

Lecture 13.2, MATH-57091 Probability and Statistics for High-School Teachers.

Artem Zvavitch

Department of Mathematical Sciences, Kent State University

December 1, 2014.

We are very often interested in trying to determine the relationship between a pair of variables.

We are very often interesting in trying to determine the relationship between a pair of variables. For example, how does amount of time a student spent on HW depends on the outcome of his work?

We are very often interesting in trying to determine the relationship between a pair of variables. For example, how does amount of time a student spent on HW depends on the outcome of his work? Or how the amount of money spent in advertising a new product related to the first month of sales?

We are very often interesting in trying to determine the relationship between a pair of variables. For example, how does amount of time a student spent on HW depends on the outcome of his work? Or how the amount of money spent in advertising a new product related to the first month of sales?

In many situation the value of the second value is given to us as outcome of the value of the first value (if you spend 50 dollars on advertisement you will get 10000 in sales during the first month).

We are very often interesting in trying to determine the relationship between a pair of variables. For example, how does amount of time a student spent on HW depends on the outcome of his work? Or how the amount of money spent in advertising a new product related to the first month of sales?

In many situation the value of the second value is given to us as outcome of the value of the first value (if you spend 50 dollars on advertisement you will get 10000 in sales during the first month). The variable whose value is determined first is called **input** or **independent** variable and the other is called **response** or **dependent** variable.

We are very often interesting in trying to determine the relationship between a pair of variables. For example, how does amount of time a student spent on HW depends on the outcome of his work? Or how the amount of money spent in advertising a new product related to the first month of sales?

In many situation the value of the second value is given to us as outcome of the value of the first value (if you spend 50 dollars on advertisement you will get 10000 in sales during the first month). The variable whose value is determined first is called **input** or **independent** variable and the other is called **response** or **dependent** variable. We will denote the independent variable by x and Y will be the corresponding value of the dependent variable.

We are very often interesting in trying to determine the relationship between a pair of variables. For example, how does amount of time a student spent on HW depends on the outcome of his work? Or how the amount of money spent in advertising a new product related to the first month of sales?

In many situation the value of the second value is given to us as outcome of the value of the first value (if you spend 50 dollars on advertisement you will get 10000 in sales during the first month). The variable whose value is determined first is called **input** or **independent** variable and the other is called **response** or **dependent** variable. We will denote the independent variable by x and Y will be the corresponding value of the dependent variable. The simplest type of relationship between this pair of variables is straight-line, or **linear** relation of the form:

$$Y = \alpha + \beta x.$$

We are very often interesting in trying to determine the relationship between a pair of variables. For example, how does amount of time a student spent on HW depends on the outcome of his work? Or how the amount of money spent in advertising a new product related to the first month of sales?

In many situation the value of the second value is given to us as outcome of the value of the first value (if you spend 50 dollars on advertisement you will get 10000 in sales during the first month). The variable whose value is determined first is called **input** or **independent** variable and the other is called **response** or **dependent** variable. We will denote the independent variable by x and Y will be the corresponding value of the dependent variable. The simplest type of relationship between this pair of variables is straight-line, or **linear** relation of the form:

$$Y = \alpha + \beta x.$$

The hope is that once we find α and β using just a few values of (x, Y) we will be possible to predict exactly the response for any value of the input variable.

We are very often interesting in trying to determine the relationship between a pair of variables. For example, how does amount of time a student spent on HW depends on the outcome of his work? Or how the amount of money spent in advertising a new product related to the first month of sales?

In many situation the value of the second value is given to us as outcome of the value of the first value (if you spend 50 dollars on advertisement you will get 10000 in sales during the first month). The variable whose value is determined first is called **input** or **independent** variable and the other is called **response** or **dependent** variable. We will denote the independent variable by x and Y will be the corresponding value of the dependent variable. The simplest type of relationship between this pair of variables is straight-line, or **linear** relation of the form:

$$Y = \alpha + \beta x.$$

The hope is that once we find α and β using just a few values of (x, Y) we will be possible to predict exactly the response for any value of the input variable. Unfortunately (or Fortunately!), life is not so easy and in many cases it does not work and the prediction is just and estimation which is valid **subject to random error**.

Simple Linear Regression.

Let us start with very simple example.

An area manager in a department store wants to study the relationship between the number of workers on duty at a certain department and the value (in hundreds) of merchandise lost to shoplifters:

Number of Workers	Loss
1	2
2	3

Simple Linear Regression.

Let us start with very simple example.

An area manager in a department store wants to study the relationship between the number of workers on duty at a certain department and the value (in hundreds) of merchandise lost to shoplifters:

Number of Workers	Loss
1	2
2	3

Clearly we should take "Number of Workers" as an input variable and "Loos" as response.

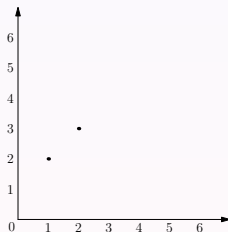
Simple Linear Regression.

Let us start with very simple example.

An area manager in a department store wants to study the relationship between the number of workers on duty at a certain department and the value (in hundreds) of merchandise lost to shoplifters:

Number of Workers	Loss
1	2
2	3

Clearly we should take "Number of Workers" as an input variable and "Loos" as response. It is also quite easy to plot the data:



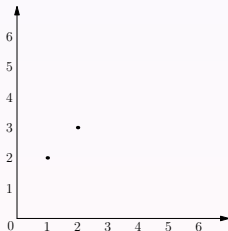
Simple Linear Regression.

Let us start with very simple example.

An area manager in a department store wants to study the relationship between the number of workers on duty at a certain department and the value (in hundreds) of merchandise lost to shoplifters:

Number of Workers	Loss
1	2
2	3

Clearly we should take "Number of Workers" as an input variable and "Loos" as response. It is also quite easy to plot the data:



The data is just two points, so simple regression model is very reasonable and easy here

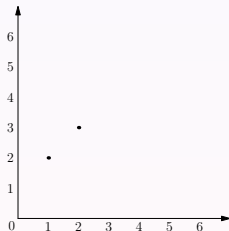
Simple Linear Regression.

Let us start with very simple example.

An area manager in a department store wants to study the relationship between the number of workers on duty at a certain department and the value (in hundreds) of merchandise lost to shoplifters:

Number of Workers	Loss
1	2
2	3

Clearly we should take "Number of Workers" as an input variable and "Loss" as response. It is also quite easy to plot the data:



The data is just two points, so simple regression model is very reasonable and easy here and using standard formula for line through two points we get

$$Y = x + 1.$$

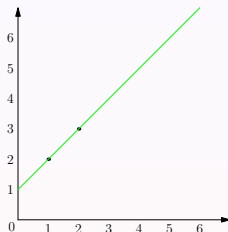
Simple Linear Regression.

Let us start with very simple example.

An area manager in a department store wants to study the relationship between the number of workers on duty at a certain department and the value (in hundreds) of merchandise lost to shoplifters:

Number of Workers	Loss
1	2
2	3

Clearly we should take "Number of Workers" as an input variable and "Loos" as response. It is also quite easy to plot the data in **scatter diagram**:



The data is just two points, so Simple Regression model is very reasonable and easy here and using standard formula for line through two points we get

$$Y = x + 1.$$

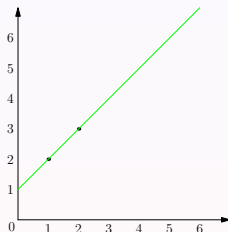
Simple Linear Regression.

Let us start with very simple example.

An area manager in a department store wants to study the relationship between the number of workers on duty at a certain department and the value (in hundreds) of merchandise lost to shoplifters:

Number of Workers	Loss
1	2
2	3

Clearly we should take "Number of Workers" as an input variable and "Loos" as response. It is also quite easy to plot the data in **scatter diagram**:



The data is just two points, so Simple Regression model is very reasonable and easy here and using standard formula for line through two points we get

$$Y = x + 1.$$

Using above we, for example, may predict the value of Y at point $x_t = 5!$

Simple Linear Regression.

A slightly different example:

Simple Linear Regression.

A slightly different example:

An area manager in a department store wants to study the relationship between the number of workers on duty at a certain department and the value (in hundreds) of merchandise lost to shoplifters:

Number of Workers	Loss
1	2
2	3
4	1

Clearly we should take "Number of Workers" as an input variable and "Loss" as response. It is also quite easy to plot the data in **scatter diagram**:

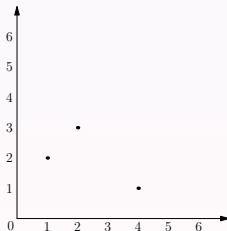
Simple Linear Regression.

A slightly different example:

An area manager in a department store wants to study the relationship between the number of workers on duty at a certain department and the value (in hundreds) of merchandise lost to shoplifters:

Number of Workers	Loss
1	2
2	3
4	1

Clearly we should take "Number of Workers" as an input variable and "Loss" as response. It is also quite easy to plot the data in **scatter diagram**:

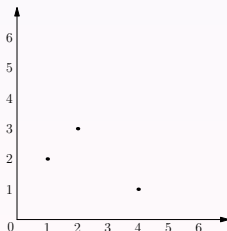


A slightly different example:

An area manager in a department store wants to study the relationship between the number of workers on duty at a certain department and the value (in hundreds) of merchandise lost to shoplifters:

Number of Workers	Loss
1	2
2	3
4	1

Clearly we should take "Number of Workers" as an input variable and "Loss" as response. It is also quite easy to plot the data in **scatter diagram**:



The data now is three points, there is "no line" through this three points!

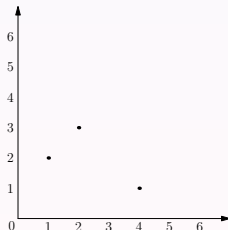
Simple Linear Regression.

A slightly different example:

An area manager in a department store wants to study the relationship between the number of workers on duty at a certain department and the value (in hundreds) of merchandise lost to shoplifters:

Number of Workers	Loss
1	2
2	3
4	1

Clearly we should take "Number of Workers" as an input variable and "Loss" as response. It is also quite easy to plot the data in **scatter diagram**:



The data now is three points, there is "no line" through this three points! But we may ask for the best line which would approximate those points (as well as some "future or previous" values).

Simple Linear Regression a more practical example

A new type of washing machine was recently introduced in 11 department stores. These stores are of roughly equal size and are located in similar types of communities. The manufacturer varied the price charged in each store, and the following data, giving the number of units sold in one month for each of the different prices, resulted

Prices (\$)	Number sold
280	44
290	41
300	34
310	38
320	33

Prices (\$)	Number sold
330	30
340	32
350	26
360	28
370	23
380	20

Simple Linear Regression a more practical example

A new type of washing machine was recently introduced in 11 department stores. These stores are of roughly equal size and are located in similar types of communities. The manufacturer varied the price charged in each store, and the following data, giving the number of units sold in one month for each of the different prices, resulted

Prices (\$)	Number sold
280	44
290	41
300	34
310	38
320	33

Prices (\$)	Number sold
330	30
340	32
350	26
360	28
370	23
380	20

Clearly we again should take Price as an input variable and "Number sold" as response. It is also quite easy to plot the data in **scatter diagram**:

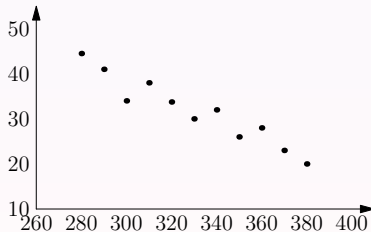
Simple Linear Regression a more practical example

A new type of washing machine was recently introduced in 11 department stores. These stores are of roughly equal size and are located in similar types of communities. The manufacturer varied the price charged in each store, and the following data, giving the number of units sold in one month for each of the different prices, resulted

Prices (\$)	Number sold
280	44
290	41
300	34
310	38
320	33

Prices (\$)	Number sold
330	30
340	32
350	26
360	28
370	23
380	20

Clearly we again should take Price as an input variable and "Number sold" as response. It is also quite easy to plot the data in **scatter diagram**:



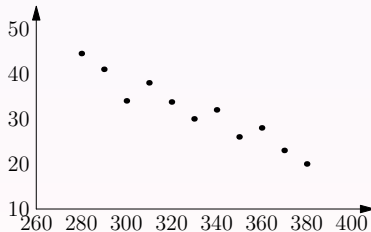
Simple Linear Regression a more practical example

A new type of washing machine was recently introduced in 11 department stores. These stores are of roughly equal size and are located in similar types of communities. The manufacturer varied the price charged in each store, and the following data, giving the number of units sold in one month for each of the different prices, resulted

Prices (\$)	Number sold
280	44
290	41
300	34
310	38
320	33

Prices (\$)	Number sold
330	30
340	32
350	26
360	28
370	23
380	20

Clearly we again should take Price as an input variable and "Number sold" as response. It is also quite easy to plot the data in **scatter diagram**:



The above indicates, that, subject to random error the assumption of straight line relationship between the number of units sold and the price appears to be reasonable!

Theory and Formulas

Suppose that the responses Y_i corresponding to the input values x_i , $i = 1, \dots, n$ are to be observed and used to create/estimate a linear regression,

Suppose that the responses Y_i corresponding to the input values x_i , $i = 1, \dots, n$ are to be observed and used to create/estimate a linear regression, as we discussed on the previous slides usually it will be impossible to find α and β such that $Y_i = \alpha + \beta x_i$

Suppose that the responses Y_i corresponding to the input values x_i , $i = 1, \dots, n$ are to be observed and used to create/estimate a linear regression, as we discussed on the previous slides usually it will be impossible to find α and β such that $Y_i = \alpha + \beta x_i$ thus we must introduce some error e which is the difference of actual value Y and $\alpha + \beta x$:

$$Y = \alpha + \beta x + e.$$

But, we also do not have all data (we have only n values for x and Y)

Suppose that the responses Y_i corresponding to the input values x_i , $i = 1, \dots, n$ are to be observed and used to create/estimate a linear regression, as we discussed on the previous slides usually it will be impossible to find α and β such that $Y_i = \alpha + \beta x_i$ thus we must introduce some error e which is the difference of actual value Y and $\alpha + \beta x$:

$$Y = \alpha + \beta x + e.$$

But, we also do not have all data (we have only n values for x and Y) thus in our calculations we will be able only to find estimates for α and β we will denote those estimators (for now) by A and B .

Suppose that the responses Y_i corresponding to the input values x_i , $i = 1, \dots, n$ are to be observed and used to create/estimate a linear regression, as we discussed on the previous slides usually it will be impossible to find α and β such that $Y_i = \alpha + \beta x_i$ thus we must introduce some error e which is the difference of actual value Y and $\alpha + \beta x$:

$$Y = \alpha + \beta x + e.$$

But, we also do not have all data (we have only n values for x and Y) thus in our calculations we will be able only to find estimates for α and β we will denote those estimators (for now) by A and B . The errors we will two to "minimize" are

$$\varepsilon_i = Y_i - (A + Bx_i),$$

i.e. ε_i represents the error that would result from using estimators A and B to predict the response at input values x_i .

Suppose that the responses Y_i corresponding to the input values x_i , $i = 1, \dots, n$ are to be observed and used to create/estimate a linear regression, as we discussed on the previous slides usually it will be impossible to find α and β such that $Y_i = \alpha + \beta x_i$ thus we must introduce some error e which is the difference of actual value Y and $\alpha + \beta x$:

$$Y = \alpha + \beta x + e.$$

But, we also do not have all data (we have only n values for x and Y) thus in our calculations we will be able only to find estimates for α and β we will denote those estimators (for now) by A and B . The errors we will two to "minimize" are

$$\varepsilon_i = Y_i - (A + Bx_i),$$

i.e. ε_i represents the error that would result from using estimators A and B to predict the response at input values x_i .

To minimize the error we will consider

$$\sum_{i=1}^n \varepsilon_i^2$$

and minimize it over A and B .

Suppose that the responses Y_i corresponding to the input values x_i , $i = 1, \dots, n$ are to be observed and used to create/estimate a linear regression, as we discussed on the previous slides usually it will be impossible to find α and β such that $Y_i = \alpha + \beta x_i$ thus we must introduce some error e which is the difference of actual value Y and $\alpha + \beta x$:

$$Y = \alpha + \beta x + e.$$

But, we also do not have all data (we have only n values for x and Y) thus in our calculations we will be able only to find estimates for α and β we will denote those estimators (for now) by A and B . The errors we will two to "minimize" are

$$\varepsilon_i = Y_i - (A + Bx_i),$$

i.e. ε_i represents the error that would result from using estimators A and B to predict the response at input values x_i .

To minimize the error we will consider

$$\sum_{i=1}^n \varepsilon_i^2$$

and minimize it over A and B . The resulting estimators of α and β are called **least-square estimators**.

Suppose that the responses Y_i corresponding to the input values x_i , $i = 1, \dots, n$ are to be observed and used to create/estimate a linear regression, as we discussed on the previous slides usually it will be impossible to find α and β such that $Y_i = \alpha + \beta x_i$ thus we must introduce some error e which is the difference of actual value Y and $\alpha + \beta x$:

$$Y = \alpha + \beta x + e.$$

But, we also do not have all data (we have only n values for x and Y) thus in our calculations we will be able only to find estimates for α and β we will denote those estimators (for now) by A and B . The errors we will two to "minimize" are

$$\varepsilon_i = Y_i - (A + Bx_i),$$

i.e. ε_i represents the error that would result from using estimators A and B to predict the response at input values x_i .

To minimize the error we will consider

$$\sum_{i=1}^n \varepsilon_i^2$$

and minimize it over A and B . The resulting estimators of α and β are called **least-square estimators**. It is a good question WHY we use squares? Why just not to consider the sum of errors?

Suppose that the responses Y_i corresponding to the input values x_i , $i = 1, \dots, n$ are to be observed and used to create/estimate a linear regression, as we discussed on the previous slides usually it will be impossible to find α and β such that $Y_i = \alpha + \beta x_i$ thus we must introduce some error e which is the difference of actual value Y and $\alpha + \beta x$:

$$Y = \alpha + \beta x + e.$$

But, we also do not have all data (we have only n values for x and Y) thus in our calculations we will be able only to find estimates for α and β we will denote those estimators (for now) by A and B . The errors we will two to "minimize" are

$$\varepsilon_i = Y_i - (A + Bx_i),$$

i.e. ε_i represents the error that would result from using estimators A and B to predict the response at input values x_i .

To minimize the error we will consider

$$\sum_{i=1}^n \varepsilon_i^2$$

and minimize it over A and B . The resulting estimators of α and β are called **least-square estimators**. It is a good question WHY we use squares? Why just not to consider the sum of errors? (it is not good, negative and positive large errors will cancel each other out).

Suppose that the responses Y_i corresponding to the input values x_i , $i = 1, \dots, n$ are to be observed and used to create/estimate a linear regression, as we discussed on the previous slides usually it will be impossible to find α and β such that $Y_i = \alpha + \beta x_i$ thus we must introduce some error e which is the difference of actual value Y and $\alpha + \beta x$:

$$Y = \alpha + \beta x + e.$$

But, we also do not have all data (we have only n values for x and Y) thus in our calculations we will be able only to find estimates for α and β we will denote those estimators (for now) by A and B . The errors we will two to "minimize" are

$$\varepsilon_i = Y_i - (A + Bx_i),$$

i.e. ε_i represents the error that would result from using estimators A and B to predict the response at input values x_i .

To minimize the error we will consider

$$\sum_{i=1}^n \varepsilon_i^2$$

and minimize it over A and B . The resulting estimators of α and β are called **least-square estimators**. It is a good question WHY we use squares? Why just not to consider the sum of errors? (it is not good, negative and positive large errors will cancel each other out). Why not to consider just a sum of absolute values?

Suppose that the responses Y_i corresponding to the input values x_i , $i = 1, \dots, n$ are to be observed and used to create/estimate a linear regression, as we discussed on the previous slides usually it will be impossible to find α and β such that $Y_i = \alpha + \beta x_i$ thus we must introduce some error e which is the difference of actual value Y and $\alpha + \beta x$:

$$Y = \alpha + \beta x + e.$$

But, we also do not have all data (we have only n values for x and Y) thus in our calculations we will be able only to find estimates for α and β we will denote those estimators (for now) by A and B . The errors we will two to "minimize" are

$$\varepsilon_i = Y_i - (A + Bx_i),$$

i.e. ε_i represents the error that would result from using estimators A and B to predict the response at input values x_i .

To minimize the error we will consider

$$\sum_{i=1}^n \varepsilon_i^2$$

and minimize it over A and B . The resulting estimators of α and β are called **least-square estimators**. It is a good question WHY we use squares? Why just not to consider the sum of errors? (it is not good, negative and positive large errors will cancel each other out). Why not to consider just a sum of absolute values? (it is much harder see the next slide).

So our goal is to minimize

$$F(A, B) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - (A + Bx_i))^2$$

over all possible A and B .

So our goal is to minimize

$$F(A, B) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - (A + Bx_i))^2$$

over all possible A and B . The idea is to use calculus. Note that to find the minimum we need to solve

$$\frac{\partial F(A, B)}{\partial A} = \frac{\partial F(A, B)}{\partial B} = 0$$

(we also in general need to check minimality conditions, but this case is simple).

So our goal is to minimize

$$F(A, B) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - (A + Bx_i))^2$$

over all possible A and B . The idea is to use calculus. Note that to find the minimum we need to solve

$$\frac{\partial F(A, B)}{\partial A} = \frac{\partial F(A, B)}{\partial B} = 0$$

(we also in general need to check minimality conditions, but this case is simple). Which is

$$\begin{cases} -2 \sum_{i=1}^n (Y_i - (A + Bx_i)) = 0 \\ -2 \sum_{i=1}^n (Y_i - (A + Bx_i))x_i = 0 \end{cases}$$

So our goal is to minimize

$$F(A, B) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - (A + Bx_i))^2$$

over all possible A and B . The idea is to use calculus. Note that to find the minimum we need to solve

$$\frac{\partial F(A, B)}{\partial A} = \frac{\partial F(A, B)}{\partial B} = 0$$

(we also in general need to check minimality conditions, but this case is simple). Which is

$$\begin{cases} -2 \sum_{i=1}^n (Y_i - (A + Bx_i)) = 0 \\ -2 \sum_{i=1}^n (Y_i - (A + Bx_i))x_i = 0 \end{cases}$$

Simplifying both equation and dividing by n we get

$$\begin{cases} \bar{Y} - A - B\bar{x} = 0 \\ \frac{1}{n} \sum_{i=1}^n (Y_i x_i) - A\bar{x} - \frac{1}{n} B \sum_{i=1}^n x_i^2 = 0 \end{cases}$$

So our goal is to minimize

$$F(A, B) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - (A + Bx_i))^2$$

over all possible A and B . The idea is to use calculus. Note that to find the minimum we need to solve

$$\frac{\partial F(A, B)}{\partial A} = \frac{\partial F(A, B)}{\partial B} = 0$$

(we also in general need to check minimality conditions, but this case is simple). Which is

$$\begin{cases} -2 \sum_{i=1}^n (Y_i - (A + Bx_i)) = 0 \\ -2 \sum_{i=1}^n (Y_i - (A + Bx_i))x_i = 0 \end{cases}$$

Simplifying both equation and dividing by n we get

$$\begin{cases} \bar{Y} - A - B\bar{x} = 0 \\ \frac{1}{n} \sum_{i=1}^n (Y_i x_i) - A\bar{x} - \frac{1}{n} B \sum_{i=1}^n x_i^2 = 0 \end{cases}$$

Now, play with algebra to get that the minimum is

$$B = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \text{ and } A = \bar{Y} - B\bar{x}$$

Estimating the regression parameters formula:

Suppose that the responses Y_i corresponding to the input values x_i , $i = 1, \dots, n$. The line

$$\gamma = \hat{\alpha} + \hat{\beta}x$$

is called the **estimated regression line**: $\hat{\beta}$ is the slope, and $\hat{\alpha}$ is the intercept of this line. Where

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \text{ and } \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$$

Estimating the regression parameters formula:

Suppose that the responses Y_i corresponding to the input values x_i , $i = 1, \dots, n$. The line

$$\gamma = \hat{\alpha} + \hat{\beta}x$$

is called the **estimated regression line**: $\hat{\beta}$ is the slope, and $\hat{\alpha}$ is the intercept of this line. Where

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \text{ and } \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$$

Also the following notation may be used to help with calculations

$$S_{xY} = \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}) \text{ and } S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

Estimating the regression parameters formula:

Suppose that the responses Y_i corresponding to the input values x_i , $i = 1, \dots, n$. The line

$$\gamma = \hat{\alpha} + \hat{\beta}x$$

is called the **estimated regression line**: $\hat{\beta}$ is the slope, and $\hat{\alpha}$ is the intercept of this line. Where

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \text{ and } \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$$

Also the following notation may be used to help with calculations

$$S_{xY} = \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}) \text{ and } S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

Thus

$$\hat{\beta} = \frac{S_{xY}}{S_{xx}} \text{ and } \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}.$$

Estimating the regression parameters formula:

Suppose that the responses Y_i corresponding to the input values x_i , $i = 1, \dots, n$. The line

$$\gamma = \hat{\alpha} + \hat{\beta}x$$

is called the **estimated regression line**: $\hat{\beta}$ is the slope, and $\hat{\alpha}$ is the intercept of this line. Where

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \text{ and } \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$$

Also the following notation may be used to help with calculations

$$S_{xY} = \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}) \text{ and } S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

Thus

$$\hat{\beta} = \frac{S_{xY}}{S_{xx}} \text{ and } \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}.$$

The following (combinatorial) formulas are also useful when computing the coefficients by hand:

$$S_{xY} = \sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y} \text{ and } S_{xx} = \sum_{i=1}^n x_i^2 - n\bar{x}^2$$

Simple Linear Regression: very simple example

An area manager in a department store wants to study the relationship between the number of workers on duty at a certain department and the value (in hundreds) of merchandise lost to shoplifters:

Number of Workers	Loss
1	2
2	3
4	1

An area manager in a department store wants to study the relationship between the number of workers on duty at a certain department and the value (in hundreds) of merchandise lost to shoplifters:

Number of Workers	Loss
1	2
2	3
4	1

Let us, try to find estimated regression line: $\bar{x} = 7/3$ and $\bar{y} = 2$,

An area manager in a department store wants to study the relationship between the number of workers on duty at a certain department and the value (in hundreds) of merchandise lost to shoplifters:

Number of Workers	Loss
1	2
2	3
4	1

Let us, try to find estimated regression line: $\bar{x} = 7/3$ and $\bar{Y} = 2$, moreover,

$$S_{XY} = (1 - 7/3)(2 - 2) + (2 - 7/3)(3 - 2) + (4 - 7/3)(1 - 2) = -2$$

An area manager in a department store wants to study the relationship between the number of workers on duty at a certain department and the value (in hundreds) of merchandise lost to shoplifters:

Number of Workers	Loss
1	2
2	3
4	1

Let us, try to find estimated regression line: $\bar{x} = 7/3$ and $\bar{y} = 2$, moreover,

$$S_{xy} = (1 - 7/3)(2 - 2) + (2 - 7/3)(3 - 2) + (4 - 7/3)(1 - 2) = -2$$

$$S_{xx} = (1 - 7/3)^2 + (2 - 7/3)^2 + (4 - 7/3)^2 \approx 4.7$$

An area manager in a department store wants to study the relationship between the number of workers on duty at a certain department and the value (in hundreds) of merchandise lost to shoplifters:

Number of Workers	Loss
1	2
2	3
4	1

Let us, try to find estimated regression line: $\bar{x} = 7/3$ and $\bar{Y} = 2$, moreover,

$$S_{xY} = (1 - 7/3)(2 - 2) + (2 - 7/3)(3 - 2) + (4 - 7/3)(1 - 2) = -2$$

$$S_{xx} = (1 - 7/3)^2 + (2 - 7/3)^2 + (4 - 7/3)^2 \approx 4.7$$

thus

$$\hat{\beta} = \frac{S_{xY}}{S_{xx}} \approx -0.43 \text{ and } \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x} = 2 - (-0.43)(7/3) \approx 3$$

An area manager in a department store wants to study the relationship between the number of workers on duty at a certain department and the value (in hundreds) of merchandise lost to shoplifters:

Number of Workers	Loss
1	2
2	3
4	1

Let us, try to find estimated regression line: $\bar{x} = 7/3$ and $\bar{Y} = 2$, moreover,

$$S_{xY} = (1 - 7/3)(2 - 2) + (2 - 7/3)(3 - 2) + (4 - 7/3)(1 - 2) = -2$$

$$S_{xx} = (1 - 7/3)^2 + (2 - 7/3)^2 + (4 - 7/3)^2 \approx 4.7$$

thus

$$\hat{\beta} = \frac{S_{xY}}{S_{xx}} \approx -0.43 \text{ and } \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x} = 2 - (-0.43)(7/3) \approx 3$$

So the estimated regression line has the following equation

$$Y = 3 - 0.43x$$

Simple Linear Regression: very simple example

An area manager in a department store wants to study the relationship between the number of workers on duty at a certain department and the value (in hundreds) of merchandise lost to shoplifters:

Number of Workers	Loss
1	2
2	3
4	1

Then the estimated regression line has the following equation

$$Y = 3 - 0.43x$$

