

STP 226 ELEMENTARY STATISTICS

CHAPTER 4

DESCRIPTIVE MEASURES IN REGRESSION AND CORRELATION

Linear Regression and **correlation** allows us to examine the relationship between two or more quantitative variables.

4.1 Linear Equations with one Independent Variable

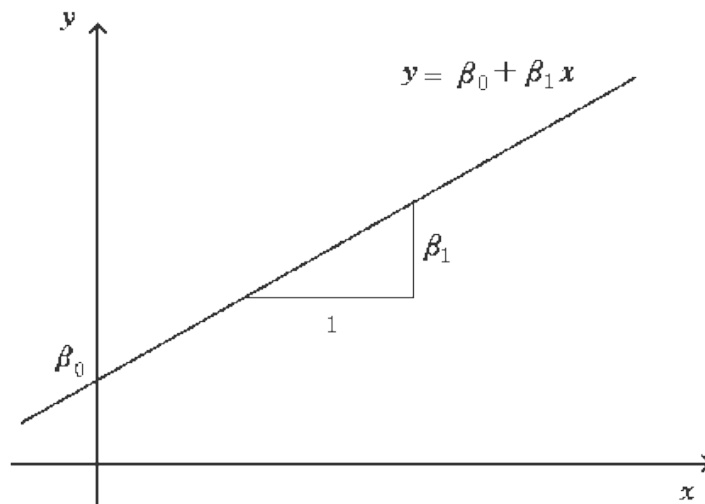
$y = b_0 + b_1x$ is a straight line where b_0 and b_1 are constants,

b_0 is the **y-intercept** and b_1 is the **slope** of the line.

Slope (b_1) = for every 1 unit horizontal increase (in x) there is a b_1 unit vertical increase/decrease (in y) depending on the line.

Slope = change in y / change in x .

For example if units of x are kg and units of y are \$ and $y = 5 + 2x$, slope = 2\$/kg, so per each increase in x by 1 kg, y increases by \$2.



The straight-line graph of the linear equation $y = b_0 + b_1x$ slopes upward if $b_1 > 0$, slopes downward if $b_1 < 0$, and is horizontal if $b_1 = 0$. Vertical line has no slope.

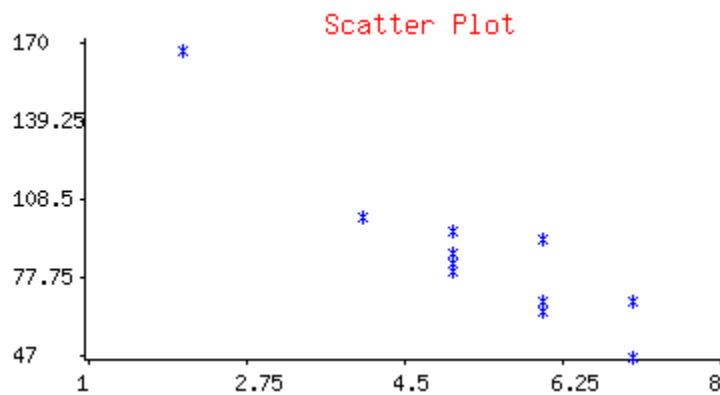
4.2 The Regression Equation

Often, in real life situations, it is not likely to have data that follow some straight line perfectly.

A **scatterplot (scatter diagram)** is useful in visualizing apparent relationships between two variables.

Example (Table 4.5) (Age and price of a Orion)

Car (Orion)	1	2	3	4	5	6	7	8	9	10	11
Age (yr): X	5	4	6	5	5	5	6	6	2	7	7
Price (\$100): Y	85	103	70	82	89	98	66	95	169	70	48

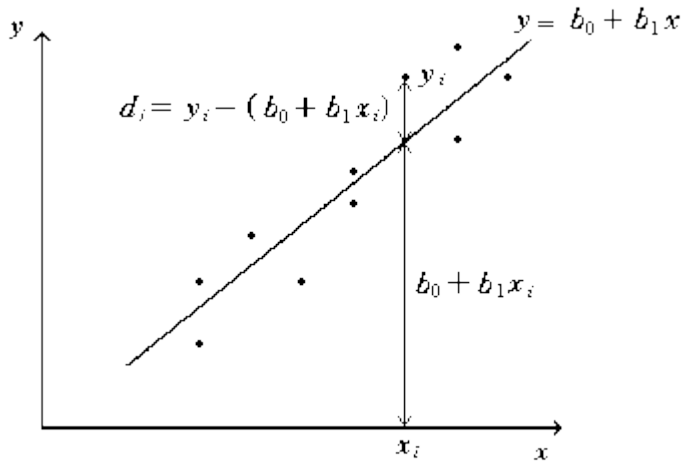


If the points seem to follow a straight line, then a straight line can be used to approximate the relationship. You observe a linear trend if your data points form an “elongated cloud “ shape. If there is no linear trend, you should not fit the linear equation to your data.

Many different lines can be drawn to approximate the relationship, however, the **least squares criterion** method gives the best line to fit the data (relationship between the two variables).

Least Squares Criterion – The straight line that best fits a set of data points is the one having the smallest possible sum of squared errors.

We define error as $e = y - \hat{y}$, where $\hat{y} = b_0 + b_1 x$ is the equation of the line. Error e is the signed vertical distance from the data point to the line $e < 0$ if the point is below the line, $e > 0$ if the point is above the line and $e = 0$ if the point is on the line.

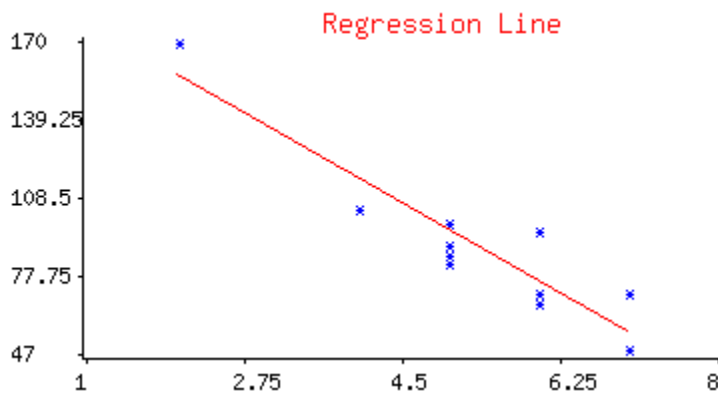


Find the difference (error) between every point and its corresponding point on the best fit line, square those errors, and sum them up. We want this sum to be minimum.

The picture above uses symbol d_i for the signed distances from the point to the line.

$$\min \sum d_i^2 = \min \sum (y_i - (b_0 + b_1 x_i))^2 = \min_{b_0, b_1} \sum e^2$$

Regression Line – The straight line that best fits a set of data points according to the **least squares criterion**.



Regression equation – The equation of the regression line:

$$\hat{y} = 195.47 - 20.6x$$

Notation used in Regression and Correlation

Definition:

$$S_{xx} = \sum (x - \bar{x})^2$$

$$S_{xy} = \sum (x - \bar{x})(y - \bar{y})$$

$$S_{yy} = \sum (y - \bar{y})^2$$

Computational Formula

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$

Formula 4.1 Regression Equation for a set of n data points is

$$\hat{y} = b_0 + b_1 x \quad , \quad \text{where}$$

$$b_1 = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad b_0 = \frac{1}{n} (\sum y - b_1 \sum x) = \bar{y} - b_1 \bar{x}$$

- **Predictor Variable and Response Variable**

For the linear **regression** equation, $y = b_0 + b_1 x$

y - **response variable** or dependent variable

x - **predictor variable/explanatory variable** or independent variable

Example (Orion Data): response variable=price, predictor variable=age

- **Extrapolation** - making predictions for values of the predictor variable outside the range of the observed values of the predictor variable.
- Grossly incorrect predictions can result from extrapolation.
- **Outlier** - a data point that lies far from the regression line, relative to other data points
- **Influential observation** - a data point whose removal causes the regression to change considerably. It is usually separated in the x-direction from the other data points. It pulls the regression line towards itself.

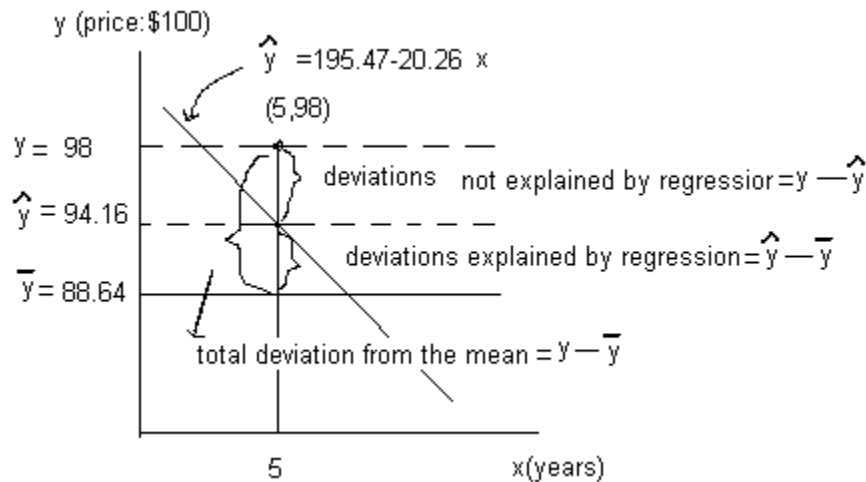
Warnings on the use of Linear Regression

- **Draw scatter diagram first**
- **Predict within the range of the data.**
- **Watch out for the influential observation**

4.3 The Coefficient of Determination (r^2)

- **One way of measuring the utility of regression equation**
determine the percentage of variation in the observed values of the response variable explained by the regression (or predictor variable).

Example (Orion Data)



X	Y	\hat{Y}	$y - \bar{y}$	$\hat{y} - \bar{y}$	$y - \hat{y}$
5	85	94.16	-3.64	5.53	-9.16
4	103	114.42	14.36	25.79	-11.42
6	70	73.90	-18.64	-14.74	-3.90
5	82	94.16	-6.64	5.53	-12.16
5	89	94.16	0.36	5.53	-5.16
5	98	94.16	9.36	5.53	3.84
6	66	73.90	-22.64	-14.74	-7.90
6	95	73.90	6.36	-14.74	21.10
2	169	154.95	80.36	66.31	14.05
7	70	53.64	-18.64	-35.00	16.36
7	48	53.64	-40.64	-35.00	-5.64

Coefficient of determination, r^2 : is the proportion of variation in the observed values of the response variable that is explained by the regression.

$$r^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

Eg. (contd.) $r^2 = 8285.0/9708.5 = 0.853$ (85.3%)

4.4 Linear Correlation

Linear Correlation Coefficient, r (Pearson product moment correlation coefficient):

A statistic used to measure the strength of linear relationship between two variables.

DEFINITION 4.6 The linear correlation coefficient, r , of n data points is defined by

$$r = \frac{1}{n-1} \frac{\sum (x-\bar{x})(y-\bar{y})}{S_x S_y}, \text{ or } r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

Eg. (contd.) $\sum x=58, \sum y=975, \sum xy=4732, \sum x^2=326, \sum y^2=96129$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{\sum xy - (\sum x)(\sum y)/n}{\sqrt{[\sum x^2 - (\sum x)^2/n][\sum y^2 - (\sum y)^2/n]}} = -0.924$$

strong negative linear relationship between the age and price of Orions.

Understanding the Linear Correlation Coefficient

- r reflects the slope of the scatter diagram
- the magnitude of r indicates the strength of the **linear** relationship
- The sign of r suggests the type of linear relationship
 $r=0$ means no linear relation, $r>0$ means positive relation, $r<0$ mean negative relation between the two variables.
- r has no units
- The sign of r is identical to the sign of the slope of the regression line.

Note: coefficient of determination, r^2 is the square of the linear correlation coefficient.

Eg. (contd.) $(-0.924)^2 = 0.854$

Warnings on the use of linear correlation coefficient.

- measures only linear relation between the variables
- high correlation does not have to imply causal relationship between x and y
- watch out for the spurious correlation (lurking variables), x and y may be associated with the third variable that produces changes in both.
- correlation is affected by the extreme observations (same as mean and standard deviation)
- Watch out for separate groups