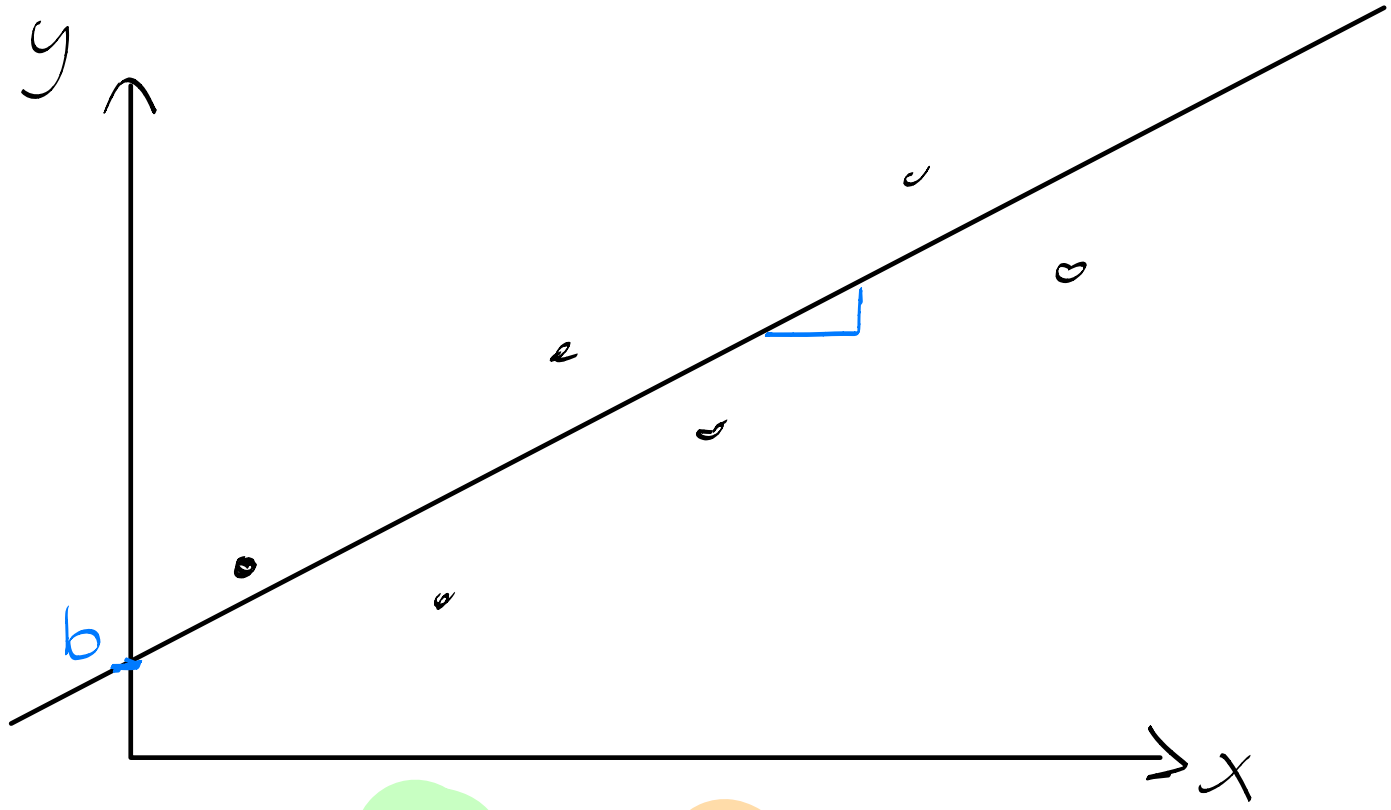
Simple Linear Regression

STA 258
University of Toronto Mississauga

Al Nosedal and Omid Jazi

Winter 2023

x	y
x_1	y_1
x_2	y_2
⋮	⋮
x_n	y_n



H/s $\quad y = mx + b$

slope — m

y - intercept — b

Formalize concept

New $\quad \hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$

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

Forrest Gump.

Measures of Linear Relationship

Recall from STA256

Covariance (sample covariance)

You can compute the covariance, S_{XY} using the following formula:

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

variance → dispersion

covariance → direction of relationship

Covariance

$$cov(X,Y) = s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{\sum_{i=1}^{n} x_i y_i}{n-1} - \frac{n\bar{x}\bar{y}}{n-1} \quad (1)$$

A measure of the direction of variability between 2 variables

It measures how 2 variables move together (not how the magnitude of a change in the variability in one variable affects the other)

$$-\infty < cov(X,Y) < +\infty$$

$cov(X,Y) > 0$: As one variable increases, the other also increases (and vice versa)

$cov(X,Y) < 0$: As one variable increases, the other decreases

$cov(X,Y) \approx 0$: No systematic linear relationship between X and Y.

Not standardized so can be difficult to interpret directly.

covariance Standardized

Coefficient of Correlation. -ve or +ve

$$\text{Cor}(X,Y) = r_{xy} = \frac{S_{xy}}{S_x S_y}$$

(+ve) (+ve)

where

r_{xy} = sample correlation coefficient

S_{xy} = sample covariance

S_x = sample standard deviation of x

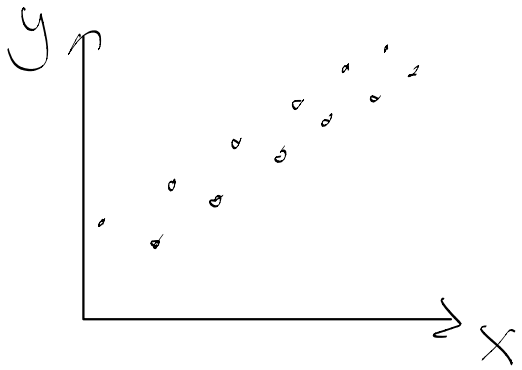
S_y = sample standard deviation of y.

$$-1 \leq r_{xy} \leq +1$$

Correlation (r)

A measure of the strength of the linear relationship between X and Y

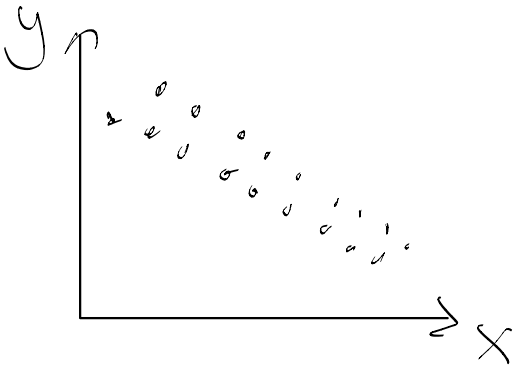
$-1 \leq r \leq +1$



$r \approx +1$

Strong +ve correlation

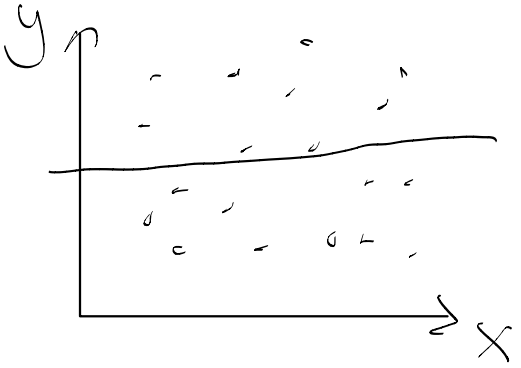
(good linear model)



$r \approx -1$

Strong -ve correlation

(good linear model)



$r \approx 0$ $(cov(X,Y) \approx 0)$

essentially no correlation

Note: $r \approx 0$ suggests a linear relationship does not exist, however some other non-linear relationship may exist

Correlation does <u>not</u> imply causation

i.e. $cor(X, Y) \approx +1$ does not necessarily imply an increase in x causes an increase in y

$cor(X, Y) \approx -1$ does not necessarily imply an increase in x causes a decrease in y

—

Examples:

ice-cream sales and murder

Regression is often used for 2 main purposes:

1/ Determine whether there is a relationship between an independent variable (X) and dependent variable (Y)

2/ To get a linear model for prediction

Example

$b_0 = \hat{\beta}_0$ $\quad$ $b_1 = \hat{\beta}_1$ better for consistent notation

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 $\bar{y}$

sample means

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

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

Solution (cont.)

Now, let's find s_x and s_y

variance

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

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

st. dev

$$s_x = 5.4313$$

Solution

variance y

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

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

sd. dev y $\quad s_y = 11.4017$

Solution

Finally, we find s_{xy} and r

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

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

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

close to -1

shows -ve corr

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);
```

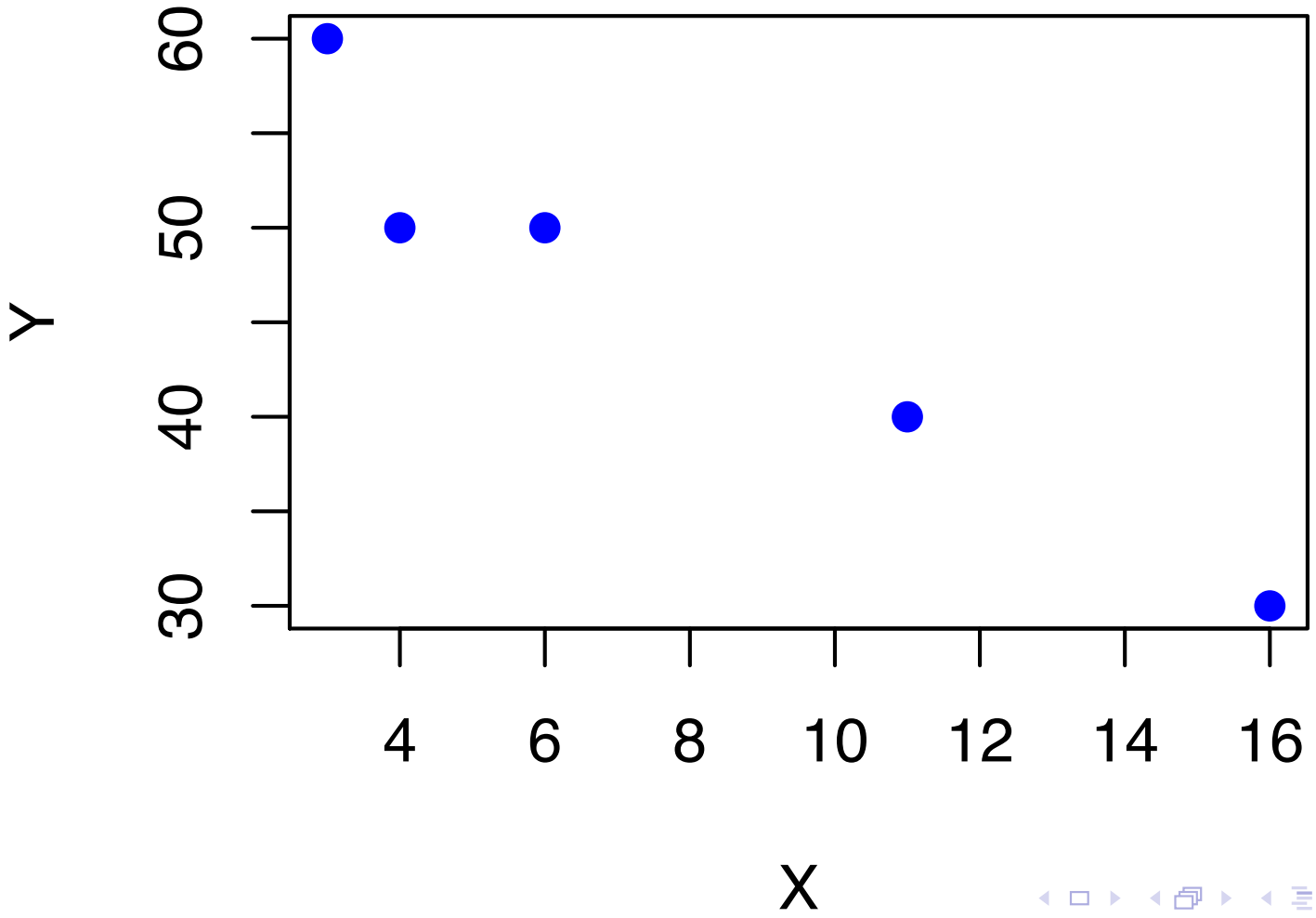
R code

```
# 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

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

$$s_x^2 = \frac{(-10)^2 + (-5)^2 + (-1)^2 + (5)^2 + (11)^2}{4} = 68$$

$$s_x = 8.2462$$

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);
```

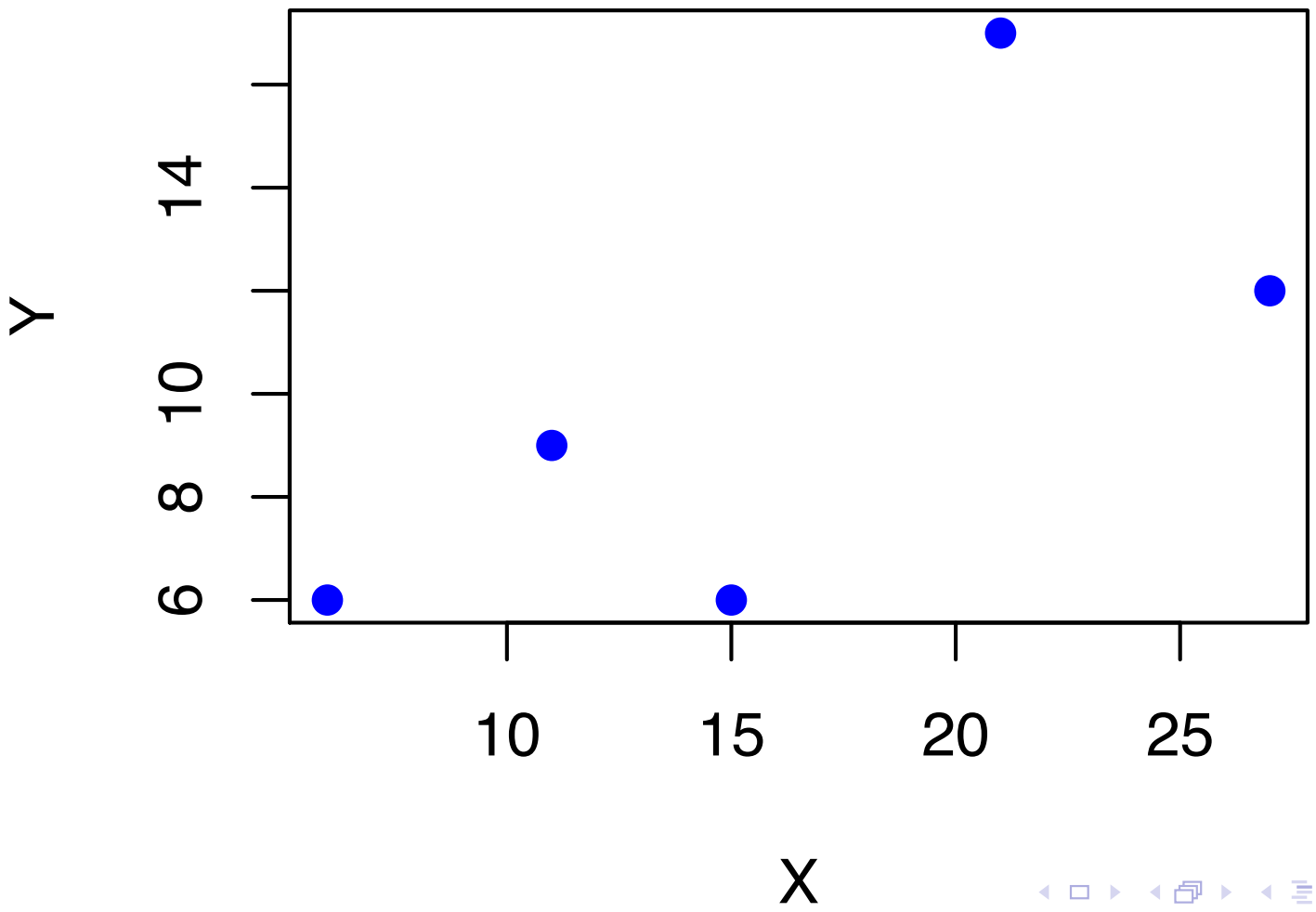
R code (Results)

```
## [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
```

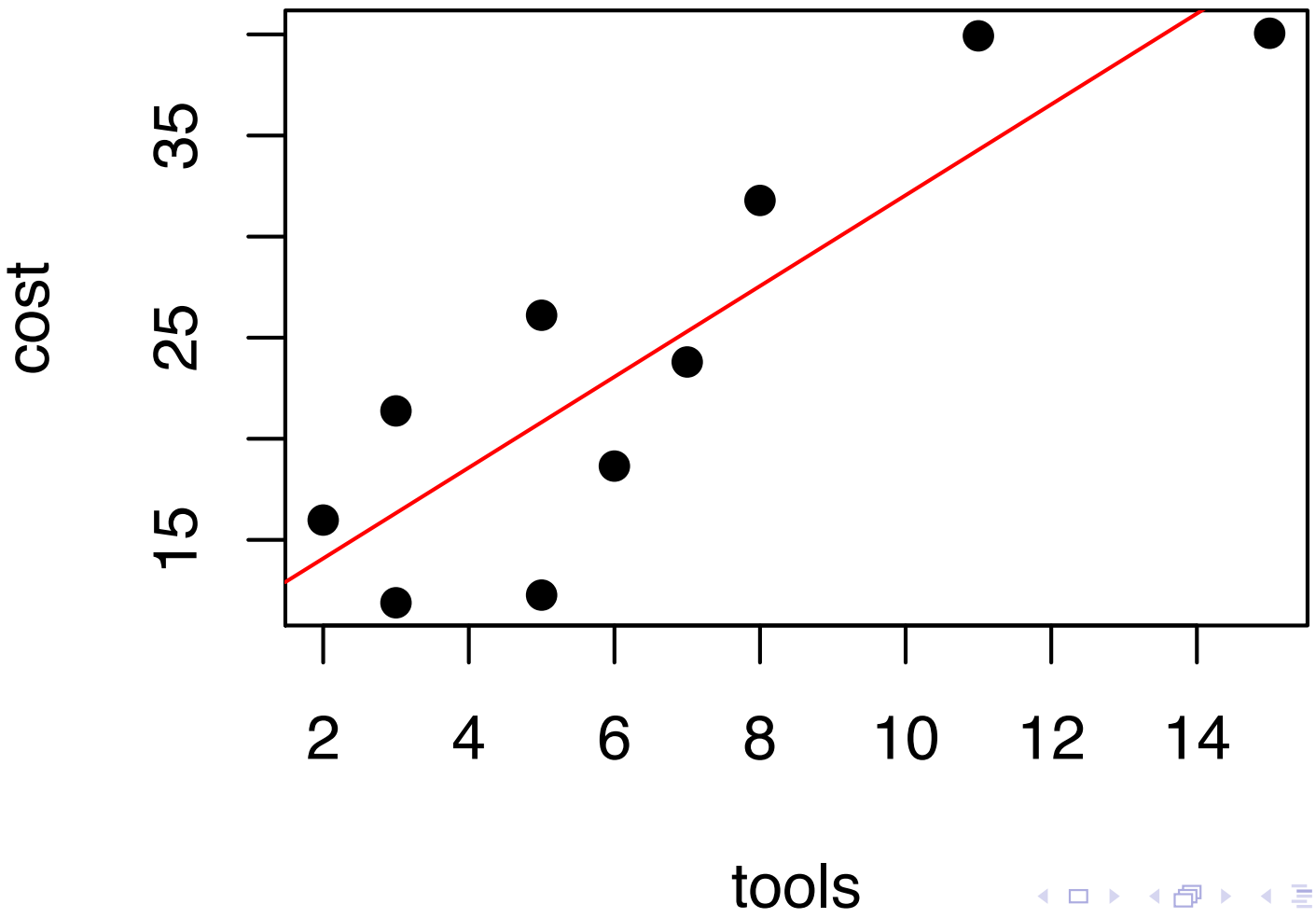
R code, another way

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

least.squares;

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

R code (Graph)

```r
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);
```

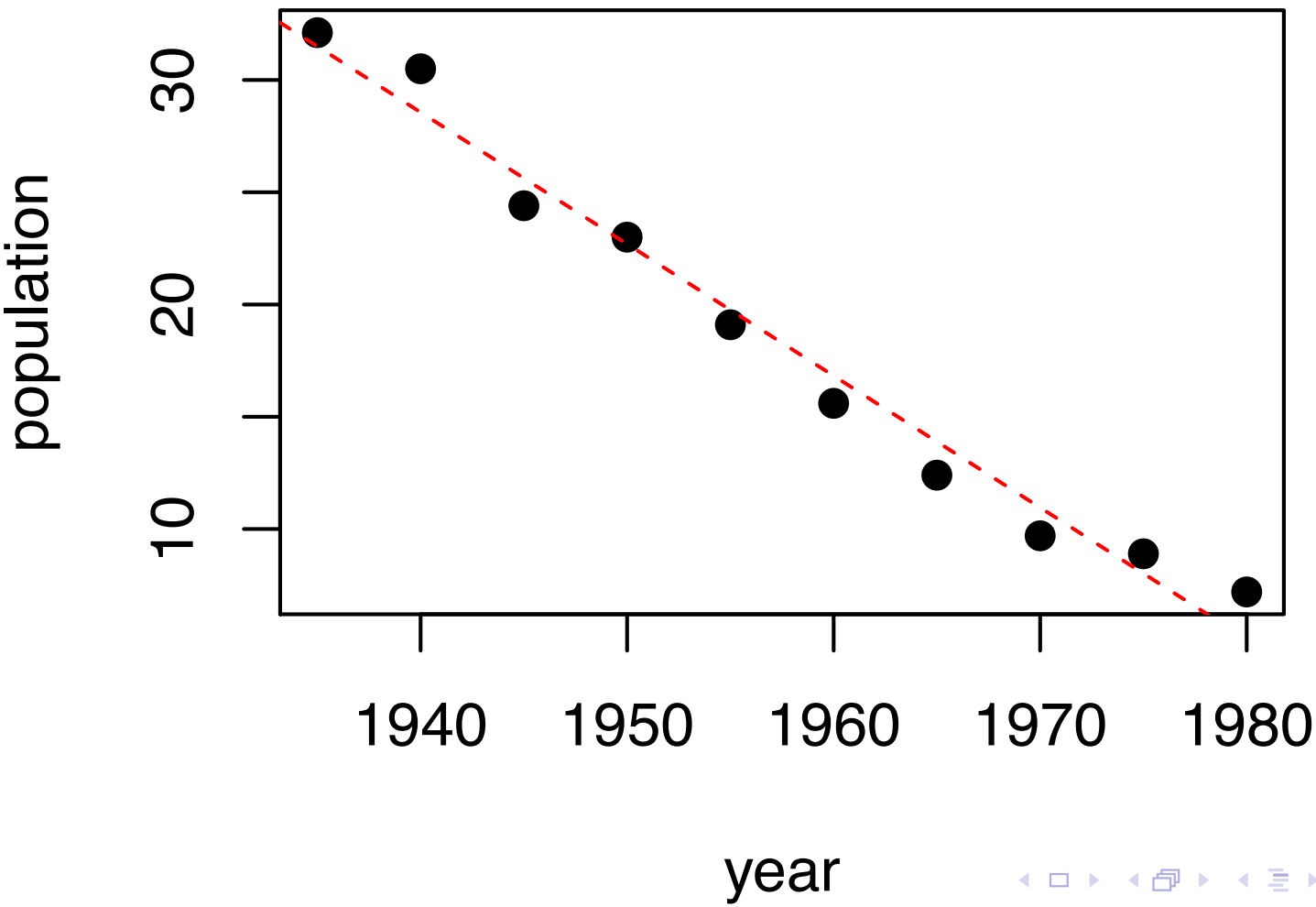
R code, least squares (Results)

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

R code (Graph)

```r
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.

Note

Correlation and Covariance are Symmetric measures

$$Cov(X, Y) = Cov(Y, X)$$

$$Cor(X, Y) = Cor(Y, X)$$

Estimating Regression Model Parameters

β_0 β_1

Least Squares Regression (General)

Interested in a linear model of the form

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$$

p parameters

where

Y : Dependent variable

$X_1, \dots, X_p$: Independent predictors (p)

$\beta_0, \dots \beta_p$: Coefficients $(p+1)$

Using sample data we get estimates of this model of the form

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_p X_p$$

where

$\hat{Y}$: predicted Y

$X_1, \dots, X_p$: Independent predictors (p)

$\hat{\beta}_0, \dots \hat{\beta}_p$: Estimated Coefficients $(p+1)$

Coefficients

β_0 : Intercept

$\beta_1, \dots, \beta_p$: Quantifies how much Y changes with a unit increase in X_j

Simple Linear Regression

Interested in a model of the form

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

where

- Y : Dependent variable
- X : Independent variable
- β_0 : Intercept
- β_1 : Coefficient of X (Slope)

} parameters

Estimate using sample data to get

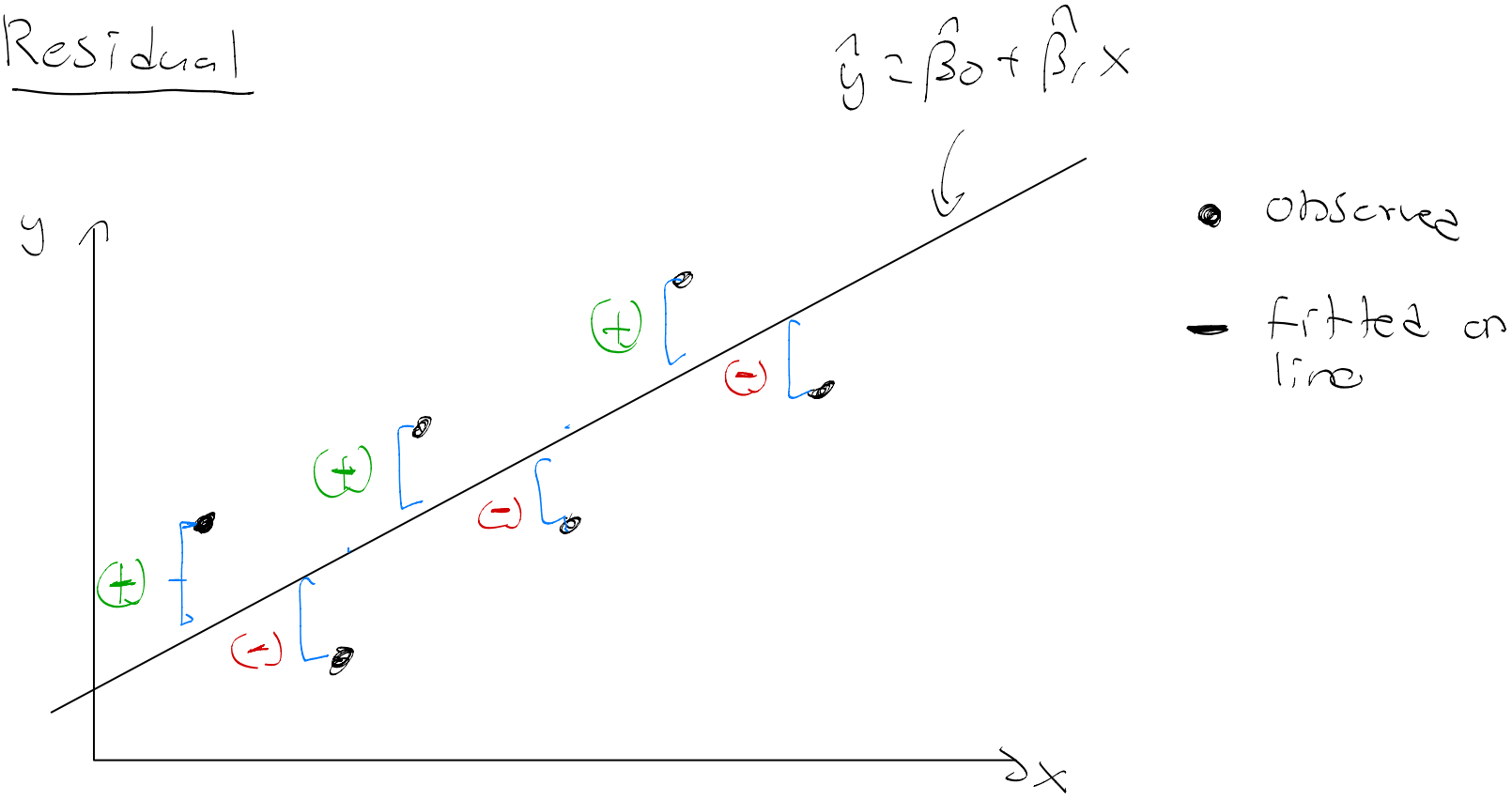
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

where

- $\hat{Y}$: predicted value of Dependent variable
- X : Independent variable
- $\hat{\beta}_0$: Estimate of Intercept
- $\hat{\beta}_1$: Estimate of Coefficient of X (" " Slope)

Obtain $\hat{\beta}_0$, $\hat{\beta}_1$ with calculus

Residual



A residual is the distance between an observed y (sample) and fitted y (regression line)

residual = (observed y) − (fitted y)

$$e_i = y_i - \hat{y}_i$$

residuals can be +ve or −ve

Sum of residuals = 0

$$\sum e_i = \sum y_i - \hat{y}_i = 0$$

Not a unique feature.

$(\bar{x}, \bar{y})$

we derive estimates of β_0 and β_1 by examining sum of squared residuals and minimize this value

objective: minimize $\sum e_i^2$

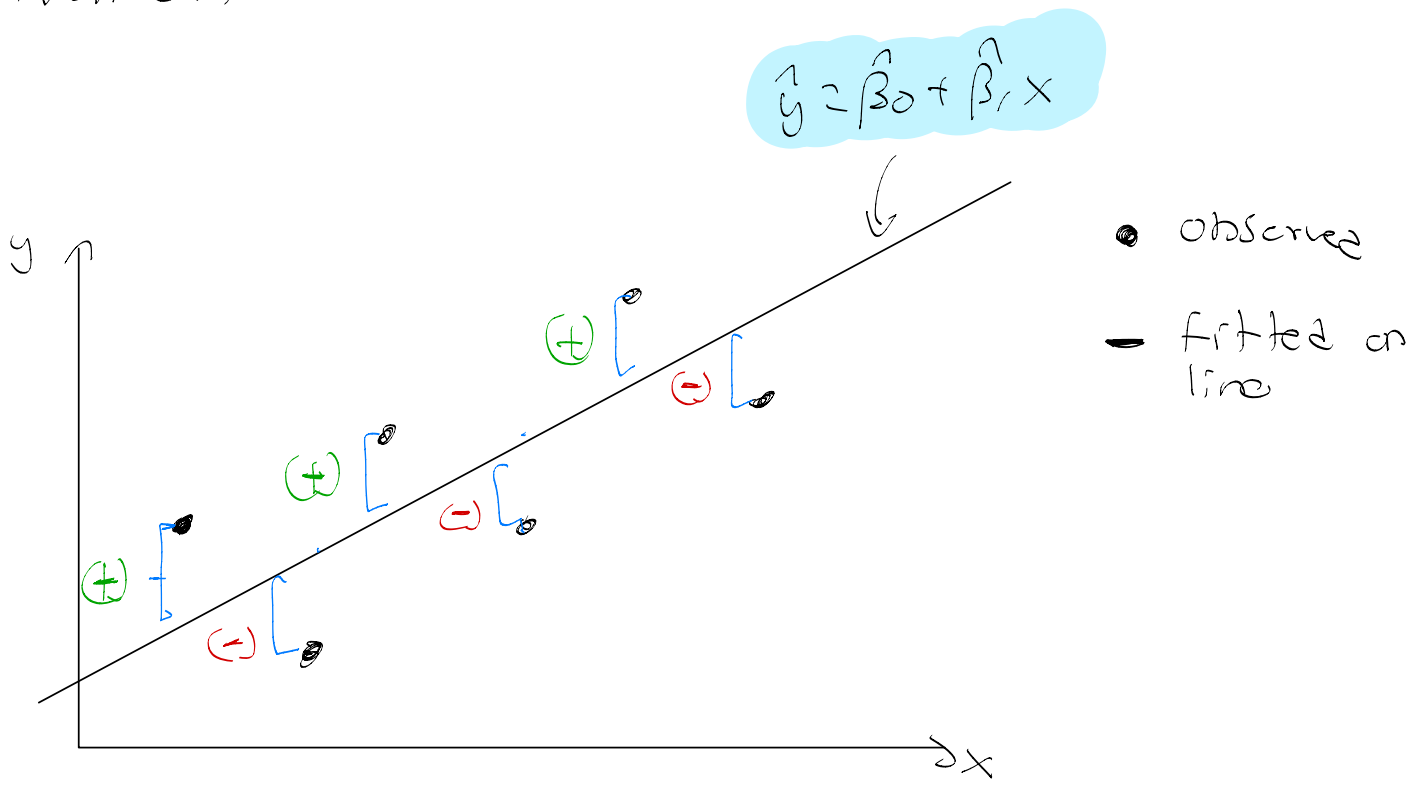
minimize $\sum (y_i - \hat{y}_i)^2$

Minimize squared residuals due to nice properties from calculus.

There is only 1 unique line which minimizes the sum of the squared residuals

↳ least squares regression line.

Derivation



Examine sum of residuals squared

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

$$= \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

$$= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

To find $\hat{\beta}_0$ and $\hat{\beta}_1$ which minimizes $\sum e_i^2$

Find $\frac{\partial}{\partial \hat{\beta}_0} \sum e_i^2$ and $\frac{\partial}{\partial \hat{\beta}_1} \sum e_i^2$

$$\frac{\partial}{\partial \hat{\beta}_0} \sum e_i^2 = \frac{\partial}{\partial \hat{\beta}_0} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

chain rule

$$= \sum_{i=1}^{n} 2\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)(-1)$$

$$= \sum_{i=1}^{n} 2\left(\hat{\beta}_0 + \hat{\beta}_1 x_i - y_i\right)$$

at min $\frac{\partial}{\partial \hat{\beta}_0} \sum e_i^2 = 0$

$$\frac{\partial}{\partial \hat{\beta}_0} \sum e_i^2 = 0$$

$$\sum_{i=1}^{n} \frac{2\left(\hat{\beta}_0 + \hat{\beta}_1 x_i - y_i\right)}{2} = 0$$

$$\sum_{i=1}^{n} \hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i = 0$$

$\bar{x} = \frac{\sum x_i}{n}$

$n\bar{x} = \sum x_i$

$n\hat{\beta}_0 \quad n\bar{x} \quad n\bar{y}$

$$n\hat{\beta}_0 + n\hat{\beta}_1 \bar{x} - n\bar{y} = 0$$

n

$$\hat{\beta}_0 + \hat{\beta}_1 \bar{x} - \bar{y} = 0$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Repeat to find expression for $\hat{\beta}_1$

$$\frac{\partial}{\partial \hat{\beta}_1} \sum e_i^2 = \frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

$$= \sum_{i=1}^{n} 2\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)\left(-x_i \right)$$

$$= \sum_{i=1}^{n} 2\left(\hat{\beta}_1 x_i^2 + \hat{\beta}_0 x_i - x_i y_i \right)$$

at min, $\frac{\partial}{\partial \hat{\beta}_1} \sum e_i^2 = 0$

$$\frac{\partial}{\partial \hat{\beta}_1} \sum e_i^2 = 0$$

$$\sum_{i=1}^{n} \frac{2}{2}\left(\hat{\beta}_1 x_i^2 + \hat{\beta}_0 x_i - x_i y_i \right) = \frac{0}{2}$$

$$\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} \hat{\beta}_0 x_i - \sum_{i=1}^{n} x_i y_i = 0$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \text{from above}$$

$$\hat{\beta}_1 \sum x_i^2 + \sum (\bar{y} - \hat{\beta}_1 \bar{x}) x_i - \sum x_i y_i = 0$$

$$\hat{\beta}_1 \sum x_i^2 + \sum (\bar{y} x_i - \hat{\beta}_1 \bar{x} x_i) - \sum x_i y_i = 0$$

$$\hat{\beta}_1 \sum x_i^2 + \bar{y} \sum x_i - \hat{\beta}_1 \bar{x} \sum x_i - \sum x_i y_i = 0$$

$\sum x_i = n\bar{x}$; $\sum x_i = n\bar{x}$

$$\hat{\beta}_1 \sum x_i^2 + n\bar{x}\bar{y} - n\hat{\beta}_1 \bar{x}^2 - \sum x_i y_i = 0$$

$$\hat{\beta}_1 \left(\sum x_i^2 - n\bar{x}^2 \right) + n\bar{x}\bar{y} - \sum x_i y_i = 0$$

$$\hat{\beta}_1 \left(\sum x_i^2 - n\bar{x}^2 \right) = \sum x_i y_i - n\bar{x}\bar{y}$$

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

Can show using manipulation that expression above for $\hat{\beta}_1$ is equivalent to

$$\hat{\beta}_1 = \frac{\overbrace{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}^{SP_{xy}}}{\underbrace{\sum_{i=1}^{n} (x_i - \bar{x})^2}_{SS_{xx}}}$$

The Method of Least Squares

The regression line

$$E(Y) = \beta_0 + \beta_1 x$$

is fitted to the data points $(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)$ by finding the line that is closest to the data in some sense. There are many ways in which closeness can be defined, but the method most generally used is to consider the vertical deviations between the line and the data points

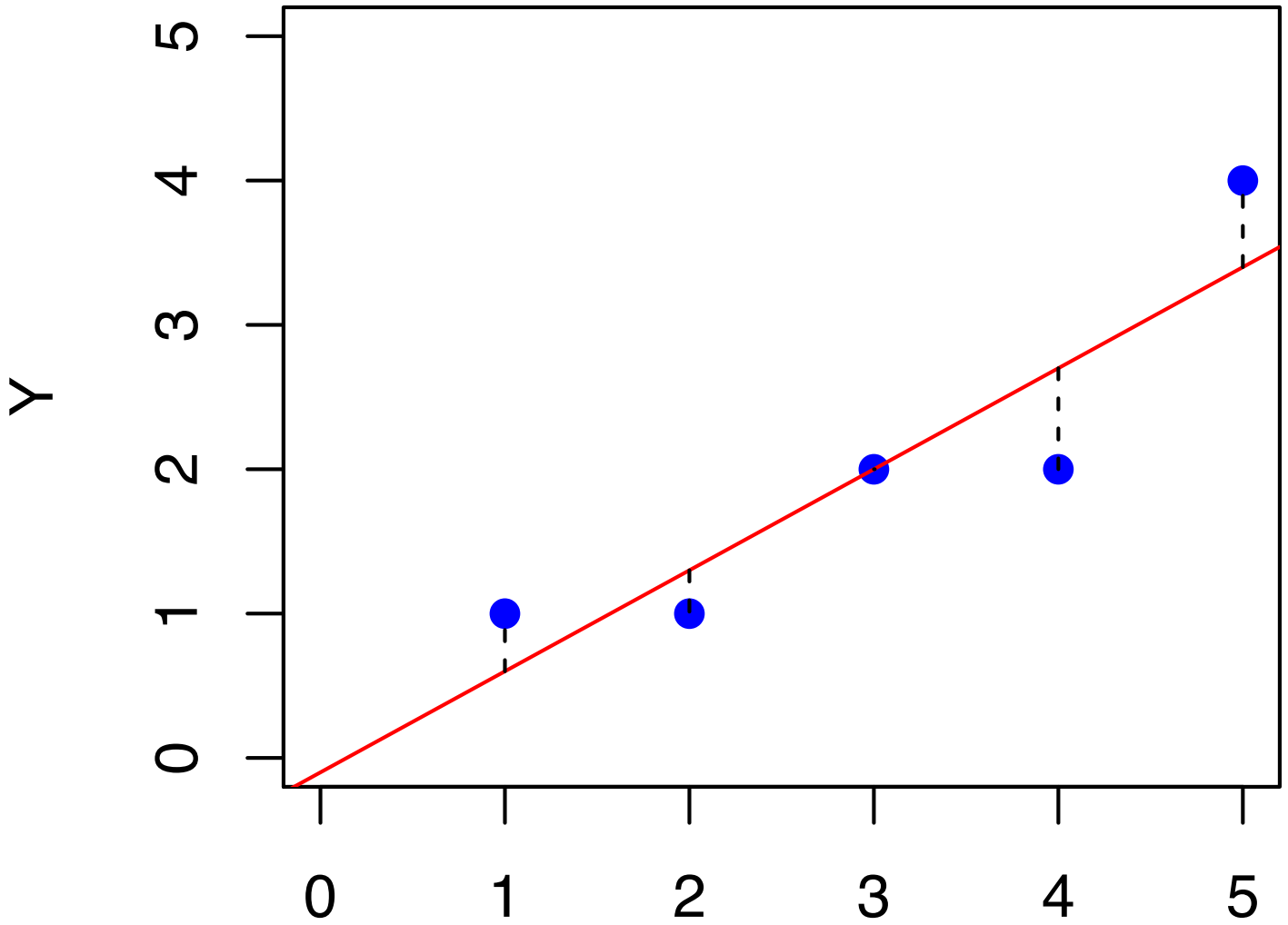
$$y_i - (\beta_0 + \beta_1 x_i), 1 \leq i \leq n.$$

The Method of Least Squares

The fitted line is chosen to be the line that minimizes the sum of the squares of these vertical deviations

$$Q = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

and this is referred to as the least squares fit.
(The quantity Q is also called the **sum of squares for error**, SSE.)


Y
0
1
2
3
4
5
0
1
2
3
4
5

The Method of Least Squares

The parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are therefore the values that minimize the quantity Q. They are found by taking partial derivatives of Q with respect to β_0 and β_1 and setting the resulting expressions equal to zero.

Partial derivatives

$$\frac{\partial Q}{\partial \beta_0} = -\sum_{i=1}^{n} 2[y_i - (\beta_0 + \beta_1 x_i)]$$

$$\frac{\partial Q}{\partial \beta_1} = -\sum_{i=1}^{n} 2x_i[y_i - (\beta_0 + \beta_1 x_i)]$$

Normal Equations

The parameter estimates are the solutions to the equations:

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i.$$

These equations are known as the **normal equations**.

Solution to Normal Equations

The normal equations can be solved to give

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

and then $\hat{\beta}_0$ can be calculated as

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Solution to Normal Equations

Notice that with the notation $S_{XX} = \sum_{i=1}^{n}(x_i - \bar{x})^2$ and $S_{XY} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ the parameter estimate $\hat{\beta}_1$ can be written as

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}.$$

(Please, verify this. HW?)

Example

Suppose an appliance store conducts a 5-month experiment to determine the effect of advertising on sales revenue. The results are shown in a table below. The relationship between sales revenue, y, and advertising expenditure, x, is hypothesized to follow a first-order linear model, that is,

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

where
y = dependent variable
x = independent variable
β_0 = y-intercept
β_1 = slope of the line
ϵ = error variable

Example

Obtaining least square reg line

Month	Advertising Expenditure x ($ hundreds)	Sales Revenue y ($ thousands)
1	1	1
2	2	1
3	3	2
4	4	2
5	5	4

a) Obtain the least squares estimates of β_0 and β_1, and state the estimated regression function.

b) Plot the estimated regression function and the data.

onth	Advertising Expenditure x ($ hundreds)	Sales Revenue y ($ thousands)
1	1	1
2	2	1
3	3	2
4	4	2
5	5	4

$\sum x_i = 15$ $\sum y_i = 10$

$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
$1-3 = -2$	$1-2 = -1$	$(-2)(-1) = 2$	$-2^2 = 4$
$2-3 = -1$	$1-2 = -1$	$(-1)(-1) = 1$	$-1^2 = 1$
$3-3 = 0$	$2-2 = 0$	0	0
$4-3 = 1$	$2-2 = 0$	0	0
$5-3 = 2$	$4-2 = 2$	$(2)(2) = 4$	$2^2 = 4$
0	0	$\sum = 7$	$\sum$
		S_{xy}	SS_{xx}

$\bar{x} = \frac{\sum x_i}{n} = \frac{15}{5} = 3$

$\bar{y} = \frac{\sum y_i}{n} = \frac{10}{5} = 2$

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{7}{10} = 0.7$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$= 2 - (0.7)(3)$

$= -0.1$

Equation of OLS reg line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{y} = -0,11 + 0,17x$$

Review: Simple Linear Regression (SLR)

Model: $y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$

Estimated (sample data): $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

- $\hat{y}$: predicted y
- $\hat{\beta}_0$: estimate of intercept
- $\hat{\beta}_1$: estimate of slope

Objective: $\min \sum (y_i - \hat{y}_i)^2$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

] calculus

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

(residuals)^2

Can use regression model for estimation and prediction

For example in previous class

$$\hat{y} = -0.1 + 0.7x$$

$\hat{\beta}_0$ $\hat{\beta}_1$

$\hat{y}$: predicted sales ($1000)

x: advertising expenditure ($100)

Estimate sales when expenditure is $250 ($x = 2.5$)

$x = 2.5$ $\quad \hat{y} = -0.1 + 0.7(2.5)$

$= 1.65$

predicted sales is $1.65 \times \underbrace{\$1000}_{\text{units}} = \$1,650.$

Calculate the residuals $\hat{y} = -0.1 + 0.7x$

$e_i = y_i - \hat{y}_i$

will show up soon

x	observed y y	predicted y $\hat{y}$	residuals $e_i = y_i - \hat{y}_i$	$(\text{residuals})^2$ e_i^2
1	1	$-0.1 + 0.7(1) = 0.6$	$1 - 0.6 = 0.4$	$(0.4)^2$
2	1	$-0.1 + 0.7(2) = 1.3$	$1 - 1.3 = -0.3$	$(-0.3)^2$
3	2	$-0.1 + 0.7(3) = 2.0$	$2 - 2 = 0$	0^2
4	2	$-0.1 + 0.7(4) = 2.7$	$2 - 2.7 = -0.7$	$(-0.7)^2$
5	4	$-0.1 + 0.7(5) = 3.4$	$4 - 3.4 = 0.6$	$(0.6)^2$
			0	Sum square error (SSE)

min

used for inference

a) Solution (R Code, one way)

```
x=c(1,2,3,4,5);
y=c(1,1,2,2,4);
x.bar=mean(x);
y.bar=mean(y);
S.xx=(x-x.bar)%*%(x-x.bar);
S.xy=(x-x.bar)%*%(y-y.bar);
b1=S.xy/S.xx;
b0=y.bar-(S.xy/S.xx)*x.bar;
```

a) Solution

$\bar{x} = 3$, $\bar{y} = 2$, $S_{XX} = 10$, $S_{XY} = 7$.
Then, the slope of the least squares line is

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = 0.7$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = -0.1$$

The least squares line is thus

$$\hat{y} = -0.1 + 0.7x$$

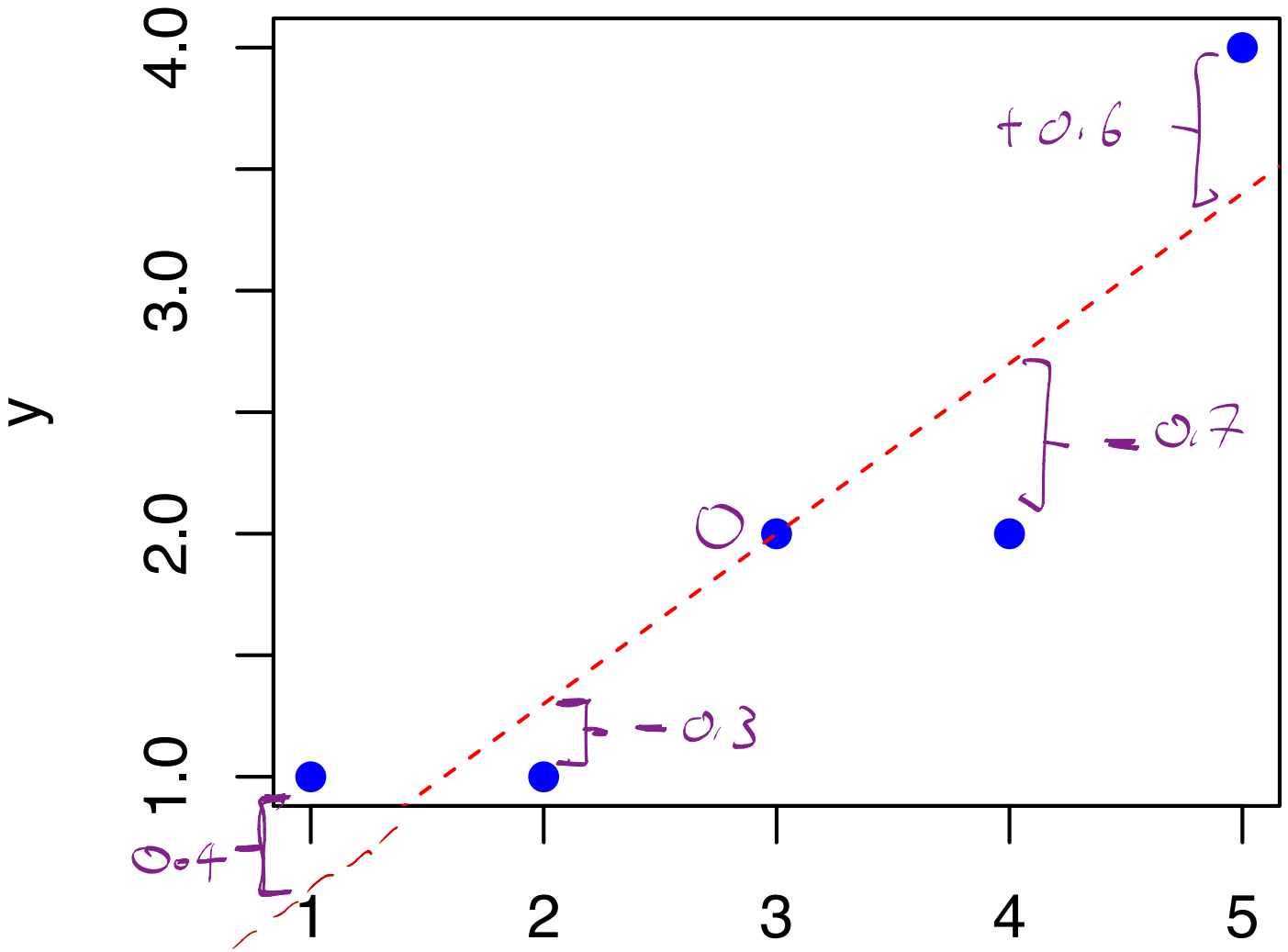
a) Solution (R Code, another way)

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

## [1] 1 2 3 4 5

y=c(1,1,2,2,4);
y;

## [1] 1 1 2 2 4

linear.reg=lm(y~x);

coef(linear.reg);

## (Intercept)           x
##        -0.1         0.7
```

b) Solution (R Code)

```r
#Scatterplot with line of best fit
plot(x, y, main="Scatterplot: Simple Linear Regression",
xlab="x", ylab="y", pch=19,col="blue");

abline(coef(linear.reg), col="red",lty=2);
```

Scatterplot: Simple Linear Regression



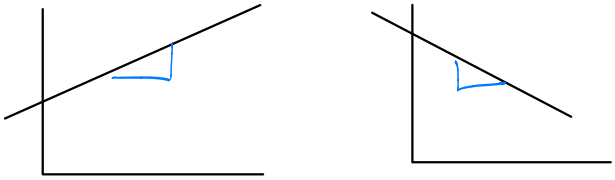
Interpreting $\hat{\beta}_0$ and $\hat{\beta}_1$

Model $Y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$

Estimate using $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$

$\hat{\beta}_0$: estimate of intercept

$\hat{\beta}_1$: estimate of slope



$\hat{\beta}_1$

For an increase in X by 1 unit, we expect y to increase / decrease by

($\hat{\beta}_1 > 0$) ($\hat{\beta}_1 < 0$)

$\hat{\beta}_1$ units on average

Let $x_1 = x^*$

Let $x_2 = x^* + 1$

At x_1, $\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x^*$

At x_2, $\hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 (x^* + 1) = \hat{\beta}_0 + \hat{\beta}_1 x^* + \hat{\beta}_1$

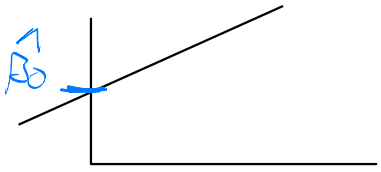
$\Delta x = x_2 - x_1 = x^* + 1 - x^* = 1$

$$\Delta y = y_2 - y_1$$

$$= \hat{\beta}_0 + \hat{\beta}_1 x^* + \hat{\beta}_1 - \left(\hat{\beta}_0 + \hat{\beta}_1 x^*\right)$$

$$= \hat{\beta}_0 + \hat{\beta}_1 x^* + \hat{\beta}_1 - \hat{\beta}_0 - \hat{\beta}_1 x^*$$

$$= \hat{\beta}_1$$



$\hat{\beta}_0$

The predicted value of Y when $X = 0$

Note: may or may not have a real-world interpretation (often due to limitations in model. Perhaps predictors are necessary)

For our example on sales and advertising

$$\hat{y} = -0.1 + 0.7x$$

$\hat{y}$: predicted sales ($1000)

$\hat{\beta}_0$

$\hat{\beta}_1$

x: advertising expenditure ($100)

$\hat{\beta}_1 = 0.7$

An increase in advertising expenditure (x) by $1 \times \underbrace{\$100}_{\text{unit}} = \$100$ would result in an increase ($\hat{\beta}_1 > 0$)

in predicted sales ($\hat{y}$) by $\underbrace{0.7}_{\hat{\beta}_1} \times \underbrace{\$1000}_{\text{units}} = \$700$

on average.

$\hat{\beta}_0 = -0.1$

When advertising expenditure is 0 ($x = 0$), we expect $-0.1 \times \$1000 = -\$100$ in sales.

Does this make sense?

Possibly.

One possible interpretation is inventory is returned. could be write offs

May not have realistic interpretation.

Correlation Coefficient

The measure ρ of linear association between two variables X and Y is estimated by the **sample correlation coefficient** r, where

$$
\begin{aligned}
r &= \frac{\text{sample covariance}}{\sqrt{(\text{sample variance of } X)(\text{sample variance of } Y)}} \\
&= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum(x_i - \bar{x})^2][\sum(y_i - \bar{y})^2]}} \\
&= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \\
&= \hat{\beta}_1\sqrt{\frac{S_{xx}}{S_{yy}}}
\end{aligned}
$$

The three deviations associated with a data point

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

$$\begin{aligned} \text{Total deviation} &= \text{Unexplained deviation (Error)} \\ &\quad + \text{Explained deviation (Regression)} \end{aligned}$$

The three deviations associated with a data point

We square all three deviations for each one of our data points, and sum over all n points. Here, cross terms drop out, and we are left with the following equation:

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

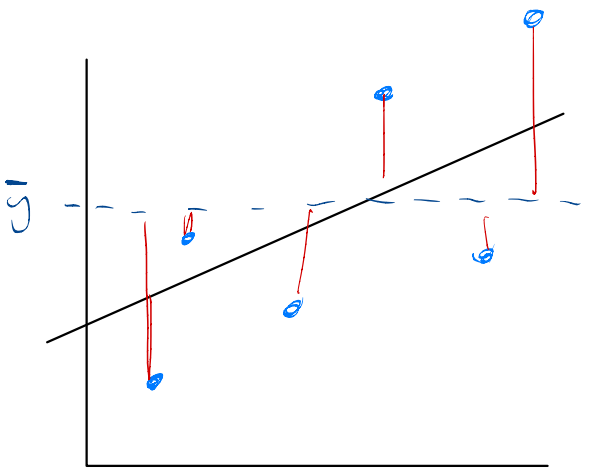
obs y — fitted y

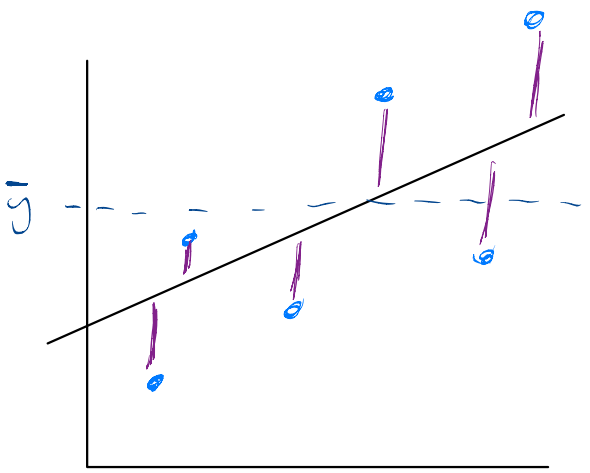
e_i^2

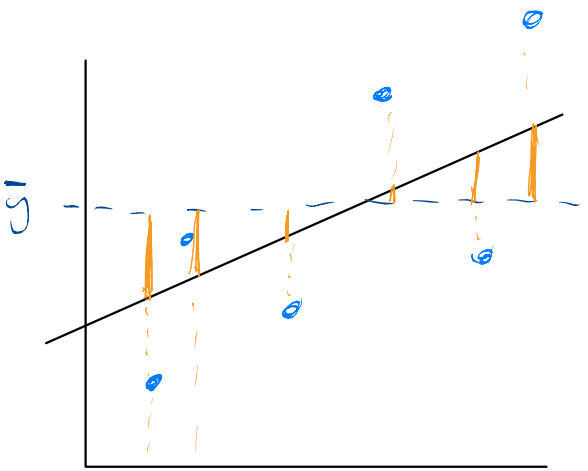
$\sum$ (residuals)^2

SST = SSE + SSR

Total sum of squares = Sum of squares for error + Sum of squares for regression.







$\sum (y_i - \bar{y})^2$ = $\sum (y_i - \hat{y}_i)^2$ + $\sum (\hat{y}_i - \bar{y})^2$

Total variation

SST

Sum square error (SSE)

$\sum (\text{residuals})^2$

$\sum e_i^2$

unexplained variation

Sum Square regression (SSR)

explained variation by regression model

Review: $y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Coefficient of correlation (r)

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

measured the strength of the linear relationship between x and y

$$-1 \leq r \leq +1$$

Coefficient of Determination (r^2)

It is the proportion of variability in y which is explained by x.

$$r^2 = \frac{SSR}{SST}$$

variation explained by regression

Total variation

$$SST = SSE + SSR \longrightarrow SSR = SST - SSE$$

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

$$= 1 - \frac{SSE}{SST}$$

unexplained variation

Total variation

$$0 \leq r^2 \leq +1$$

Coefficient of determination

We define the coefficient of determination as the sum of squares due to the regression divided by the total sum of squares.

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

The coefficient of determination can be interpreted as the proportion of the variation in Y that is explained by the regression relationship of Y with X (or the proportion of the total corrected sum of squares explained by the regression).

Example

Car dealers across North America use the so-called Blue Book to help them determine the value of used cars that their customers trade in when purchasing new cars. The book lists the trade-in values for all models of cars. It provides alternative values for each car model according to its condition and optional features. However, the Blue Book does not indicate the value determined by the odometer reading, despite the fact that a critical factor for used-car buyers is how far the car has been driven.

Example

To examine this issue, a used-car dealer randomly selected 100 3-year old Toyota Camrys that were sold at auction during the past month. Each car was in top condition and equipped with all the features that come standard with this car. The dealer recorded the price ($1000) and the number of miles (thousands) on the odometer. The dealer wants to find the regression line and coefficient of detemination. Describe what this statistic tells you about the regression model.

Reading our data

```
# url of camrys data;
camrys_url =
"https://mcs.utm.utoronto.ca/~nosedal/data/camrys.txt"

# importing data into R;
camrys= read.table(camrys_url,header=TRUE);

names(camrys);
attach(camrys);
```

R code

lm(y ~ x, data_source)

```
# Fitting Linear Regression Model;

model=lm(camrys$Price~Odometer,data=camrys);
```

y

x

data source

camrys

Price

Odometer

```
summary(model);

##
## Call:
## lm(formula = camrys$Price ~ Odometer, data = camrys)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68679 -0.27263  0.00521  0.23210  0.70071
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.248727   0.182093   94.72   <2e-16
## Odometer    -0.066861   0.004975  -13.44   <2e-16
##
## (Intercept) ***
## Odometer    ***
## ---
## Signif. codes:
```

```
> summary(model);

Call:
lm(formula = camrys$Price ~ Odometer, data = camrys)

Residuals:
     Min       1Q   Median       3Q      Max
-0.68679 -0.27263  0.00521  0.23210  0.70071

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.248727   0.182093   94.72   <2e-16 ***
Odometer    -0.066861   0.004975  -13.44   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3265 on 98 degrees of freedom
Multiple R-squared:  0.6483,    Adjusted R-squared:  0.6447
F-statistic: 180.6 on 1 and 98 DF,  p-value: < 2.2e-16
```

R output on Camry example

- Obtain equation of regression line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$\hat{y}$: predicted price

$\hat{\beta}_0$: estimate of y-intercept

$\hat{\beta}_1$: estimate of slope

x: 1000's miles on odometer

$\hat{\beta}_0 = 17.248$

$\hat{\beta}_1 = -0.0669$

Equation of regression line

$$\hat{y} = 17.248 - 0.0669x$$

- From R output, what is r^2?

use multiple R^2: $r^2 = 0.6483 = 64.83\%$

- Provide an interpretation of the r^2.

Approximately 64.83% of the variability in sale price (y) is explained by the number of miles on odometer (x) through the regression model.

Relationship between r and r^2

r^2 can be obtained by squaring r.

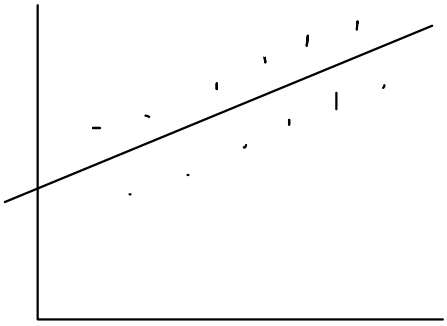
$(0 \le r^2 \le 1)$

we can obtain r from r^2 by

$$r = \pm\sqrt{r^2}$$

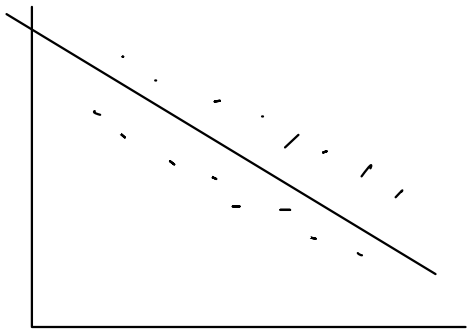
$(-1 \le r \le 1)$

sign depends on the sign of $\hat{\beta}_1$



$\hat{\beta}_1 > 0$

$r > 0$ $\quad r = +\sqrt{r^2}$



$\hat{\beta}_1 < 0$

$r < 0$ $\quad r = -\sqrt{r^2}$

For the camry example

$\hat{y} = 17.248 - 0.0669x$ $\quad r^2 = 0.6483$

$$r = -\sqrt{r^2} = -\sqrt{0.6483} = -0.8052$$

Interpretation

We found that r^2 is equal to 0.6483 (Multiple R-squared). This statistic tells us that 64.83% of the variation in the auction selling prices is explained by the variation in the odometer readings. The remaining 35.17% is unexplained. In general, the higher the value of r^2, the better the model fits the data. From the HT of β_1 we already know that there is evidence of a linear relationship. The coefficient of determination merely supplies us with a measure of the strength of that relationship.