Chap. 3 Notes: Linear Relationships: Regression and Correlation

J. Harner   A. Billings

Department of Statistics
West Virginia University

Stat 211 Fall 2007

Outline

Introduction:

Bivariate data:

Bivariate data consists of data for two variables for each individual in the sample.

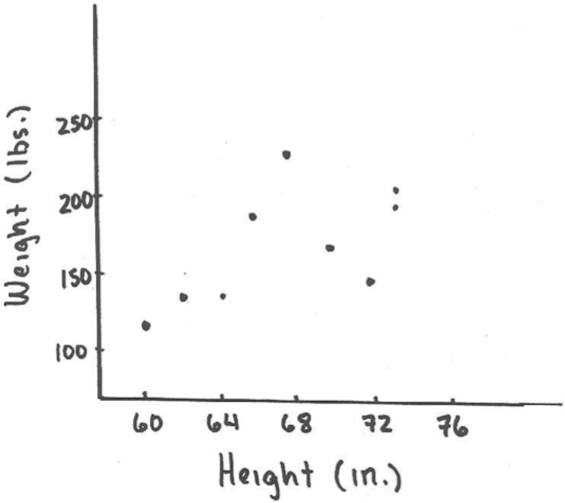
Example: height and weight.

Example: gender and income.

Scatter Plots:

- A scatter plot is a graph of all ordered pairs of numeric bivariate data on a coordinate axis system.
- Each plotted point represents the value of two variables for a single individual.

Example: Students' heights and weights;



Pearson Correlation Coefficient : r:

Recall: Bivariate data consists of measurements on two variables on each individual.

Example: Height and weight..

Age and Income.

- The **Pearson Correlation Coefficient**, r , is computed as,

$$\mathrm{r} = \frac{SS(XY)}{\sqrt{SS(X) \, . \, SS(Y)}}$$

Where,

$$\begin{aligned} SS(XY) &= \sum XY - \frac{(\sum X).(\sum Y)}{n} \\ SS(X) &= \sum X^2 - \frac{(\sum X)^2}{n} \\ SS(Y) &= \sum Y^2 - \frac{(\sum Y)^2}{n} \end{aligned}$$

- The "SS" is called "Sum of Square".

Example: Data from 8 randomly selected individuals was collected on "number of hours of TV per day" (X) and "Cholesterol level" (Y).

Compute the Pearson Correlation Coefficient r.

Solution:

X	Y	X^2	Y^2	XY
3.5	215	12.25	46225	752.5
1.5	180	2.25	32400	270.0
2.0	205	4.00	42025	40.0
1.0	175	1.00	30625	175.0
2.0	190	4.00	36100	380.0
2.5	200	6.25	40000	500.0
305	212	9.00	44944	636.0
3.0	220	9.00	48400	660.0
18.5	1597	47.75	320719	3783.5
$\sum X$	$\sum Y$	$\sum X^2$	$\sum Y^2$	$\sum XY$

- Now, compute sums of squares:

$$
\begin{aligned}
SS(XY) &= \sum XY - \frac{(\sum X).(\sum Y)}{n} \\
&= 3783.5 - \frac{(18.5).(1597)}{8} \\
&= 3783.5 - 3693.06 \\
&= 90.44 \\
SS(X) &= \sum X^2 - \frac{(\sum Y^2)}{n} \\
&= 47.75 - \frac{18.5^2}{8} \\
&= 47.75 - 42.78 \\
&= 4.97
\end{aligned}
$$

$$
\begin{aligned}
SS(Y) &= \sum Y^2 - \frac{(\sum Y)^2}{n} \\
&= 320719 - \frac{1597^2}{8} \\
&= 320719 - 318801.12 \\
&= 1917.88
\end{aligned}
$$

- Now, plug into formula for r,

$$
\begin{aligned}
\mathrm{r} &= \frac{SS(XY)}{\sqrt{SS(X).SS(Y)}} \\
&= \frac{90.44}{\sqrt{(4.97).(1917.88)}} \\
&= \frac{90.44}{97.63} \\
&= 0.926
\end{aligned}
$$

Interpretation:

I. **IF** we conclude that X and X are correlated (hyp. test), it does not necessarily imply a "Causal relationship".

   a. A third variable, W , may affect both X and Y.
   b. Coincidence.

II. r measures the strength of a linear (straight line) relationship.

III. Outliers may disproportionately affect the value of r.

IV. $-1 \leq \mathrm{r} \leq 1$

V. $r \simeq 0$ means that X and Y are uncorrelated.

VI. $r > 0$ (positive correlation)
As X increases (decreases) the corresponding value of Y tends to increase (decrease).

VII. $r < 0$ (negative correlation)
As X increases (decreases) the corresponding value of the Y tends to decrease (increase).

Scatter diagram and Correlation:

strong nonlinear relationship, linear correlation

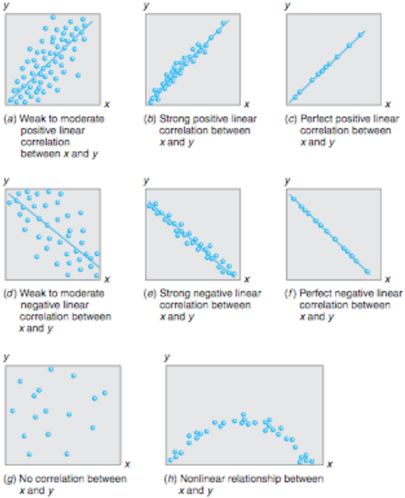
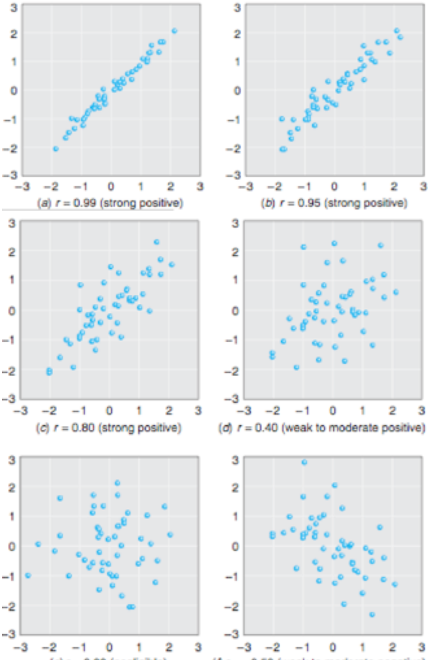


Figure 3.4 Scatterplots of types of linear relationships.



(a) $r = 0.99$ (strong positive)

(b) $r = 0.95$ (strong positive)

(c) $r = 0.80$ (strong positive)

(d) $r = 0.40$ (weak to moderate positive)

(e) $r = 0.00$ (negligible)

(f) $r = -0.50$ (weak to moderate negative)

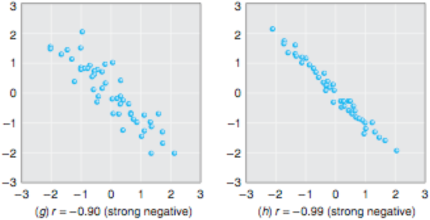


Figure 3.6 Scatterplots illustrating different levels of correlation. In each plot both variables have mean 0 and standard deviation 1.

Regression Analysis - Introduction

- If we determine that two variables X and Y are correlated, we can use the value of one variable, X , to "predict" the corresponding value of the other variable, Y .

Example: Use height (X) to predict weight (Y)

Use education level (X) to predict income (Y).

- We use a "regression equation" to make prediction,

$$y = b_0 + b_1 X$$

where,

$$\begin{aligned} b_0 &= y - intercept \\ b_1 &= slope(regressioncoefficient) \end{aligned}$$

- Book uses:[1]

$$y = mx + b$$

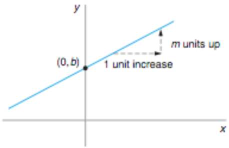


Figure 3.15 The line $y = mx + b$.

[1] refer to page 157 of the Text Book

Computing a Regression Equation:

- Given a set of bivariate data (X , Y), we can compute b_0 and b_1 as,

$$\begin{aligned} b_1 &= \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} \\ &= \frac{SS(XY)}{SS(X)} \end{aligned}$$

and,

$$b_0 = \bar{Y} - b_1 \bar{X}$$

Example: (Contd. from Sect 3.2 notes)

$$\begin{aligned} X &= \text{hours of TV per day} \\ Y &= \text{cholesterol level} \end{aligned}$$

From Sect. 3.2, we computed

$$\begin{aligned} SS(XY) &= \sum XY - \frac{(\sum X)(\sum Y)}{n} \\ &= 90.4 \\ SS(X) &= \sum X^2 - \frac{(\sum X)^2}{n} \\ &= 4.97 \end{aligned}$$

So,

$$\begin{aligned} b_1 &= \frac{SS(XY)}{SS(X)} \\ &= \frac{90.44}{4.97} \\ &= 18.2 \end{aligned}$$

Next,

$$\begin{aligned} \bar{X} &= \frac{\sum X}{n} = \frac{18.5}{8} \\ &= 2.31 \\ \bar{Y} &= \frac{\sum Y}{n} = \frac{1597}{8} \\ &= 199.63 \end{aligned}$$

So,

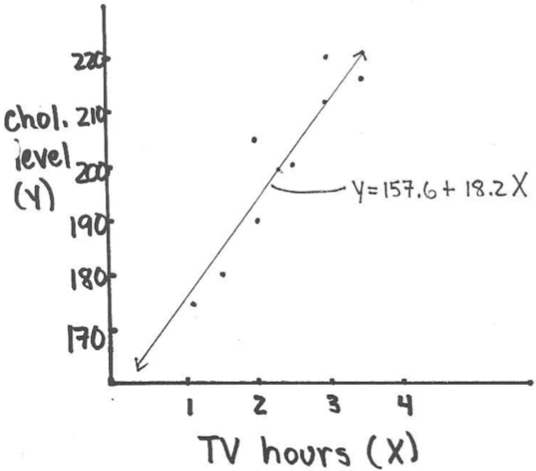
$$\begin{aligned} b_0 &= \bar{Y} - b_1\bar{X} \\ &= 199.63 - (18.2)(2.31) \\ &= 199.63 - 42.04 \\ &= 157.59 \end{aligned}$$

- So our regression equation is,

$$\begin{aligned} y &= b_0 + b_1 X \\ y &= 157.6 + 18.2X \end{aligned}$$

Plotting the regression Line on the scatter Plot:

Example: (Contd.)



- Notes:

  1. Regression Line is the "line of best fit" in the the sense that this line is the line which is closest to all data points simultaneously.
  2. The regression line $y = b_0 + b_1 X$ passes through $(\bar{X}, \bar{Y})$.
  3. Don't predict y - values for X's outside of the range of the data.

Predicting values of y:

Example: (Contd.)
Predict the cholesterol level (Y) for a person who watches 2 hours TV a day (X).

- For X = 2, the predicted Y is,

$$
\begin{aligned}
\widehat{Y} &= 157.6 + 18.2X \\
\widehat{Y} &= 157.6 + 18.2(2) \\
&= 194.0
\end{aligned}
$$

Interpretation:

- The predicted y - value is the mean cholesterol level of all persons who watch 2 hours TV a day.

Residuals:

- For each y - value, we can compute a *residuals* - the difference between the observed y - value and the predicted y - value.

$$\text{residual} = y - \hat{y}$$

Example: (Contd.)
For X = 2.2, the observed y = 200 and,

$$\begin{aligned} \hat{y} &= 157.6 + 18.2X \\ &= 157.6 + 18.2(2.5) \\ &= 203.1 \end{aligned}$$

- Thus, the residual for this observation is,

$$\begin{aligned} \textit{residual} &= y - \hat{y} \\ &= 200 - 203.1 \\ &= -3.1 \end{aligned}$$

- Note:

A residual may be positive or negative (or zero)

Interpretation:

- the observed y - value for X = 2.5 is 3.1 units less than the predicted y - value for X = 2.5.

Causation

- Read Sect. 3.4 of the Text Book[2]

[2] refer to page 187 of the Text Book