STA 218: Statistics for Management

Al Nosedal.
University of Toronto.

Fall 2018

My momma always said: "Life was like a box of chocolates. You never know what you're gonna get."

Forrest Gump.

Problem

How much do people with a bachelor's degree (but no higher degree) earn? Here are the incomes of 15 such people, chosen at random by the Census Bureau in March 2002 and asked how much they earned in 2001. Most people reported their incomes to the nearest thousand dollars, so we have rounded their responses to thousands of dollars: 110 25 50 50 55 30 35 30 4 32 50 30 32 74 60.

How could we find the "typical" income for people with a bachelor's degree (but no higher degree)?

Measuring center: the mean

The most common measure of center is the ordinary **arithmetic average, or mean**. To find the mean of a set of observations, add their values and divide by the number of observations. If the n observations are $x_1, x_2, ..., x_n$, their mean is

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$

or in more compact notation,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Income Problem

$\bar{x} = \frac{110+25+50+50+55+30+...+32+74+60}{15} = 44.466$

Do you think that this number represents the "typical" income for people with a bachelor's degree (but no higher degree)?

Measuring center: the median

The **median** M is the **midpoint** of a distribution, the number such that half the observations are smaller and the other half are larger.
To find the median of the distribution:
Arrange all observations in order of size, from smallest to largest.
If the number of observations n is odd, the median M is the center observation in the ordered list. Find the location of the median by counting $\frac{n+1}{2}$ observations up from the bottom of the list.
If the number of observations n is even, the median M is the mean of the two center observations in the ordered list. Find the location of the median by counting $\frac{n+1}{2}$ observations up from the bottom of the list.

Income Problem (Median)

We know that if we want to find the median, M, we have to order our observations from smallest to largest: 4 25 30 30 30 32 32 35 50 50 50 55 60 74 110. Let?s find the location of M
location of $M = \frac{n+1}{2} = \frac{15+1}{2} = 8$
Therefore, $M = x_8 = 35$ (x_8= 8th observation on our ordered list).

Measuring center: Mode

Another measure of location is the **mode**. The mode is defined as follows. The mode is the **value that occurs with greatest frequency**. Note: situations can arise for which the greatest frequency occurs at two or more different values. In these instances more than one mode exists.

Income Problem (Mode)

Using the definition of mode, we have that:

$\text{mode}_1 = 30$
and
$\text{mode}_2 = 50$

Note that both of them have the greatest frequency, 3.

Example: New York travel times.

Here are the travel times in minutes of 20 randomly chosen New York workers:
10 30 5 25 40 20 10 15 30 20 15 20 85 15 65 15 60 60 40 45.
Compare the mean and median for these data. What general fact does your comparison illustrate?

Solution

Mean:

$\bar{x} = \frac{10+30+5+...+60+60+40+45}{20} = 31.25$

Median:

First, we order our data from smallest to largest

5 10 10 15 15 15 15 20 20 20 25 30 30 40 40 45 60 60 65 85 .

location of $M = \frac{n+1}{2} = \frac{20+1}{2} = 10.5$

Which means that we have to find the mean of x_{10} and x_{11}.

$M = \frac{x_{10}+x_{11}}{2} = \frac{20+25}{2} = 22.5$

Comparing the mean and the median

The mean and median of a symmetric distribution are close together. In a skewed distribution, the mean is farther out in the long tail than is the median. Because the mean cannot resist the influence of extreme observations, we say that it is not a resistant measure of center.

Measures of Central Location for Categorical Data

For ordinal and nominal data, the calculation of the **mean** is **not valid**. Because the calculation of the median begins by placing the data in order, this statistic is appropriate for ordinal data. The mode, which is determined by counting the frequency of each observation, is appropriate for nominal data. However, nominal data do not have a "center", so we cannot interpret the mode of nominal data in that way. It is generally **pointless** to compute the **mode** of nominal data.

Geometric Mean

Suppose you make a 2-year investment of \$1,000, and it grows by 100% to \$2,000 during the first year. During the second year, however, the investment suffers a 50% loss, from \$2,000 back to \$1,000. The rates of return for years 1 and 2 are $R_1 = 100\%$ and $R_2 = -50\%$, respectively. The arithmetic mean is computed as

$$\bar{R} = \frac{R_1 + R_2}{2} = \frac{100 + (-50)}{2} = 25\%$$

Geometric mean

But this figure is misleading. Because there was no change in the value of the investment from the beginning to the end of the 2-year period, the "average" compounded rate of return is 0%. As you will see, this is the value of the **geometric mean**.
Let R_i denote the rate of return (in decimal form) in period i ($i = 1, 2, ..., n$). The **geometric mean** R_g of the returns $R_1, R_2, ..., R_n$ is defined such that

$$(1 + R_g)^n = (1 + R_1)(1 + R_2) \dots (1 + R_n)$$

Geometric mean

Solving for R_g, we produce the following formula:

$$R_g = [(1 + R_1)(1 + R_2)...(1 + R_n)]^{1/n} - 1$$

The geometric mean of our investment illustration is

$$R_g = [(1+R_1)(1+R_2)]^{1/2} - 1 = \sqrt{(1 + 1)(1 + [-0.50])} - 1 = 1 - 1 = 0$$

The geometric mean is therefore 0%.

Exercise 4.10

An investment of \$1,000 you made 4 years ago was worth \$1,200 after the first year, \$1,200 after the second year, \$1,500 after the third year, and \$2,000 today.

a. Compute the annual rates of return.
b. Compute the mean and median of the rates of return.
c. Compute the geometric mean.

Solution a)

$$1000(1 + R_1) = 1200$$
$$1 + R_1 = \frac{1200}{1000}$$
$$R_1 = 1.20 - 1 = 0.20$$

Solution a)

$$1200(1 + R_2) = 1200$$
$$1 + R_2 = \frac{1200}{1200}$$
$$R_2 = 1 - 1 = 0$$

Solution a)

$$1200(1 + R_3) = 1500$$
$$1 + R_3 = \frac{1500}{1200}$$
$$R_3 = 1.25 - 1 = 0.25$$

Solution a)

$$1500(1 + R_4) = 2000$$
$$1 + R_4 = \frac{2000}{1500}$$
$$R_4 \approx 1.33 - 1 = 0.33$$

Solution b)

Mean.

$$\bar{R} = \frac{R_1+R_2+R_3+R_4}{4}$$
$$\bar{R} = \frac{0.20+0+0.25+0.33}{4}$$
$$\bar{R} = \frac{0.78}{4} = 0.195$$

Solution b)

Median. First, we order our observations from smallest to largest:
0, 0.20, 0.25, 0.33.
location of $M = \frac{n+1}{2} = \frac{4+1}{2} = \frac{5}{2} = 2.5$
(which implies that M will be the average between the second and third observations in our **ordered list**).
$M = \frac{R_{(2)}+R_{(3)}}{2} = \frac{0.20+0.25}{2} = 0.225.$

Solution c)

Geometric mean.

$$R_g = [(1 + R_1)(1 + R_2)...(1 + R_n)]^{1/n} - 1$$
$$R_g = [(1 + 0.20)(1 + 0)(1 + 0.25)(1 + 0.33)]^{1/4} - 1$$
$$R_g = [(1.20)(1)(1.25)(1.33)]^{1/4} - 1$$
$$R_g = [1.995]^{1/4} - 1$$
$$R_g = 1.188 - 1$$
$$R_g = 0.188$$

Measures of Variability: Range

The simplest measure of variability is the range.
Range= Largest value - smallest value
Range= MAX - min

Measures of Variability: Variance

The variance s^2 of a set of observations is an average of the squares of the deviations of the observations from their mean. In symbols, the variance of n observations $x_1, x_2, ..., x_n$ is

$$s^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+...+(x_n-\bar{x})^2}{n-1}$$

or, more compactly,

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

Measures of Variability: Standard Deviation

The standard deviation s is the square root of the variance s^2:

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

Example

Consider a sample with data values of 10, 20, 12, 17, and 16. Compute the variance and standard deviation.

Solution

First, we have to calculate the mean, $\bar{x}$:

$\bar{x} = \frac{10+20+12+17+16}{5} = 15.$

Now, let's find the variance s^2:

$s^2 = \frac{(10-15)^2+(20-15)^2+(12-15)^2+(17-15)^2+(16-15)^2}{5-1}.$

$s^2 = \frac{64}{4} = 16.$

Finally, let's find the standard deviation s:

$s = \sqrt{16} = 4.$

Empirical Rule

Knowing the mean and standard deviation allows the statistics practitioner to extract useful bits of information. If the histogram is bell shaped, we can use the **Empirical Rule**.
Approximately 68% of the observations fall within one standard deviation of the mean.
Approximately 95% of the observations fall within two standard deviations of the mean.
Approximately 99.7% of the observations fall within three standard deviations of the mean.

Chebysheff's Theorem

The proportion of observations in any sample or population that lie within k standard deviations of the mean is at least

$$1 - \frac{1}{k^2} \qquad \text{for } |k| > 1.$$

Example

The results of a national survey showed that on average, adults sleep 6.9 hours per night. Suppose that the standard deviation is 1.2 hours.

a. Use Chebysheff's theorem to calculate the percentage of individuals who sleep between 4.5 and 9.3 hours.

b. Assume that the number of hours of sleep follows a bell-shaped distribution. Use the empirical rule to calculate the percentage of individuals who sleep between 4.5 and 9.3 hours per day. How does this result compare to the value that you obtained using Chebysheff's theorem in part (a)?

Solution

a) At least 75% of adults in our sample sleep between 4.5 hours [the mean minus two standard deviations $= 6.9 - 2(1.2) = 4.5$] and 9.3 hours per night [the mean plus two standard deviations $= 6.9 + 2(1.2) = 9.3$].
b) Using the empirical rule, approximately 95% of adults sleep between 4.5 and 9.3 hours per day.

The quartiles Q_1 and Q_3

To calculate the quartiles:
Arrange the observations in increasing order and locate the median M in the ordered list of observations.
The first quartile Q_1 is the median of the observations whose position in the ordered list is to the left of the location of the overall median.
The third quartile Q_3 is the median of the observations whose position in the ordered list is to the right of the location of the overall median.

Income Problem (Q_1)

Data:
4 25 30 30 30 32 32 35 50 50 50 55 60 74 110.
From previous work, we know that $M = x_8 = 35$.
This implies that the first half of our data has $n_1 = 7$ observations.
Let us find the location of Q_1:
location of $Q_1 = \frac{n_1+1}{2} = \frac{7+1}{2} = 4$.
This means that $Q_1 = x_4 = 30$.

Income Problem (Q_3)

Data:
4 25 30 30 30 32 32 35 50 50 50 55 60 74 110.
From previous work, we know that $M = x_8 = 35$.
This implies that the first half of our data has $n_2 = 7$ observations.
Let us find the location of Q_3:
location of $Q_3 = \frac{n_2+1}{2} = \frac{7+1}{2} = 4$.
This means that $Q_3 = 55$.

Five-number summary

The five-number summary of a distribution consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five-number summary is
$min\ Q_1\ M\ Q_3\ MAX$.

Income Problem (five-number summary)

Data: 4 25 30 30 30 32 32 35 50 50 50 55 60 74 110. The five-number summary for our income problem is given by:
4 30 35 55 110

R Code

```
# Step 1. Entering Data;

income=c(4,25,30,30,30,32,32,35,50,50,50,55,60,74,110);
```

R Code

```
# Step 2. Finding five-number summary;

fivenum(income);
```

R Code

```
## [1]   4.0  30.0  35.0  52.5 110.0
```

Note. Sometimes, R will give you a slightly different five-number summary.

R Code (just for fun)

```
# Ordering our list of observations
# from largest to smallest;

sort(income, decreasing=TRUE);
```

R Code (just for fun)

```
##   [1] 110   74   60   55   50   50   50   35   32   32   30   30   30
```

Box plot

A boxplot is a graph of the five-number summary.
A central box spans the quartiles Q_1 and Q_3.
A line in the box marks the median M.
Lines extended from the box out to the smallest and largest observations.

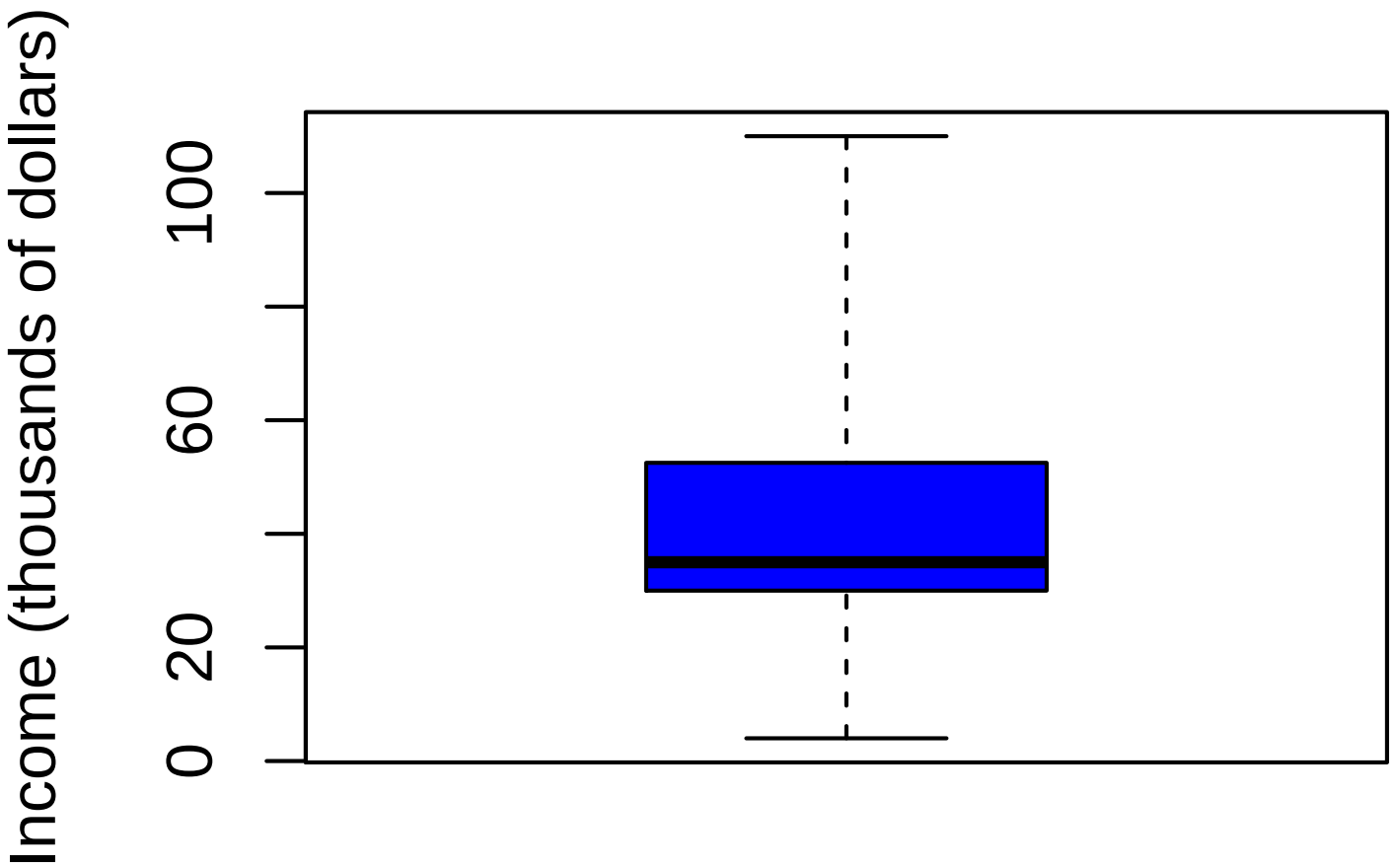
R code

```
# Step 1. Entering data;
income=c(4, 25, 30, 30, 30, 32, 32, 35,
         50, 50, 50, 55, 60, 74, 110);

# Step 2. Making boxplot

boxplot(income,col="blue",range=0,
        ylab="Income (thousands of dollars)")
# "regular" boxplot. It doesn't identify
# suspected outliers.
```

R code



Measures of Variability: IQR

A measure of variability that overcomes the dependency on extreme values is the interquartile range (IQR).
IQR = third quartile - first quartile
$\text{IQR} = Q_3 - Q_1$

1.5 IQR Rule

Identifying suspected outliers. Whether an observation is an outlier is a matter of judgement: does it appear to clearly stand apart from the rest of the distribution? When large volumes of data are scanned automatically, however, we need a rule to pick out suspected outliers. The most common rule is the 1.5 IQR rule. A point is a suspected outlier if it lies more than 1.5 IQR below the first quartile Q_1 or above the third quartile Q_3.

A high income.

In our income problem, we noted the influence of one high income of \$110,000 among the incomes of a sample of 15 college graduates. Does the 1.5 IQR rule identify this income as a suspected outlier?

Solution

Data: 4 25 30 30 30 32 32 35 50 50 50 55 60 74 110.

Q_1 and Q_3 are given by:

Q_1 =30 and Q_3 =55

$Q_3 + 1.5\ \text{IQR} = 55 + 1.5(25) = 92.5$

Since $110 > 92.5$ we conclude that 110 is an outlier.

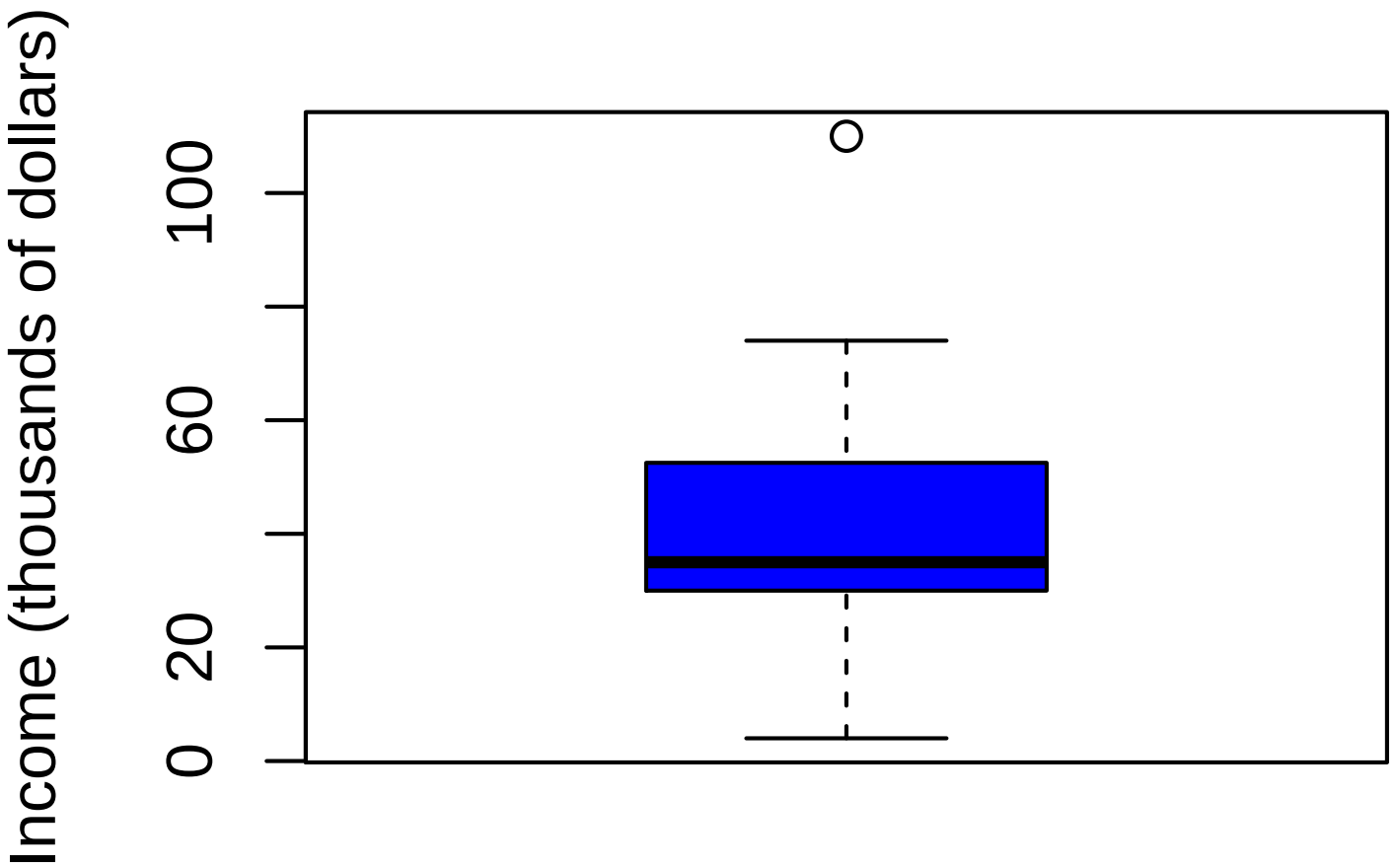
R code

```
# Step 1. Entering data;
income=c(4, 25, 30, 30, 30, 32, 32, 35,
         50, 50, 50, 55, 60, 74, 110);

# Step 2. Making boxplot

boxplot(income,col="blue",
        ylab="Income (thousands of dollars)")
# this version identifies
# suspected outliers.
```

R code



Problem

Ebby Halliday Realtors provide advertisements for distinctive properties and estates located throughout the United States. The prices listed for 22 distinctive properties and estates are shown here. Prices are in thousands.
1500 895 719 619 625 4450 2200
1280 700 619 725 739 799 2495
1395 2995 880 3100 1699 1120 1250 912.
a)Provide a five-number summary.
b)The highest priced property, $ 4,450,000, is listed as an estate overlooking White Rock Lake in Dallas, Texas. Should this property be considered an outlier?

Solution

a) min $= 619$
$Q_1 = 725$
M $= 1016$
$Q_3 = 1699$
MAX $= 4450$
b) IQR $= 1699$ - $725 = 974$
Q_3+1.5 IQR $= 1699 + 1.5\ (974) = 1699 + 1461 = 3160$.
Since $4450 > 3160$, we conclude that 4450 is an outlier.

Measures of Linear Relationship

Covariance (sample covariance)
You can compute the covariance, S_{XY} using the following formula:

$$S_{XY} = \frac{\sum_{i=1}^{n} x_i y_i}{n-1} - \frac{n\bar{x}\bar{y}}{n-1} \tag{1}$$

Measures of Linear Relationship

Coefficient of Correlation.

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

where

r_{xy} = sample correlation coefficient

S_{xy} = sample covariance

S_x = sample standard deviation of x

S_y = sample standard deviation of y.

Example

Five observations taken for two variables follow.

x_i	y_i
4	50
6	50
11	40
3	60
16	30

a. Develop a scatter diagram with x on the horizontal axis.
b. Compute the sample covariance.
c. Compute and interpret the sample correlation coefficient.

Solution

First, let's find $\bar{x}$ and $\bar{y}$

$$\bar{x} = \frac{4 + 6 + 11 + 3 + 16}{5} = 8$$

$$\bar{y} = \frac{50 + 50 + 40 + 60 + 30}{5} = 46$$

Solution (cont.)

Now, let's find s_x and s_y

$$s_x^2 = \frac{(4-8)^2 + (6-8)^2 + (11-8)^2 + (3-8)^2 + (16-8)^2}{4}$$

$$s_x^2 = \frac{(-4)^2 + (-2)^2 + (3)^2 + (-5)^2 + (8)^2}{4} = \frac{118}{4} = 29.5$$

$$s_x = 5.4313$$

Solution

$$s_y^2 = \frac{(50-46)^2 + (50-46)^2 + (40-46)^2 + (60-46)^2 + (30-46)^2}{4}$$

$$s_y^2 = \frac{(4)^2 + (4)^2 + (-6)^2 + (14)^2 + (-16)^2}{4} = \frac{520}{4} = 130$$

$$s_y = 11.4017$$

Solution

Finally, we find s_{xy} and r

$$\sum_{i=1}^{n} x_i y_i = (4)(50)+(6)(50)+(11)(40)+(3)(60)+(16)(30) = 1600$$

$$s_{xy} = \frac{\sum x_i y_i}{n-1} - \frac{n\bar{x}\bar{y}}{n-1} = \frac{1600}{4} - \frac{(5)(8)(46)}{4} = -60$$

$$r = \frac{s_{xy}}{s_x s_y} = \frac{-60}{(5.4313)(11.4017)} = -0.9688$$

R code

```
# Step 1. Entering data;

X=c(4,6,11,3,16);

Y=c(50,50,40,60,30);
```

R code

```
# Step 2. Finding means;

mean(X);

mean(Y);
```

R code

```
# Step 3. Finding variances;

var(X);

var(Y);
```

R code

```
# Step 4. Finding standard deviations;

sd(X);

sd(Y);
```

R code

```
# Step 5. Finding covariance and correlation;

cov(X,Y);

cor(X,Y);
```

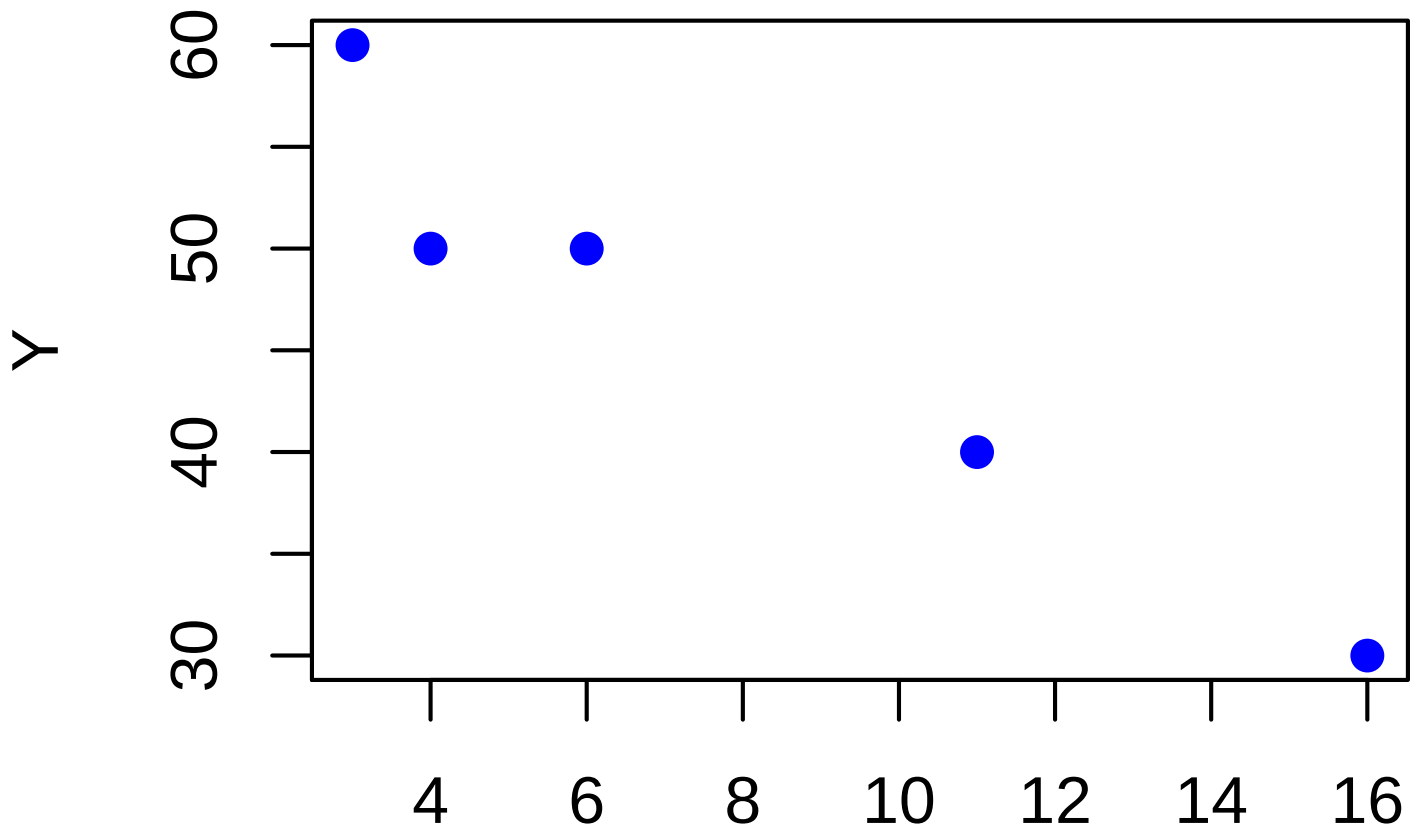
R code

```r
# Making scatterplot

plot(X,Y,pch=19,col="blue");

# pch=19 tells R to draw solid circles;
```

R code



Example

Five observations for two variables follow.

x_i	y_i
6	6
11	9
15	6
21	17
27	12

a. Develop a scatter diagram for these data.
b. Compute the sample covariance.
c. Compute and interpret the sample correlation coefficient.

Solution

First, let's find $\bar{x}$ and $\bar{y}$

$$\bar{x} = \frac{6 + 11 + 15 + 21 + 27}{5} = 16$$

$$\bar{y} = \frac{6 + 9 + 6 + 17 + 12}{5} = 10$$

Solution (cont.)

Now, let's find s_x and s_y

$$s_x^2 = \frac{(6-16)^2 + (11-16)^2 + (15-16)^2 + (21-16)^2 + (27-16)^2}{4}$$

$$s_x^2 = \frac{(-10)^2 + (-5)^2 + (-1)^2 + (5)^2 + (9)^2}{4} = 68$$

$$s_x = 8.2462$$

Solution (cont.)

$$s_y^2 = \frac{(6-10)^2 + (9-10)^2 + (6-10)^2 + (17-10)^2 + (22-10)^2}{4}$$

$$s_y^2 = \frac{(-4)^2 + (-1)^2 + (-4)^2 + (7)^2 + (12)^2}{4} = 21.5$$

$$s_y = 4.6368$$

Solution (cont.)

Finally, we find s_{xy} and r

$$\sum_{i=1}^{n} x_i y_i = (6)(6) + (11)(9) + (15)(6) + (21)(17) + (27)(12) = 906$$

$$s_{xy} = \frac{\sum x_i y_i}{n-1} - \frac{n\bar{x}\bar{y}}{n-1} = \frac{906}{4} - \frac{(5)(16)(10)}{4} = 26.5$$

$$r = \frac{s_{xy}}{s_x s_y} = \frac{26.5}{(8.2462)(4.6368)} = 0.6930$$

R code

```
# Step 1. Entering data;

X=c(6,11,15,21,27);

Y=c(6,9,6,17,12);
```

R code

```
# Step 2. Finding covariance and correlation;

cov(X,Y);

cor(X,Y);
```

R code (Results)

```
## [1] 26.5
## [1] 0.6930622
```

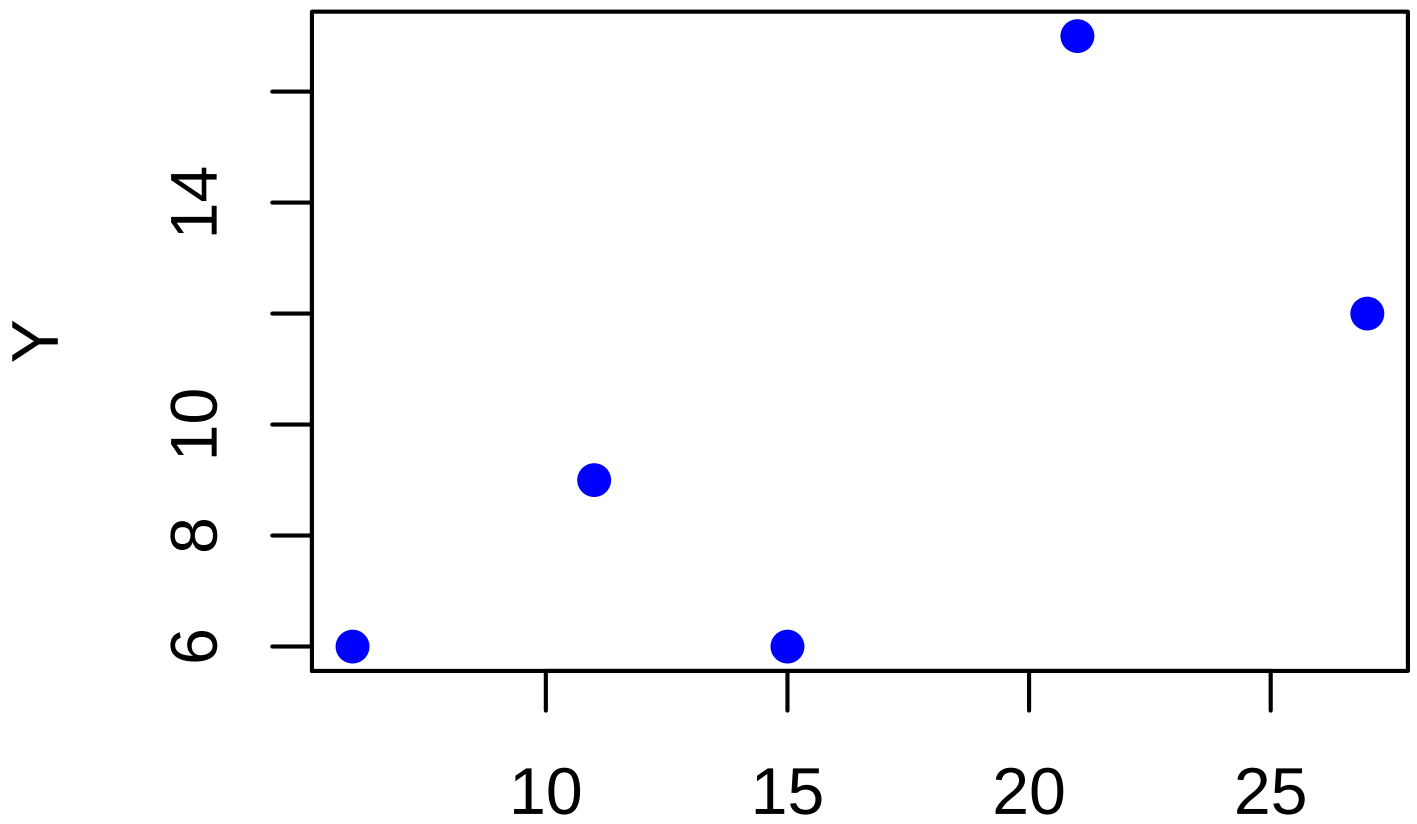
R code

```r
# Making scatterplot

plot(X,Y,pch=19,col="blue");

# pch=19 tells R to draw solid circles;
```

R code



Example

Calculate the coefficient of correlation for the following sets of data.
Set 1.

x_i	y_i
1	1
2	2
3	3
4	4
5	5

R code

```
# Step 1. Entering data;

X=c(1,2,3,4,5);

Y=c(1,2,3,4,5);
```

R code

```
# Step 2. Finding covariance and correlation;

cov(X,Y);

cor(X,Y);
```

R code (Results)

```
## [1] 2.5
## [1] 1
```

Example

Set 2.

x_i	y_i
-1	1
-2	2
-3	3
-4	4
-5	5

Example

Set 3.

x_i	y_i
-3	9
-2	4
-1	1
0	0
1	1
2	4
3	9

Facts about correlation

The correlation r measures the strength and direction of the linear association between two quantitative variables x and y. Although you calculate a correlation for any scatterplot, **r measures only straight-line relationships**.
Correlation indicates the direction of a linear relationship by its sign: $r > 0$ for a positive association and $r < 0$ for a negative association. Correlation always satisfies $-1 \leq r \leq 1$ and indicates the strength of a relationship by how close it is to -1 or 1. Perfect correlation, $r = \pm 1$, occurs only when the points on a scatterplot lie exactly on a straight line.

Least Squares Method

The least squares method produces a straight line drawn through the points so that the sum of squared deviations between the points and the line is minimized. The line is represented by the equation:

$$\hat{y} = b_0 + b_1 x$$

where b_0 is the y-intercept, and b_1 is the slope, and $\hat{y}$ (y hat) is the value of y determined by the line.

Least squares method

The coefficients b_0 and b_1 are derived using Calculus so that we minimize the sum of squared deviations: $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

Least Squares Line Coefficients

$$b_1 = r\frac{s_y}{s_x}$$

$$b_0 = \bar{y} - b_1\bar{x}.$$

Example

A tool die maker operates out of a small shop making specialized tools. He is considering increasing the size of his business and needs to know more about his costs. One such cost is electricity, which he needs to operate his machines and lights. He keeps track of his daily electricity costs and the number of tools that he made that day. These data are listed next. Determine the fixed and variable electricity costs using the Least Squares Method.

Day	Number of tools (X)	Electricity costs (Y)
1	7	23.80
2	3	11.89
3	2	15.89
4	5	26.11
5	8	31.79
6	11	39.93
7	5	12.27
8	15	40.06
9	3	21.38
10	6	18.65

R code

```
# Step 1. Entering Data;

tools=c(7,3,2,5,8,11,5,15,3,6);

cost=c(23.80,11.89,15.98,26.11,31.79,
39.93,12.27,40.06,21.38,18.65);
```

R code

```
# Step 2. Finding Slope;

Sx=sd(tools);

Sy=sd(cost);

r=cor(tools,cost);

b1=r*(Sy/Sx);

b1;

## [1] 2.245882
```

R code

```
# Step 3. Finding y-intercept;

x.bar=mean(tools);

y.bar=mean(cost);

b0=y.bar - b1*x.bar;

b0;

## [1] 9.587765
```

R code, another way

```
least.squares=lm(cost ~ tools);

least.squares;

##
## Call:
## lm(formula = cost ~ tools)
##
## Coefficients:
## (Intercept)        tools
##       9.588        2.246
```

R code (Graph)

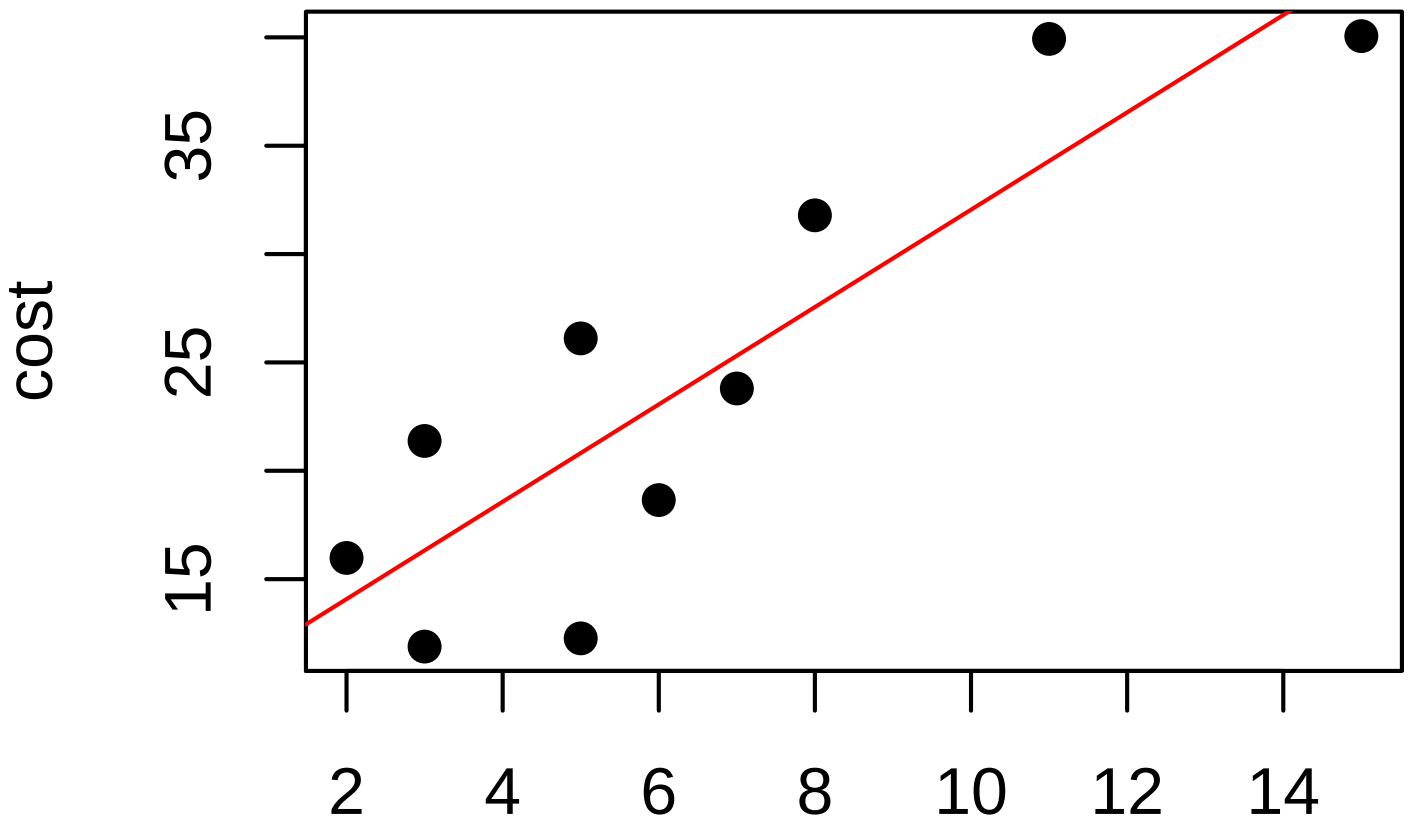
```
plot(tools,cost,pch=19);

abline(least.squares$coeff,col="red");

# pch=19 tells R to draw solid circles;

# abline tells R to add trendline;
```

Scatterplot



Interpretation

The slope measures the marginal rate of change in the dependent variable. In this example, the slope is 2.25, which means that in this sample, for each one-unit increase in the number of tools, the marginal increase in the electricity cost is \$2.25 per tool.

The y-intercept is 9.57; that is, the line strikes the y-axis at 9.57. However, when x=0, we are producing no tools and hence the estimated fixed cost of electricity is \$9.57 per day.

Interpretation

The coefficient of correlation is 0.8711, which tells us that there is a positive linear relationship between the number of tools and the electricity cost. The coefficient of correlation tells us that the linear relationship is quite strong and thus the estimates of the fixed and variable cost should be good.

The **coefficient of determination** measures the amount of variation in the dependent variable that is explained by the variation in the independent variable. In our example, the coefficient of correlation was calculated to be $r = 0.8711$. Thus, the coefficient of determination is $r^2 = (0.8711)^2 = 0.7588$. This tells us that 75.88% of the variation in electrical costs is explained by the number of tools. The remaining 24.12% is unexplained.

Facts about Least Squares Method

1. The distinction between explanatory and response variables is essential in Least Squares Method.
2. The least-squares line (trendline) always passes through the point $(\bar{x}, \bar{y})$ on the graph of y against x.
3. The square of the correlation, r^2, is the fraction of the variation in the values of y that is explained by the variation in x.

Example

The number of people living on American farms declined steadily during last century. Here are data on the farm population (millions of persons) from 1935 to 1980:

Year	Population
1935	32.11
1940	30.5
1945	24.4
1950	23.0
1955	19.1
1960	15.6
1965	12.4
1970	9.7
1975	8.9
1980	7.2

Example

a) Make a scatterplot of these data and find the least-squares regression line of farm population on year.
b) According to the regression line, how much did the farm population decline each year on the average during this period? What percent of the observed variation in farm population is accounted for by linear change over time?
c) Use the regression equation (trendline) to predict the number of people living on farms in 1990. Is this result reasonable? Why?

R code

```
# Step 1. Entering Data;

year=seq(1935,1980,by=5);

population=c(32.11,30.5,24.4,23.0,19.1,
15.6,12.4,9.7,8.9,7.2);

# seq creates a sequence of numbers;

# which starts at 1935 and ends at 1980;

# we want a distance of 5 between numbers;
```

R code, least squares

```
least.squares=lm(population ~ year);

least.squares;

cor(year,population);
```

R code, least squares (Results)

```
##
## Call:
## lm(formula = population ~ year)
##
## Coefficients:
## (Intercept)         year
##   1167.1418      -0.5869
## [1] -0.9884489
```

R code (Graph)

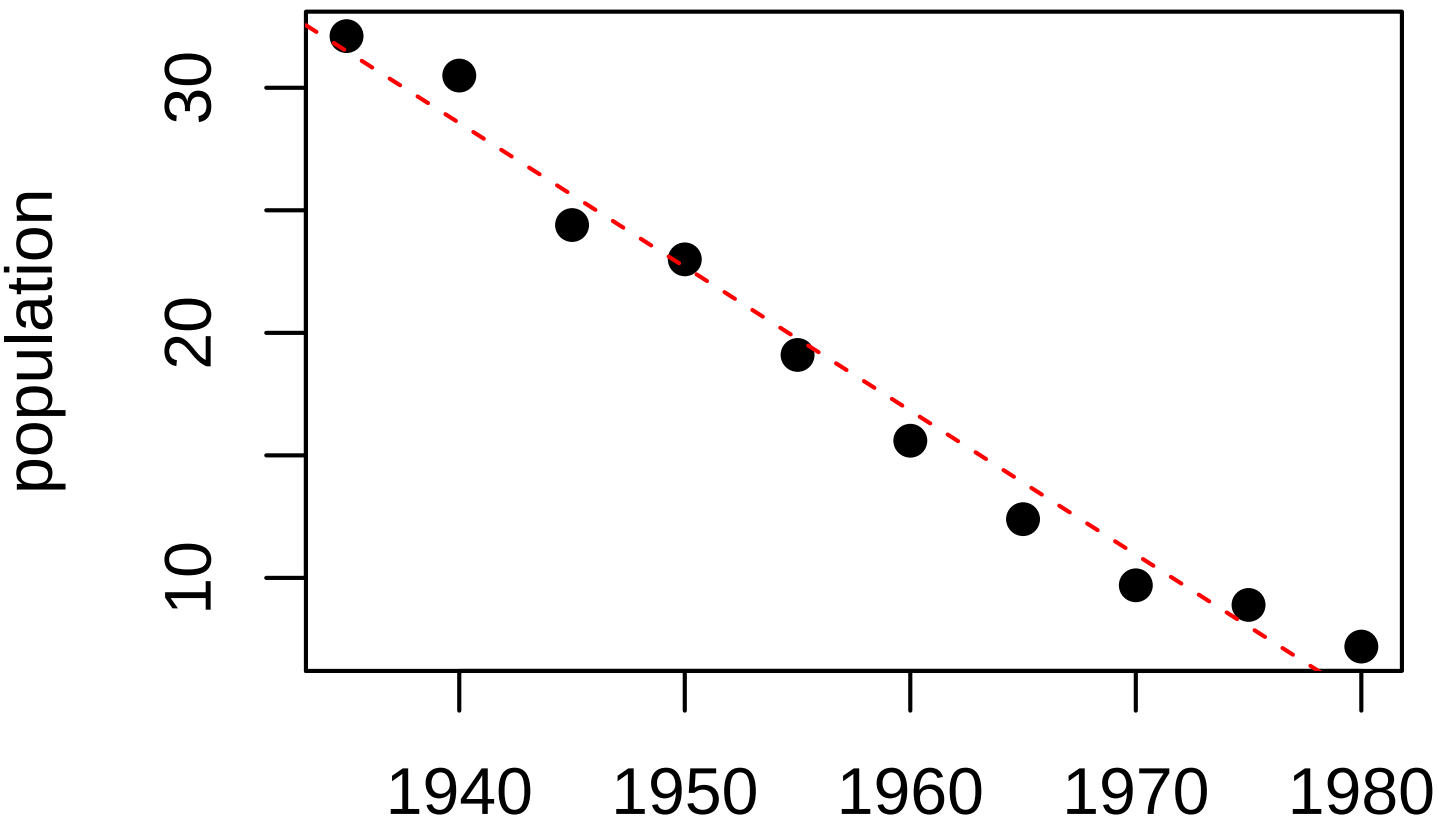
```
plot(year,population,pch=19);

abline(least.squares$coeff,col="red",lty=2);

# pch=19 tells R to draw solid circles;
# lty=2 tells R to draw a dashed line;

# abline tells R to add trendline;
```

R code (Graph)



Solution

a) The scatterplot shows a strong negative association with a straight-line pattern. The regression line (trendline) is $\hat{y} = 1167.14 - 0.587x$.
b) This is the slope - about 0.587 million $(587,000)$ per year during this period. Because $r \approx -0.9884$, the regression line explains $r^2 \approx 97.7\%$ of the variation in population.
c) Substituting, $x = 1990$ gives
$\hat{y} = 1167.14 - 0.587(1990) = -0.99$, an impossible result because a population must be greater than or equal to 0. The rate of decrease in the farm population dropped in the 1980s. Beware of extrapolation.

Association does not imply causation

An association between an explanatory variable x and a response variable y, even if it is very strong, is not by itself good evidence that changes in x actually cause changes in y.

Example

Measure the number of television sets per person x and the average life expectancy y for the world's nations. There is a high positive correlation: nations with many TV sets have higher life expectancies.
The basic meaning of causation is that by changing x we can bring about a change in y. Could we lengthen the lives of people in Rwanda by shipping them TV sets? No. Rich nations have more TV sets than poor nations. Rich nations also have longer life expectancies because they offer better nutrition, clean water, and better health care. There is no cause-and-effect tie between TV sets and length of life.