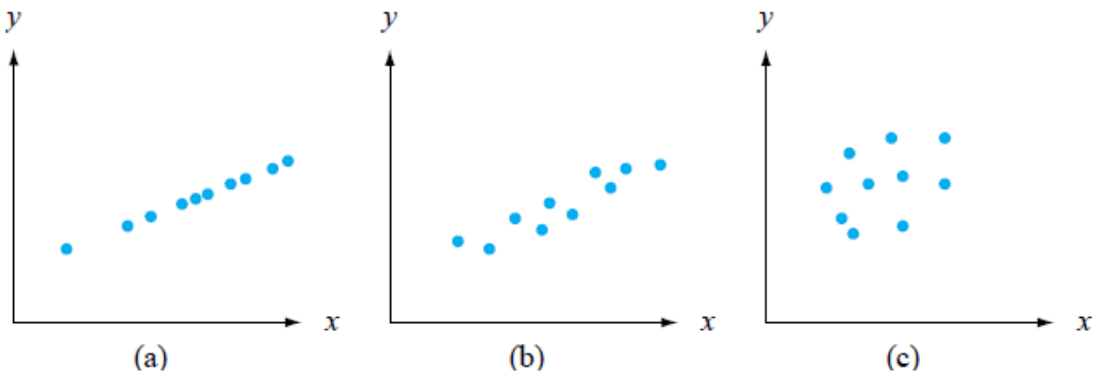
Residuals (error sum of squares)

- The **residuals** are the differences $y_1 - \hat{y}_1, y_2 - \hat{y}_2, y_3 - \hat{y}_3, \ldots . y_n - \hat{y}_n$, between the observed and fitted *y* values.
  - A positive number, if the point lies above the line and a negative number, if it lies below the line.
  - When the estimated regression line is obtained via the principle of least squares, the sum of the residuals should in theory be zero.
  - *In practice, the sum may deviate a bit from zero due to rounding.*
- The **error sum of squares** (equivalently, residual sum of squares), denoted by **SSE**, is:
- $\boldsymbol{SSE = \sum(y_i - \hat{y}_i)^2 = \sum[y_i - (\widehat{\beta_0} + \widehat{\beta_1} x)]^2 \; or \; \sum y_i^2 - \widehat{\beta_0}\sum y_i - \widehat{\beta_1}\sum x_i y_i}$
- and the estimate of σ^2 is $\boldsymbol{\hat{\sigma}^2 = s^2 = \frac{SSE}{n-2} = \frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$
  - *The divisor n-2 in s^2 is the number of degrees of freedom (df) associated with SSE and the estimate s^2. This is because to obtain s^2, the two parameters $\boldsymbol{\beta_0}$ **and** $\boldsymbol{\beta_1}$ must first be estimated, which results in a loss of 2 df*

Devore, J. L. (2011). Probability and Statistics for Engineering and the Sciences. Brooks/C

The Coefficient of Determination

- The points in the first plot all fall exactly on a straight line.
- The points in the second can be estimated by least squares line.
- The points in the third, there is substantial variation about the least squares line relative to overall *y* variation, so the simple linear regression model fails to explain variation in *y* by relating *y* to *x*.
- **The error sum of squares SSE can be interpreted as a measure of how much variation in *y* is left unexplained by the model—i.e., how much cannot be attributed to a linear relationship.**
- SSE = 0 (for a)
- SSE = small (for b)
- SSE = large (for c)



Using the model to explain y variation: (a) data for which all variation is explained; (b) data for which most variation is explained; (c) data for which little variation is explained

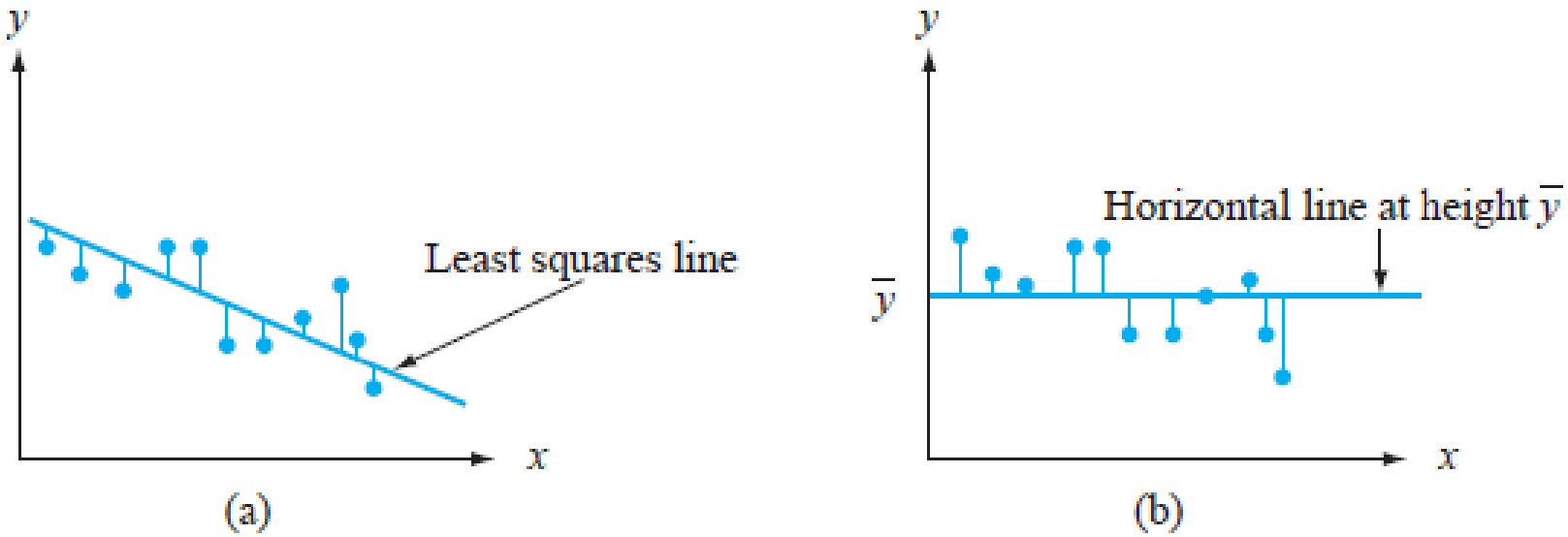
Total Sum of Squares (SST)

- A quantitative measure of the total amount of variation in observed *y* values is given by the **total sum of squares (SST).**

$$SST = S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$$

- Total sum of squares is the sum of squared deviations about the sample mean of the observed *y* values.
  - In SST the same number $\bar{y}$ is subtracted from each y_i,
  - whereas SSE involves subtracting each different predicted value $\hat{y}_i$ from the corresponding observed y_i.
- SST is the sum of squared deviations about the horizontal line at height $\bar{y}$.

Total Sum of Squares (SST)



Sums of squares illustrated: (a) SSE = sum of squared deviations about the least squares line; (b) SST = sum of squared deviations about the horizontal line

Devore, J. L. (2011). Probability and Statistics for Engineering and the Sciences. Brooks/C

Total Sum of Squares (SST)

- As the sum of squared deviations about the least squares line is smaller than the sum of squared deviations about *any* other line, **SSE<SST**, unless the horizontal line itself is the least squares line.
- The ratio **SSE/SST** is the proportion of total variation that cannot be explained by the simple linear regression model, and
- **1- SSE/SST** (a number between 0 and 1) is the proportion of observed *y* variation explained by the model, which is **coefficient of determination,** denoted by $\mathbf{r^2}$,

Devore, J. L. (2011). Probability and Statistics for Engineering and the Sciences. Brooks/C

Total Sum of Squares (SST)

- ***It is interpreted as the proportion of observed y variation that can be explained by the simple linear regression model***
  - If $\mathbf{r^2}$ is more Simple Linear Regression is successful in explaining *y* variation.
  - If $\mathbf{r^2}$ is small, an analyst will usually want to search for an alternative model (either a nonlinear model or a multiple regression model that involves more than a single independent variable)
- It also can be written in a slightly different way by introducing a third sum of squares—**regression sum of squares,** SSR—given by
- $\boldsymbol{SSR = \sum(\hat{y}_i - \bar{y})^2 = SST - SSE}$
- Regression sum of squares is interpreted as the amount of total variation that *is* explained by the model.
- $r^2 = 1 - SSE/SST = \frac{SST - SSE}{SST} = SSR/SST$

Devore, J. L. (2011). Probability and Statistics for Engineering and the Sciences. Brooks/C

Regression Coefficients (Summary)

Least Squares Estimators

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

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

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}$$

Least Squares Estimates

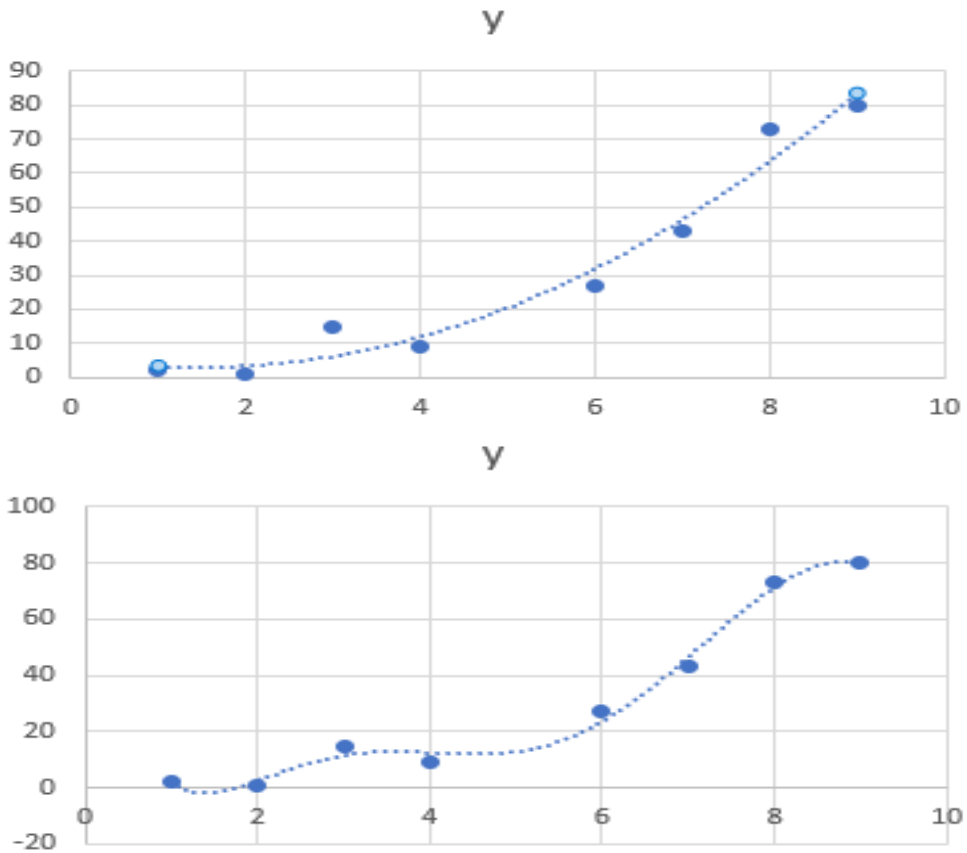
$$a = \bar{y} - b.\bar{x} \; \boldsymbol{and} \; b = \frac{S_{xy}}{S_{xx}}$$

Estimated Regression Line

$$\hat{\boldsymbol{y}} = \boldsymbol{a} + \boldsymbol{bx}$$

Regression (Overfitting)

- Which of these two models would be a better fit to the data?



Overfitting means that your model makes not accurate predictions. It is based on the training data in data science

The total sum of squares (TSS) measures how much variation there is in the observed data, while the **residual sum of squares (RSS)** measures the variation in the error between the observed data and modeled (fit)values.

innovate achieve lead

BITS Pilani
Pilani Campus

Non-Linear Regression

Non-Linear Regression

- The situation when the observed value of the dependent variable *Y* deviated from the linear regression function by a random amount, and the data in not best fit by linear probabilistic model.
- Then we have two ways for this situation:
  - The first way is to replace $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$ by a non-linear function of *x.*
    - intrinsically linear
    - Polynomial Regression
  - The second is to use a regression function involving more than a single independent variable *(Multiple regression)*

Assessing Model Adequacy

1. First step is to plot of the observed pairs (x_i, y_i) to decide the form of a mathematical relationship between x and y.
2. Choose the function: linear or non-linear regression method.
3. Once a function of the chosen form has been fitted, it is checked for the best fit by superimposing the graph of the best-fit function on the scatter plot of the data.
4. Any tilt or curvature of the best-fit function may obscure some aspects of the fit that should be investigated.

Residuals and Standardized Residuals

A more effective approach to assessment of model adequacy is to compute the fitted or predicted values $\hat{y}_i$ and the residuals $e_i = y_i - \hat{y}_i$ and then plot various functions of these computed quantities. We then examine the plots either to confirm our choice of model or for indications that the model is not appropriate. Suppose the simple linear regression model is correct, and let $y = \hat{\beta}_0 + \hat{\beta}_1 x$ be the equation of the estimated regression line. Then the ith residual is $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$. To derive properties of the residuals, let $e_i = Y_i - \hat{Y}_i$, represent the ith residual as a random variable (rv) before observations are actually made. Then

$$E(Y_i - \hat{Y}_i) = E(Y_i) - E(\hat{\beta}_0 + \hat{\beta}_1 x_i) = \beta_0 + \beta_1 x_i - (\beta_0 + \beta_1 x_i) = 0 \quad (13.1)$$

Residuals and Standardized Residuals

Because $\hat{Y}_i (= \hat{\beta}_0 + \hat{\beta}_1 x_i)$ is a linear function of the Y_j's, so is $Y_i - \hat{Y}_i$ (the coefficients depend on the x_j's). Thus the normality of the Y_j's implies that each residual is normally distributed. It can also be shown that

$$V(Y_i - \hat{Y}_i) = \sigma^2 \cdot \left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}\right] \qquad (13.2)$$

Replacing σ^2 by s^2 and taking the square root of Equation (13.2) gives the estimated standard deviation of a residual.

Let's now standardize each residual by subtracting the mean value (zero) and then dividing by the estimated standard deviation.

The standardized residuals

The **standardized residuals** are given by

$$e_i^* = \frac{y_i - \hat{y}_i}{s\sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}}} \quad i = 1, \ldots, n \qquad (13.3)$$

Example:

The data is:

x_i	y_i	$\hat{y}_i$	e_i	e_i^*
100	150	125.6	24.4	.75
125	140	168.4	−28.4	−.84
125	180	168.4	11.6	.35
150	210	211.1	−1.1	−.03
150	190	211.1	−21.1	−.62
200	320	296.7	23.3	.66
200	280	296.7	−16.7	−.47
250	400	382.3	17.7	.50
250	430	382.3	47.7	1.35
300	440	467.9	−27.9	−.80
300	390	467.9	−77.9	−2.24
350	600	553.4	46.6	1.39
400	610	639.0	−29.0	−.92
400	670	639.0	31.0	.99

The Line estimated line of regression is

$$y = -45.55 + 1.71x \; and \; r^2 = .961$$

Devore, J. L. (2011). Probability and Statistics for Engineering and the Sciences. Brooks/C

Diagnostic Plots

Diagnostic Plots

The basic plots that many statisticians recommend for an assessment of model validity and usefulness are the following:

1. e_i^* (or e_i) on the vertical axis versus x_i on the horizontal axis
2. e_i^* (or e_i) on the vertical axis versus $\hat{y}_i$ on the horizontal axis
3. $\hat{y}_i$ on the vertical axis versus y_i on the horizontal axis
4. A normal probability plot of the standardized residuals

Plots 1 and 2 are called **residual plots** (against the independent variable and fitted values, respectively), whereas Plot 3 is fitted against observed values.

If Plot 3 yields points close to the 45° line [slope +1 through (0, 0)], then the estimated regression function gives accurate predictions of the values actually observed. Thus Plot 3 provides a visual assessment of model effectiveness in making predictions. Provided that the model is correct, neither residual plot should exhibit distinct patterns. The residuals should be randomly distributed about 0 according to a normal distribution,

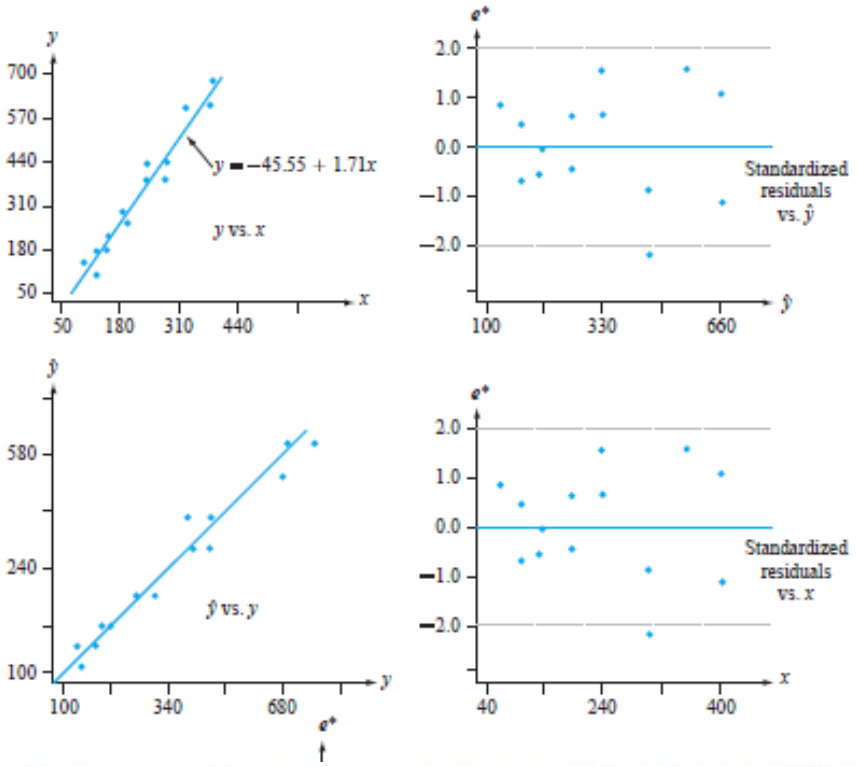


Figure presents a scatter plot of the data and the four plots just recommended. The plot of $\hat{y}$ versus y confirms the impression given by r^2 that x is effective in predicting y and also indicates that there is no observed y for which the predicted value is terribly far off the mark. Both residual plots show no unusual pattern or discrepant values. There is one standardized residual slightly outside the interval $(-2, 2)$, but this is not surprising in a sample of size 14. The normal probability plot of the standardized residuals is reasonably straight. In summary, the plots leave us with no qualms about either the appropriateness of a simple linear relationship or the fit to the given data.

Difficulties and Remedies

1. A nonlinear probabilistic relationship between x and y is appropriate.
2. The variance of ϵ (and of Y) is not a constant σ^2 but depends on x.
3. The selected model fits the data well except for a very few discrepant or outlying data values, which may have greatly influenced the choice of the best-fit function.
4. The error term ϵ does not have a normal distribution.
5. When the subscript i indicates the time order of the observations, the ϵ_i's exhibit dependence over time.
6. One or more relevant independent variables have been omitted from the model.

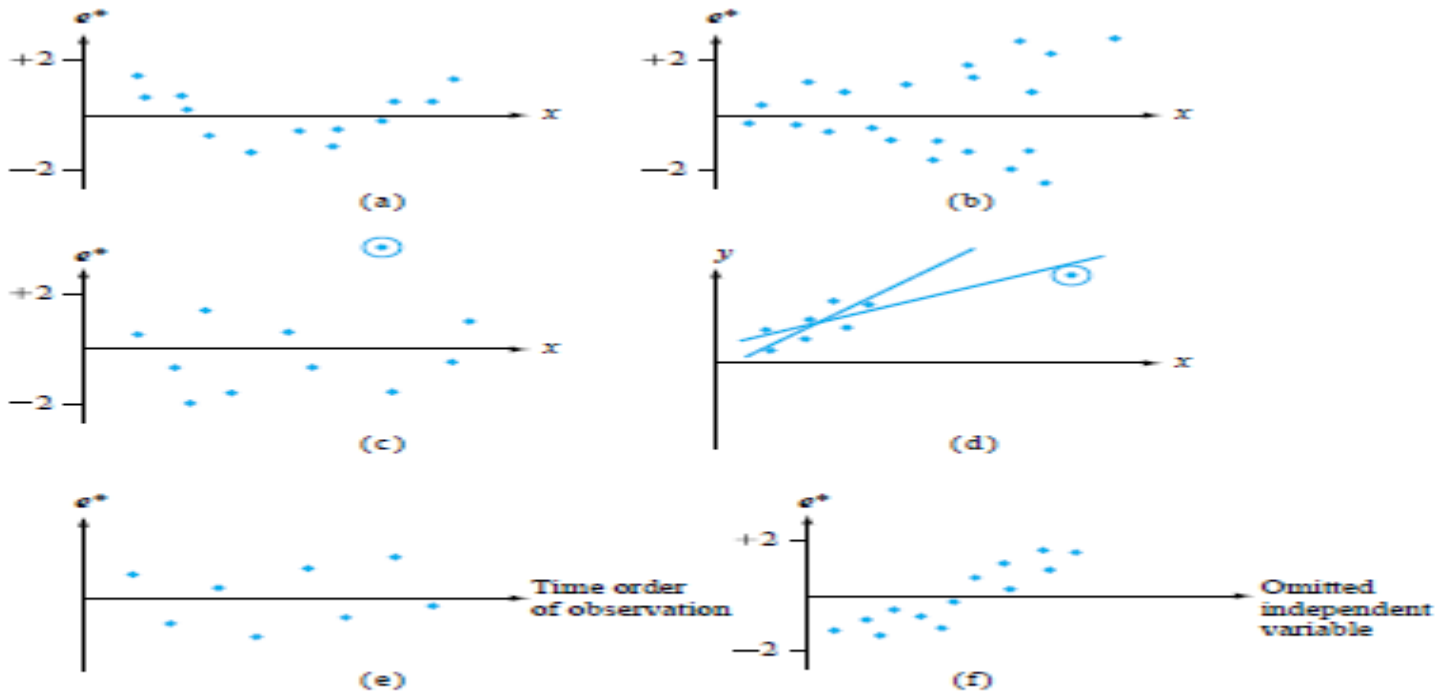


Figure 13.2 Plots that indicate abnormality in data: (a) nonlinear relationship; (b) nonconstant variance; (c) discrepant observation; (d) observation with large influence; (e) dependence in errors; (f) variable omitted

Regression with Transformed Variables

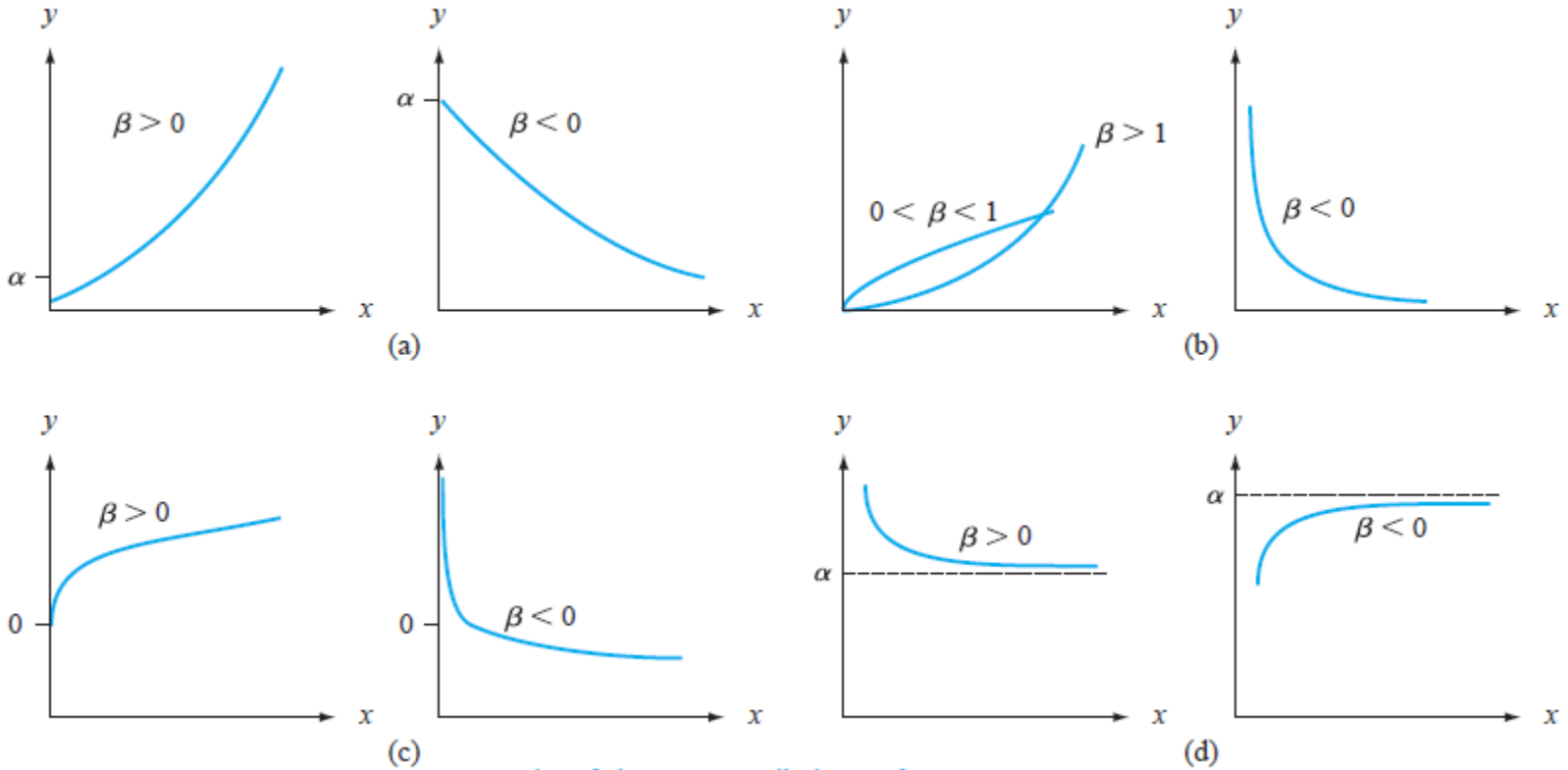
A function relating y to x is **intrinsically linear** if, by means of a transformation on x and/or y, the function can be expressed as $y' = \beta_0 + \beta_1 x'$, where $x' =$ the transformed independent variable and $y' =$ the transformed dependent variable.

Useful Intrinsically Linear Functions*

Function	Transformation(s) to Linearize	Linear Form
a. Exponential: $y = \alpha e^{\beta x}$	$y' = \ln(y)$	$y' = \ln(\alpha) + \beta x$
b. Power: $y = \alpha x^{\beta}$	$y' = \log(y), x' = \log(x)$	$y' = \log(\alpha) + \beta x'$
c. $y = \alpha + \beta \cdot \log(x)$	$x' = \log(x)$	$y = \alpha + \beta x'$
d. Reciprocal: $y = \alpha + \beta \cdot \frac{1}{x}$	$x' = \frac{1}{x}$	$y = \alpha + \beta x'$

*When $\log(\cdot)$ appears, either a base 10 or a base e logarithm can be used.

Intrinsically linear functions



Graphs of the intrinsically linear functions

Regression with Transformed Variables

Examples of functions that are not intrinsically linear are:

$$y = \alpha + \gamma e^{\beta x}$$

$$y = \alpha + \gamma x^{\beta}$$

Regression with Transformed Variables

- The major advantage of an intrinsically linear model is that the parameters $\boldsymbol{\beta_0}$and $\boldsymbol{\beta_1}$ of the transformed model can be immediately estimated using the principle of least squares simply by substituting *x'* and *y'* into the estimating formulas.

- Parameters of the original nonlinear model can then be estimated by transforming back $\widehat{\boldsymbol{\beta_0}}$and/or $\widehat{\boldsymbol{\beta_1}}$if necessary

$$\hat{\beta}_1 = \frac{\sum x'_i y'_i - \sum x'_i \sum y'_i / n}{\sum (x'_i)^2 - (\sum x'_i)^2 / n}$$

$$\hat{\beta}_0 = \frac{\sum y'_i - \hat{\beta}_1 \sum x'_i}{n} = \bar{y}' - \hat{\beta}_1 \bar{x}'$$

Devore, J. L. (2011). Probability and Statistics for Engineering and the Sciences. Brooks/C

Example

S.No	x	y	x' = ln(x)	y'=ln(y)	x'*x'	y'*y'	x'*y'	Beta1	Beta0	Alpha	line = alpha.x^Beta1
1	600	2.35	6.396929655	0.854415328	40.92070901	0.730025553	5.465634751	-5.39974349	35.66931485	3.09732E+15	3.088036872
2	600	2.65	6.396929655	0.97455964	40.92070901	0.949766492	6.234189462				3.088036872
3	600	3.00	6.396929655	1.098612289	40.92070901	1.206948961	7.027745529				3.088036872
4	600	3.60	6.396929655	1.280933845	40.92070901	1.640791516	8.194043702				3.088036872
5	500	6.40	6.214608098	1.85629799	38.62135382	3.445842229	11.53616452				8.264962651
6	500	7.80	6.214608098	2.054123734	38.62135382	4.219424313	12.76557399				8.264962651
7	500	9.80	6.214608098	2.282382386	38.62135382	5.209269354	14.18411206				8.264962651
8	500	16.50	6.214608098	2.803360381	38.62135382	7.858829425	17.42178613				8.264962651
9	400	21.50	5.991464547	3.068052935	35.89764742	9.412948813	18.38213039				27.57592616
10	400	24.50	5.991464547	3.198673118	35.89764742	10.23150971	19.16473658				27.57592616
11	400	26.00	5.991464547	3.258096538	35.89764742	10.61519305	19.5207699				27.57592616
12	400	33.00	5.991464547	3.496507561	35.89764742	12.22556513	20.94920109				27.57592616
Sum	6000	157.1	74.412009	26.2260157	461.758841	67.7461145	160.846088	-5.4	35.669	3.1E+15	155.7157027
Average	500	13.0917	6.2010008	2.18550131	38.47990342	5.64550955	13.4038407	-5.4	35.669	3.1E+15	12.97630856

y = alpha*x^Beta

innovate achieve lead

BITS Pilani
Pilani Campus

Polynomial Regression

Polynomial Regression

- The non-linear yet intrinsically linear models involved functions of the independent variable x were either strictly increasing or strictly decreasing.
- In many situations, either theoretical reasoning or else a scatter plot of the data suggests that the true regression function has one or more peaks or valleys, i.e., at least one relative minimum or maximum.
- In such cases, a polynomial function may provide a satisfactory approximation to the true regression function.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots .. + \beta_k x^k$$

Polynomial Regression

The *k*th-degree polynomial regression model equation is

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \epsilon$$

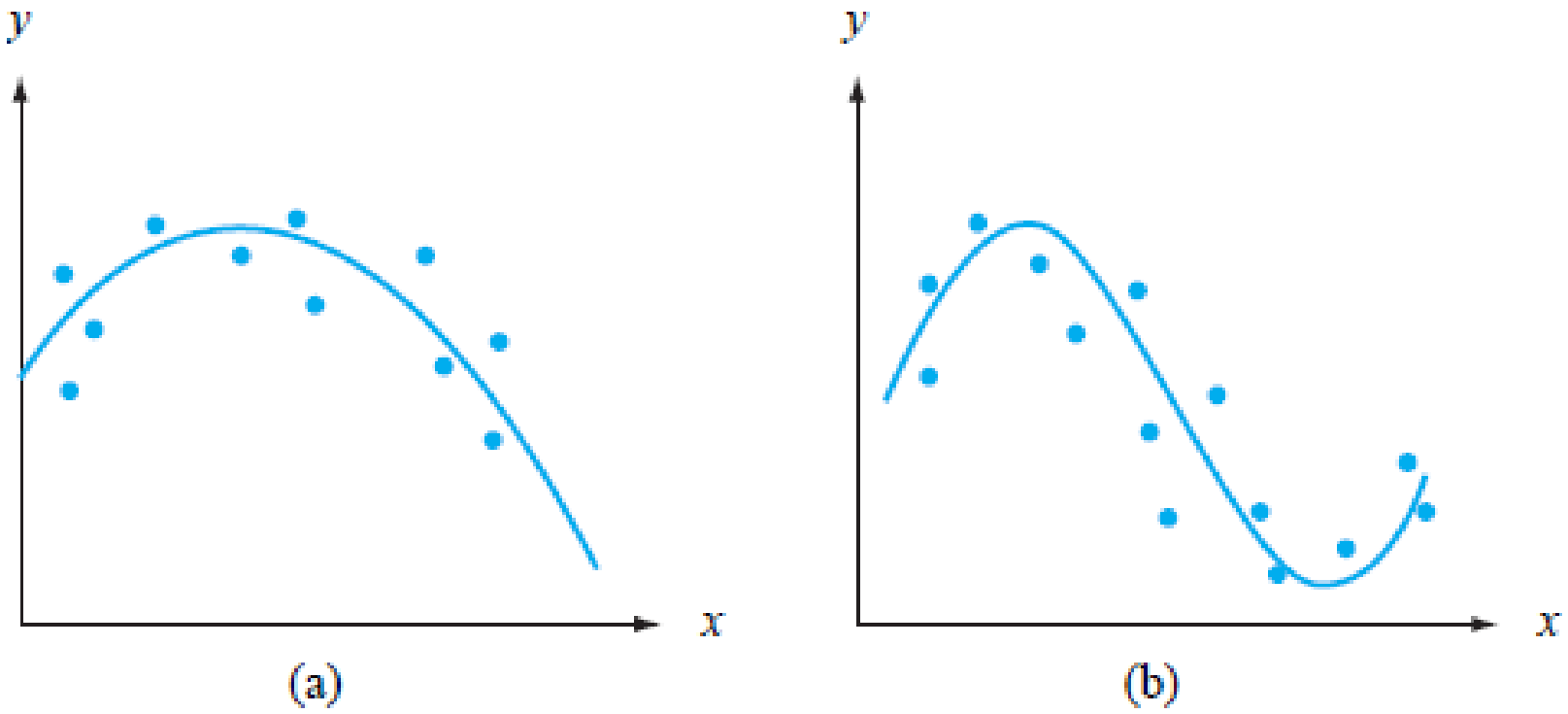
where ϵ is a normally distributed random variable with

$$\mu_\epsilon = 0 \qquad \sigma_\epsilon^2 = \sigma^2$$

$$\begin{aligned}
&b_0 n + b_1 \Sigma x_i + b_2 \Sigma x_i^2 + \cdots + b_k \Sigma x_i^k = \Sigma y_i \\
&b_0 \Sigma x_i + b_1 \Sigma x_i^2 + b_2 \Sigma x_i^3 + \cdots + b_k \Sigma x_i^{k+1} = \Sigma x_i y_i \\
&\quad\vdots \qquad\qquad\quad \vdots \qquad\qquad\quad \vdots \\
&b_0 \Sigma x_i^k + b_1 \Sigma x_i^{k+1} + \cdots + b_k \Sigma x_i^{2k} = \Sigma x_i^k y_i
\end{aligned}$$

Very rarely in practice is it necessary to go beyond $k = 3$.

Polynomial Regression



(a) Quadratic regression model; (b) cubic regression model

innovate achieve lead

BITS Pilani
Pilani Campus

Multiple Regression

Multiple (Linear) Regression

- Multiple regression analysis is a straightforward extension of simple regression analysis which allows more than one independent variable.

- In multiple regression, the objective is to build a probabilistic model that relates a dependent variable *y* to more than one independent or predictor variable.

- Let *k* represent the number of predictor variables (k>2) and denote these predictors by $x_1, x_2, x_3, \ldots x_k$.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots .. + \beta_k x_k + \epsilon$$

Multiple (Linear) Regression

- $\Sigma y_1^2 = \Sigma y_1^2 - (\Sigma y_1)^2 / n$
- $\Sigma y_2^2 = \Sigma y_2^2 - (\Sigma y_2)^2 / n$
- $\Sigma y_1 x = \Sigma y_1 x - (\Sigma y_1 \Sigma x) / n$
- $\Sigma y_2 x = \Sigma y_2 x - (\Sigma y_2 \Sigma x) / n$
- $\Sigma y_1 y_2 = \Sigma y_1 y_2 - (\Sigma y_1 \Sigma y_2) / n$

Multiple (Linear) Regression

S.No	x	y1	y2	y1^2	y2^2	x*y1	x*y2	y1*y2
1	4	3	23	9	529	12	92	69
2	5	8	32	64	1024	40	160	256
3	2	9	28	81	784	18	56	252
4	6	4	60	16	3600	24	360	240
5	7	3	62	9	3844	21	434	186
6	8	1	43	1	1849	8	344	43
7	7	6	60	36	3600	42	420	360
8	3	3	37	9	1369	9	111	111
9	5	2	24	4	576	10	120	48
10	5	5	64	25	4096	25	320	320
11	7	2	28	4	784	14	196	56
12	8	1	66	1	4356	8	528	66
13	5	7	35	49	1225	35	175	245
14	2	5	37	25	1369	10	74	185
15	4	0	59	0	3481	0	236	0
16	6	2	32	4	1024	12	192	64
17	5	6	76	36	5776	30	380	456
18	7	5	25	25	625	35	175	125
19	9	0	55	0	3025	0	495	0
20	8	3	34	9	1156	24	272	102
21	7	5	54	25	2916	35	378	270
22	9	1	57	1	3249	9	513	57
Sum	**129**	**81**	**991**	**433**	**50257**	**421**	**6031**	**3511**
Regression Sums				**134.8**	**5617**	**-53.95**	**220.14**	**-137.7**

Multiple (Linear) Regression

- $\Sigma y_1^2 = \Sigma y_1^2 - (\Sigma y_1)^2 / n$

  $= 433 - (81)^2 / 22 =$ **134.7727**

- $\Sigma y_2^2 = \Sigma y_2^2 - (\Sigma y_2)^2 / n$

  $= 50257 - (991)^2 / 22 =$ **5616.9545**

- $\Sigma y_1 x = \Sigma y_1 x - (\Sigma y_1 \Sigma x) / n$

  $= 421 - (81*129) / 22 =$ **-53.9545**

- $\Sigma y_2 x = \Sigma y_2 x - (\Sigma y_2 \Sigma x) / n$

  $= 6031 - (991*129) / 22 =$ **220.1364**

- $\Sigma y_1 y_2 = \Sigma y_1 y_2 - (\Sigma y_1 \Sigma y_2) / n$

  $= 3511 - (81*991) / 22 =$ **-137.6818**

Multiple (Linear) Regression

- $b_1 = [(\Sigma y_2^2)(\Sigma y_1 x) - (\Sigma y_1 y_2)(\Sigma y_2 x)] \; / \; [(\Sigma y_1^2)(\Sigma y_2^2) - (\Sigma y_1 y_2)^2]$
- $b_2 = [(\Sigma y_1^2)(\Sigma y_2 x) - (\Sigma y_1 y_2)(\Sigma y_1 x)] \; / \; [(\Sigma y_1^2)(\Sigma y_2^2) - (\Sigma y_1 y_2)^2]$
- $b_0 = \bar{x} - b_1\bar{y}_1 - b_2\bar{y}_2$
- $\hat{y} = b_0 + b_1 * y_1 + b_2 * y_2$
- $\hat{Y} = 0.45898 + (-0.0018)y_1 + (0.12013)y_2$
- As it fits a line, it is a linear model.
- There are also non-linear regression models involving multiple variables, such as logistic regression, quadratic regression, and probit models.

Logistic Regression

Logistic Regression

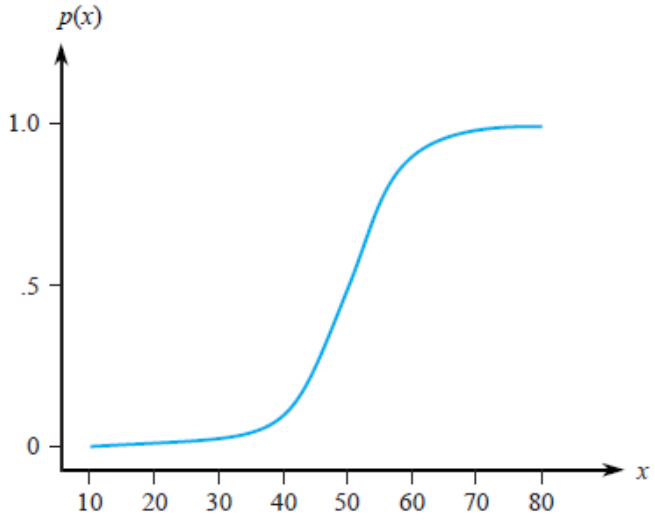
- Logistic Regression is used when the outcome variable is categorical.
- The independent variables could be either categorical or continuous.
- The slope coefficient in the Logistic Regression Model has a relationship with the OR (*Odd Ratio*)
- This regression type is used when the output expected is binary (success or failure/ 0 or 1, Yes or No).
- A function that has been used for this type of regression is the **logit function:**

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

Logistic Regression

- The graph of $p(x)$ for particular values of $\beta_0 \; and \; \beta_1$ with $\beta_1 > 0$. As x increases, the probability of success increases.
- For β_1 negative, the success probability would be a decreasing function of x.

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$



A graph of a logit function

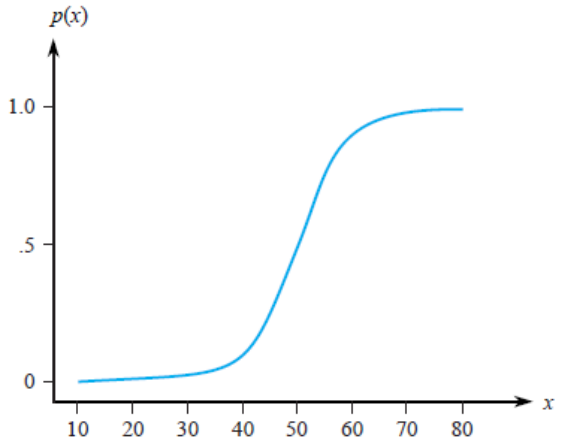
Logistic Regression

- *Logistic regression* means assuming that $p(x)$ is related to x by the logit function.
- We know,

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

Rearranging, we get

$$\frac{\boldsymbol{p(x)}}{\boldsymbol{1 - p(x)}} = \boldsymbol{e^{\beta_0+\beta_1 x}}$$



A graph of a logit function

Devore, J. L. (2011). Probability and Statistics for Engineering and the Sciences. Brooks/C

innovate achieve lead

BITS Pilani

Pilani Campus

Correlation

Correlation

- Co-relation
  - The sample correlation coefficient *r is* a measure of, how strongly related two variables *x* and *y* are, in a sample
- Most popularly seen correlation coefficient: Pearson Product-Moment Correlation

Types of Correlation

1. Positive correlation
   - High values of X tend to be associated with high values of Y.
   - As X increases, Y increases
2. Negative correlation
   - High values of X tend to be associated with low values of Y.
   - As X increases, Y decreases
3. No correlation
4. No consistent tendency for values on Y to increase or decrease as X increases

Covariance

- 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 co-variance is:

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

Devore, J. L. (2011). Probability and Statistics for Engineering and the Sciences. Brooks/C

Calculating by hand…

$$\hat{r} = \frac{\text{covariance}(x,y)}{\sqrt{\text{var } x}\sqrt{\text{var } y}} = \frac{\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}}{\sqrt{\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}\sqrt{\dfrac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}}$$

Devore, J. L. (2011). Probability and Statistics for Engineering and the Sciences. Brooks/C

Simple formula…

$$\hat{r} = \frac{\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\cancel{n-1}}}{\sqrt{\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\cancel{n-1}}}\sqrt{\dfrac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\cancel{n-1}}}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

Numerator of covariance

$$\hat{r} = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

Numerators of variance

Devore, J. L. (2011). Probability and Statistics for Engineering and the Sciences. Brooks/C

Example

S.No	x	y	x^2	y^2	x*y	Sxx	Syy	Sxy	r
1	0.066	4.600	0.004356	21.16	0.3036	0.025516	434.5375	2.3826	0.715535541
2	0.088	11.600	0.007744	134.56	1.0208				
3	0.120	9.500	0.0144	90.25	1.14				
4	0.050	6.300	0.0025	39.69	0.315				
5	0.162	13.800	0.026244	190.44	2.2356				
6	0.186	15.400	0.034596	237.16	2.8644				
7	0.057	2.500	0.003249	6.25	0.1425				
8	0.100	11.800	0.01	139.24	1.18				
9	0.112	8.000	0.012544	64	0.896				
10	0.055	7.000	0.003025	49	0.385				
11	0.154	20.600	0.023716	424.36	3.1724				
12	0.074	16.600	0.005476	275.56	1.2284				
13	0.111	9.200	0.012321	84.64	1.0212				
14	0.140	17.900	0.0196	320.41	2.506				
15	0.071	2.800	0.005041	7.84	0.1988				
16	0.110	13.000	0.0121	169	1.43				
Sum	**1.656**	**170.6**	**0.1969**	**2253.6**	**20.04**	**0.0255**	**434.54**	**2.3826**	**0.7155**
Average	**0.1035**	**10.663**	**0.0123**	**140.85**	**1.2525**	**0.0255**	**434.54**	**2.3826**	**0.7155**

Correlation Coefficient

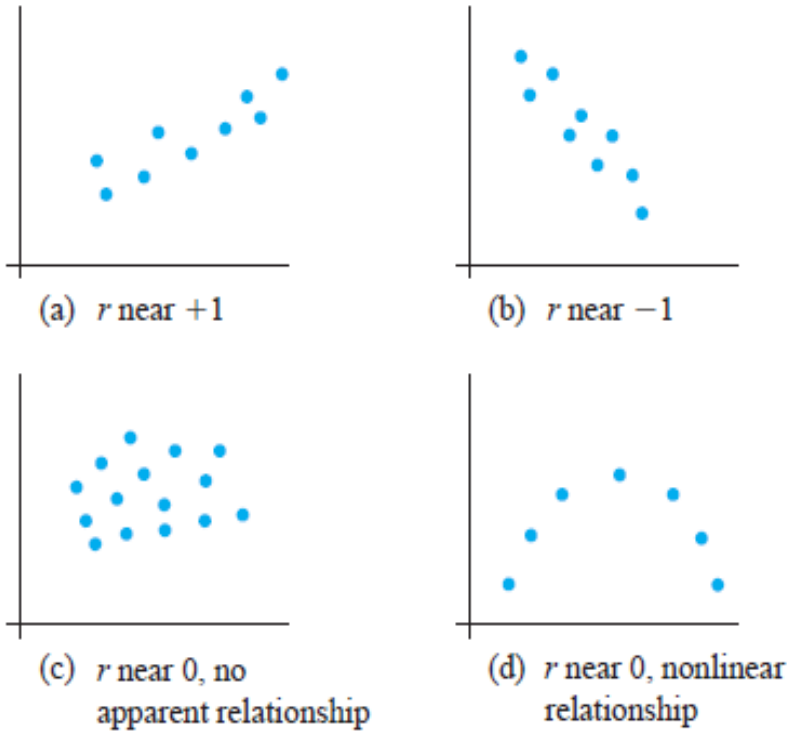
Properties of r

The most important properties of r are as follows:

1. The value of r does not depend on which of the two variables under study is labeled x and which is labeled y.
2. The value of r is independent of the units in which x and y are measured.
3. $-1 \leq r \leq 1$
4. $r = 1$ if and only if (iff) all (x_i, y_i) pairs lie on a straight line with positive slope, and $r = -1$ iff all (x_i, y_i) pairs lie on a straight line with negative slope.
5. The square of the sample correlation coefficient gives the value of the coefficient of determination that would result from fitting the simple linear regression model—in symbols, $(r)^2 = r^2$.

Correlation Coefficient

- Thus, ***r measures the degree of linear relationship* among variables**.
  - A value of *r* near 0 is not evidence of the lack of a strong relationship, but only the absence of a linear relation, so that such a value of *r* must be interpreted with caution.



Data plots for different values of *r*

Weak	Moderate	Strong
$-.5 \le r \le .5$	either $-.8 < r < -.5$ or $.5 < r < .8$	either $r \ge .8$ or $r \le -.8$

Correlation

High positive correlation

Zero correlation

High negative correlation

stronger

+1.00

perfect positive

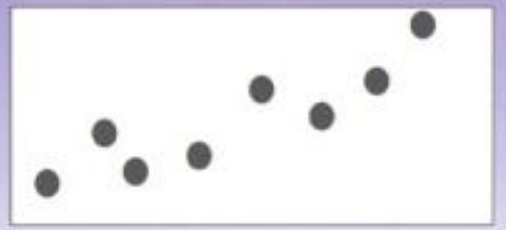
as one event increases, the second exactly increases

weaker

+.50

positive

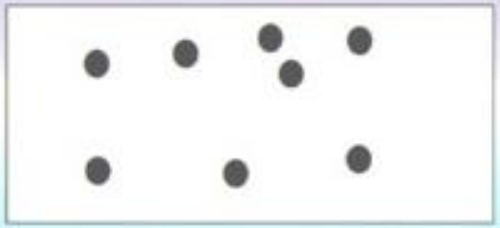
as one event increases, the second sometimes increases



0

zero correlation

no relationship between the events

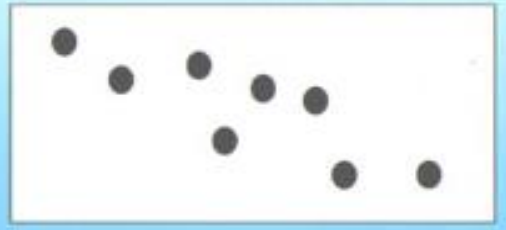


weaker

negative

-.50

as one event increases, the second sometimes decreases

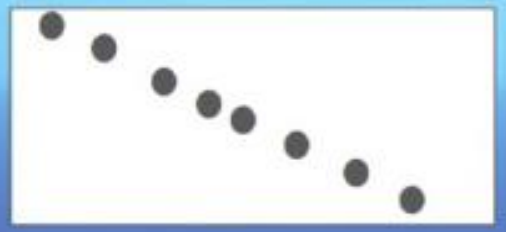


stronger

perfect negative

as one event increases, the second exactly decreases

-1.00



The Pearson Product Moment Correlation Coefficient

- The correlation coefficient is the single number that represents the degree of relation between two variables.
- The Pearson Product-Moment Correlation Coefficient (symbolized by *r*) is the most common measure of correlation; researchers calculate it when both the X variable and the Y variable are interval or ratio scale measurements.
- The formula for *r* is:

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

Devore, J. L. (2011). Probability and Statistics for Engineering and the Sciences. Brooks/C

Assumptions of Pearson's Correlation Coefficient

- There is linear relationship between two variables, i.e. when the two variables are plotted on a scatter diagram a straight line will be formed by the points.
- Cause and effect relation exists between different forces operating on the item of the two-variable series.

Example:

TABLE 8-1 CALCULATING A PEARSON PRODUCT–MOMENT CORRELATION COEFFICIENT USING THE RAW-SCORE FORMULA

	SAT (X)	X^2	GPA (Y)	Y^2	XY
Student 1	980	960,400	2.02	4.0804	1979.60
Student 2	1070	1,144,900	2.45	6.0025	2621.50
Student 3	1020	1,040,400	2.63	6.9169	2682.60
Student 4	1240	1,537,600	3.11	9.6721	3856.40
Student 5	880	774,400	2.09	4.3681	1839.20
Student 6	1110	1,232,100	2.75	7.5625	3052.50
Student 7	1350	1,822,500	3.72	13.8384	5022.00
Student 8	1080	1,166,400	2.38	5.6644	2570.40
Student 9	1230	1,512,900	3.06	9.3636	3763.80
Student 10	1470	2,160,900	3.48	12.1104	5115.60
	$\Sigma = 11{,}430$	13,352,500	27.69	79.5793	32503.60
	$M = 1143$		$M = 2.769$		

$$(\Sigma X)(\Sigma Y) = (11{,}430)(27.69) = 316{,}496.70$$

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

$$= \frac{(10)(32{,}503.60) - 316{,}496.70}{\sqrt{[10(13{,}352{,}500.00) - (11{,}430)(11{,}430)][10(79.5793) - (27.69)(27.69)}}$$

$$= \frac{8539.30}{\sqrt{(2{,}880{,}100)(29.0569)}} = \frac{8539.30}{\sqrt{83686777.69}} = \frac{8539.30}{9148.0478} = .93$$

Limitations of Pearson's Coefficient

- Always assume linear relationship
- Interpreting the value of "r" is difficult.
- Value of Correlation Coefficient is affected by the extreme values.
- Time consuming method.

Advantages of Pearson's Coefficient

- It summarizes in one value, the degree of correlation & direction of correlation also.

Coefficient of Determination

- The convenient way of interpreting the value of correlation coefficient is to use the square of coefficient of correlation value which is called Coefficient of Determination (r^2).
- Coefficient of Determination = Explained variation / Total variation

- Suppose: r = 0.9, r^2 = 0.81 this would mean that 81% of the variation in the dependent variable has been explained by the independent variable.
- The maximum value of r^2 is 1.

Devore, J. L. (2011). Probability and Statistics for Engineering and the Sciences. Brooks/C

Example

- Suppose: r = 0.60

  r = 0.30

- It does not mean that the first correlation is twice as strong as the second;
- the 'r' can be understood by computing the value of r^2 .

  When r = 0.60 $r^2 = 0.36$ -----(1)

  r = 0.30 $r^2 = 0.09$ -----(2)

- This implies that in the first case 36% of the total variation is explained whereas in second case 9% of the total variation is explained .

Practice Problems

Example 1

- An educational economist wants to establish the relationship between an individual's income and education. He takes a random sample of 10 individuals and asks for their income (in $1000s) and education (in years). The results are shown below. Find the least squares regression line.

Education	11	12	11	15	8	10	11	12	17	11
Income	25	33	22	41	18	28	32	24	53	26

Solution:

- The least squares regression line is

$$\hat{y} = -13.93 + 3.74x$$

- Interpretation of coefficients:
- The sample slope $\hat{\beta}_1 = 3.74$ tells us that on average for each additional year of education, an individual's income rises by \$3.74 thousand.

- The y-intercept is $\hat{\beta}_0 = -13.93$. This value is the expected (or average) income for an individual who has 0 education level (which is meaningless here).

Example 2

- Car dealers across North America use the red book to determine a cars selling price on the basis of important features.
- To examine this issue 100 three-year old cars in mint condition were randomly selected. Their selling price and odometer reading were observed.

Odometer	Price
37388	5318
44758	5061
45833	5008
30862	5795
…..	…
34212	5283
33190	5259
39196	5356
36392	5133

Example 3

- In baseball, the fans are always interested in determining which factors lead to successful teams. The table below lists the team batting average and the team winning percentage for the 14 league teams at the end of a recent season.

Team-B-A	Winning%
0.254	0.414
0.269	0.519
0.255	0.500
0.262	0.537
0.254	0.352
0.247	0.519
0.264	0.506
0.271	0.512
0.280	0.586
0.256	0.438
0.248	0.519
0.255	0.512
0.270	0.525
0.257	0.562

y = winning % and x = team batting average

Solution:

a) LS Regression Line

$$\sum x_i = 3.642, \sum x_i^2 = 0.949$$

$$\sum y_i = 7.001, \sum y_i^2 = 3.549$$

$$\sum x_i y_i = 1.824562$$

$$SS_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 1.824562 - \frac{(3.642)(7.001)}{14} = 0.0033$$

$$SS_x = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 0.948622 - \frac{(3.642)^2}{14} = 0.00118$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_x} = \frac{0.003302}{0.001182} = 0.7941$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 0.5 - (0.7941)0.26 = 0.2935$$

The least squares regression line is

$$\hat{y} = 0.2935 + 0.7941x$$

The meaning $\hat{\beta}_1 = 0.7941$ is for each additional batting average of the team, the winning percentage increases by an average of 79.41%.

b) Standard Error of Estimate

$$SSE = S_{yy} - \left(\frac{S_{xy}^2}{S_{xx}}\right) = \left(\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}\right) - \left(\frac{S_{xy}^2}{S_{xx}}\right)$$

$$= (3.548785 - \frac{7.001^2}{14}) - \frac{0.003302^2}{0.00182} = 0.03856$$

So,

$$s_\varepsilon^2 = \frac{SSE}{n-2} = \frac{0.03856}{14-2} = 0.00321 \text{ and } s_\varepsilon = \sqrt{s_\varepsilon^2} = 0.0567$$

Since s_ε=0.0567 is small, we would conclude that "s" is relatively small, indicating that the regression line fits the data quite well.

c) Do the data provide sufficient evidence at the 5% significance level to conclude that higher team batting average lead to higher winning percentage?

$$H_0: \quad \beta_1 = 0$$
$$H_A: \quad \beta_1 > 0$$

$$\text{Test statistic: } t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} = 1.69 \qquad \text{(p-value=.058)}$$

Conclusion: Do not reject H_0 at $\alpha = 0.05$. The higher team batting average does not lead to higher winning percentage.

d)Coefficient of Determination

$$R^2 = \frac{SS_{xy}^2}{SS_x - SS_y} = 1 - \frac{SSE}{SS_y} = 1 - \frac{0.03856}{0.04778} = 0.1925$$

The 19.25% of the variation in the winning percentage can be explained by the batting average.

e) Predict with 90% confidence the winning percentage of a team whose batting average is 0.275.

$$\hat{y} = 0.2935 + 0.7941(0.275) = 0.5119$$

$$\hat{y} \pm t_{\alpha/2,n-2} s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_g - \bar{x})^2}{SS_x}} =$$

$$0.5119 \pm (1.782)(0.0567)\sqrt{1 + \frac{1}{14} + \frac{(0.275 - 0.2601)^2}{0.001182}}$$

90% PI for y: 0.5119 ± 0.1134

$$(0.3985, 0.6253)$$

The prediction is that the winning percentage of the team will fall between 39.85% and 62.53%.

IMP Note to Self

Thank you