Linear Regression

Finding the Line of Best Fit

- Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).
- When there is a single input variable (x), the method is referred to as simple linear regression. When there are multiple input variables, it is referred as multiple linear regression.

Application Areas

- Company sales or profit predictions.
- In business to evaluate trends and make estimates and predictions.
- Predicting housing price based on the area and prices of other houses.
- Stock Market predictions.
- Bus company cost function.
- Credit card industry to minimize the risk portfolio.
- Engine performance from the test data.

Linear Regression: Finding the Line of Best Fit

Year	1998	1999	2000	2001	2002	2003
Gross receipts	48.00	51.45	54.04	55.94	60.49	64.10

Source: 2006 Statistical Abstracts of the United States.

Example 1:

- Independent variable is t – number of years since 1900
- Dependent variable is G – gross receipts of movie industry, in billions of dollars

Linear Regression: Finding the Best Fit Line

Example 1

Year	1998	1999	2000	2001	2002	2003
Gross receipts	48.00	51.45	54.04	55.94	60.49	64.10

Source: 2006 Statistical Abstracts of the United States.

Scatter Plot

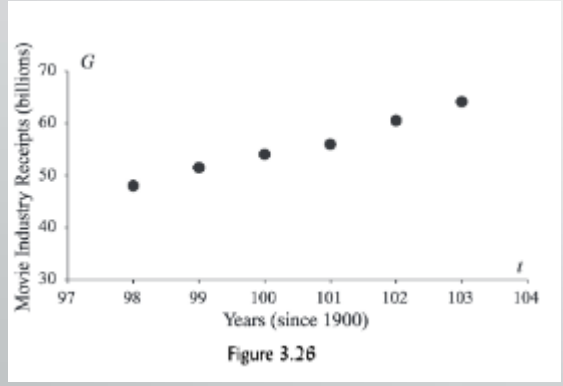


Figure 3.26

Possible Line Fits

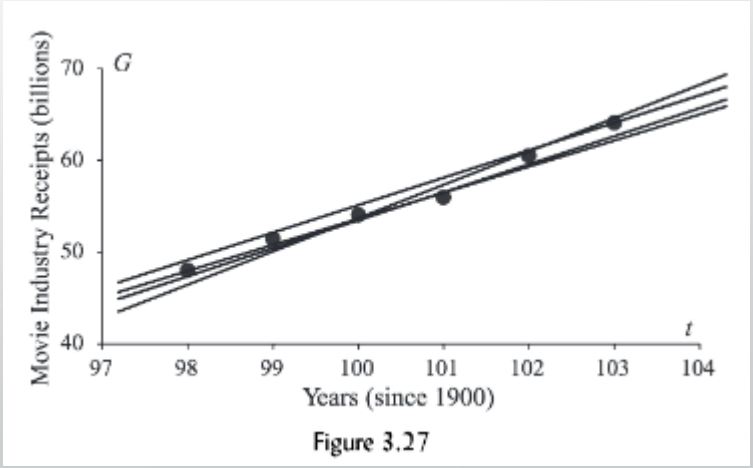


Figure 3.27

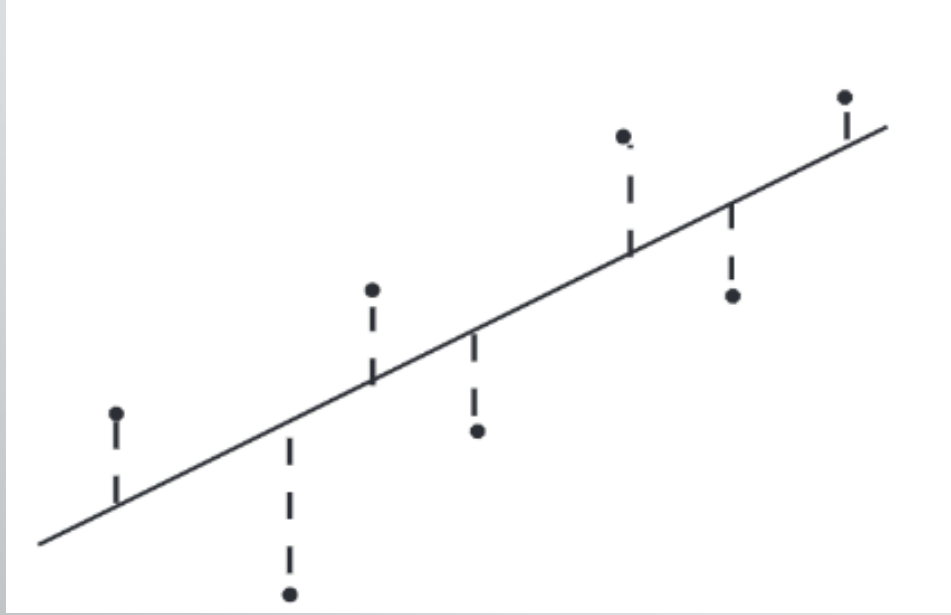
Linear Regression: Finding the Best Fit Line
Example 1

Year	1998	1999	2000	2001	2002	2003
Gross receipts	48.00	51.45	54.04	55.94	60.49	64.10

Source: 2006 Statistical Abstracts of the United States.

- How to determine the line that fits this data set in the best possible way?
- Line that passes as close as possible to ALL data points.
- Note: This line may not necessarily contain any of the points in the data set

The Least Squares Criterion



- The **least-squares criterion -** the line that best fits a set of data points is the one having the smallest possible sum of squared errors
- Note if we sum these errors some will be positive, others will be negative so they would cancel out – something to be avoided
- So we SQUARE all these differences(errors) before summing them up

Example 2 – Line of Best Fit

x	0	1	2	3	4	5	6	7	8
y	6	12	14	16	21	28	27	35	44

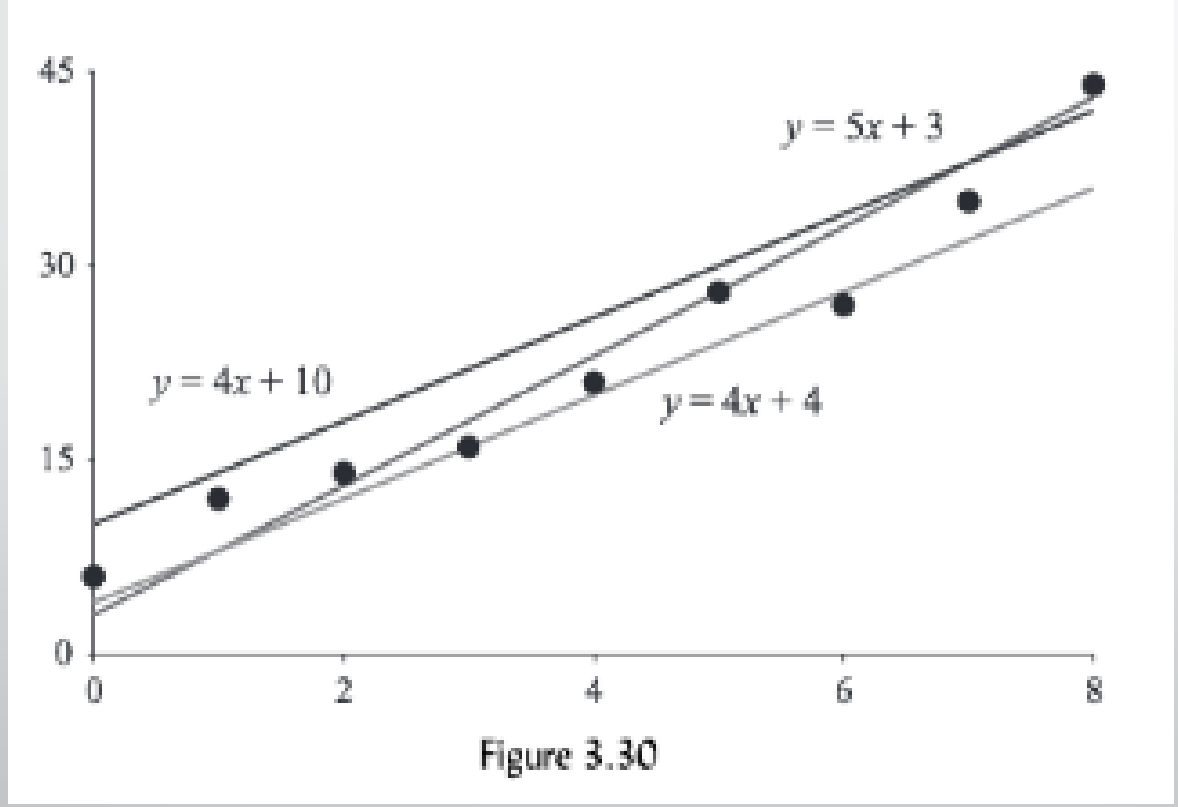


Figure 3.30

Example 2 – Line of Best Fit

- Which of the 3 lines captures the pattern in the data in the best possible way?
- Need to compute the sum of the squares

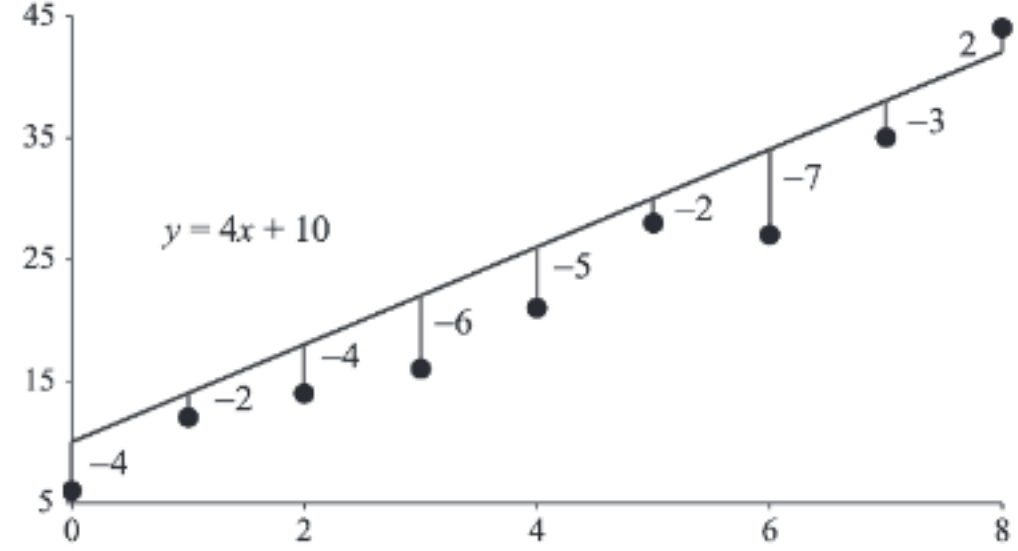


Figure 3.31

x	y	$y_L = 4x + 10$	$y - y_L$	$(y - y_L)^2$
0	6	10	−4	16
1	12	14	−2	4
2	14	18	−4	16
3	16	22	−6	36
4	21	26	−5	25
5	28	30	−2	4
6	27	34	−7	49
7	35	38	−3	9
8	44	42	2	4
				163

Example 2 – Line of Best Fit (continued)

$y = 4x + 4$					$y = 5x + 3$				
x	y	y_L	$y - y_L$	$(y - y_L)^2$	x	y	y_L	$y - y_L$	$(y - y_L)^2$
0	6	4	2	4	0	6	3	3	9
1	12	8	4	16	1	12	8	4	16
2	14	12	2	4	2	14	13	1	1
3	16	16	0	0	3	16	18	−2	4
4	21	20	1	1	4	21	23	−2	4
5	28	24	4	16	5	28	28	0	0
6	27	28	−1	1	6	27	33	−6	36
7	35	32	3	9	7	35	38	−3	9
8	44	36	8	64	8	44	43	1	1
				115					80

Correlation between two variables

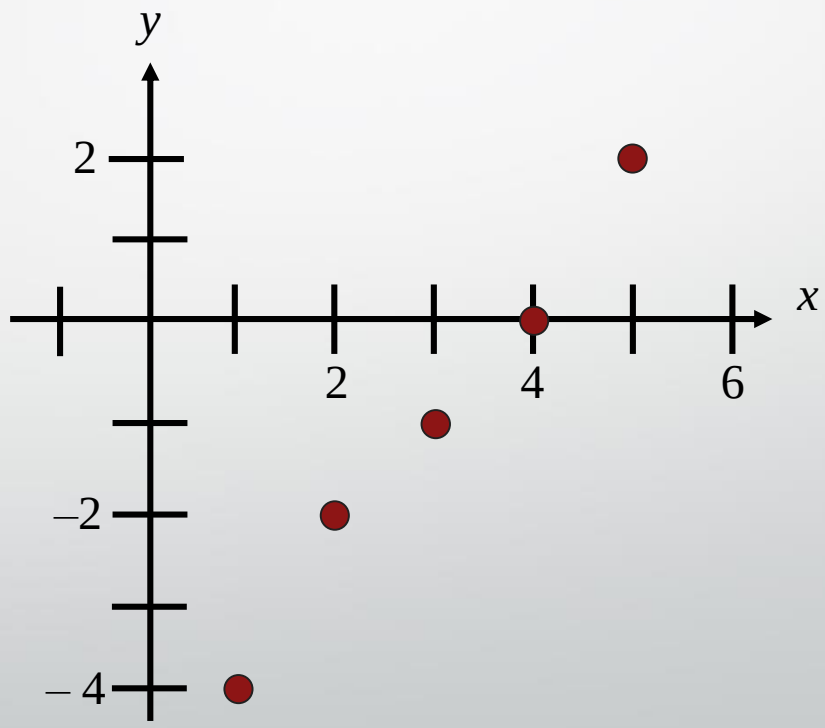
- A relationship between two variables.
- The data can be represented by ordered pairs (x, y)
  - x is the **independent variable**
  - y is the **dependent variable**

Correlation

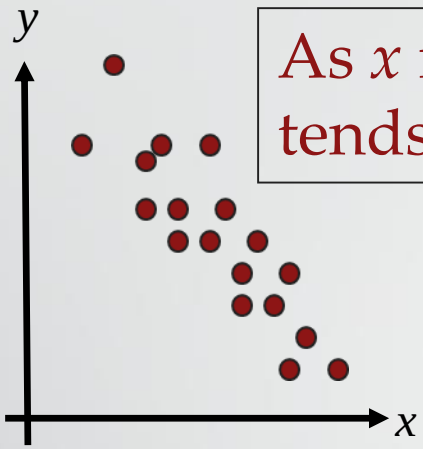
A **scatter plot** can be used to determine whether a linear (straight line) correlation exists between two variables.

Example:

x	1	2	3	4	5
y	– 4	– 2	– 1	0	2

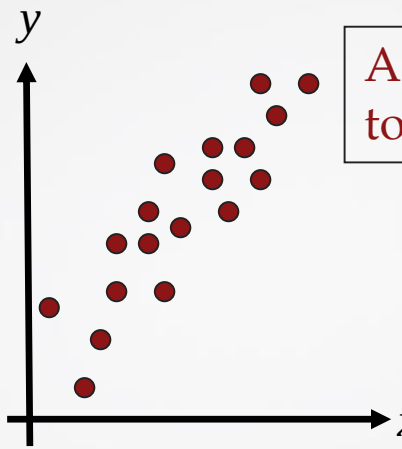


Types of Correlation



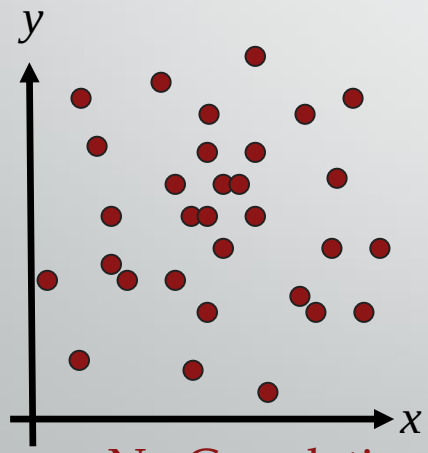
As x increases, y tends to decrease.

Negative Linear Correlation

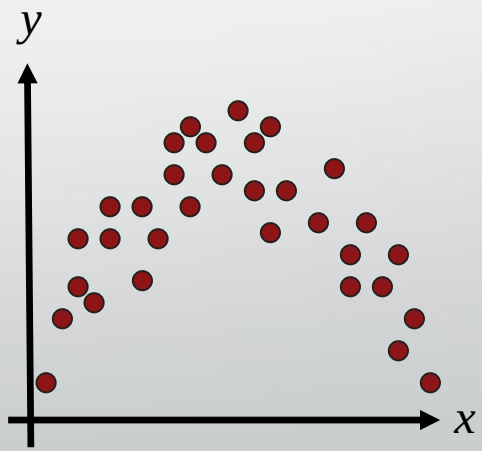


As x increases, y tends to increase.

Positive Linear Correlation



No Correlation



Nonlinear Correlation