Lesson 6: Two Quantitative Variables

Agenda

We will study

- Two quantitative (statistics) variables and how one variable can be used to predict the other one
- Review mathematics (e.g., linear function) needed in the statistics concept
- Use computing to model the statistics variables: finding the prediction function and the associated errors for a given sample

Two Quantitative Variables

1. Two quantitative variables (Motivation)
2. Scatterplot of These Variables
3. Linear Functions and Their Lines
4. Regression Line and Prediction
5. Best Line to Describe the Relation of The Two Variables

1. Two Quantitative Variables (motivation)

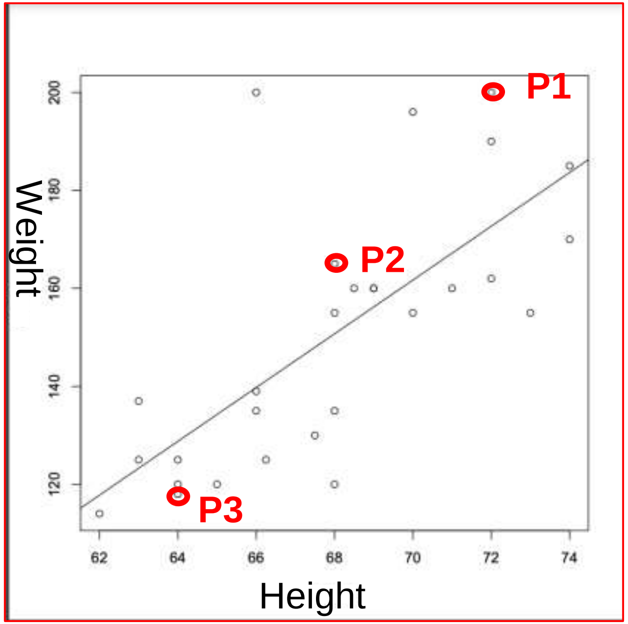
We studied two categorical variables, and now we will study two quantitative variables of a population and how these two variables are correlated.

1. Two Quantitative Variables (motivation)

Example. Consider a *sample* of 28 people from our whole *population* and two variables of the population: *height* and *weight*.
(Partial data of the variables -- table. The graph: plot of all 28 people (each point indicates a person's height and weight)).

Person	Height	Weight
P1	72	200
P2	**68**	**165**
P3	64	118
P4	**66**	135

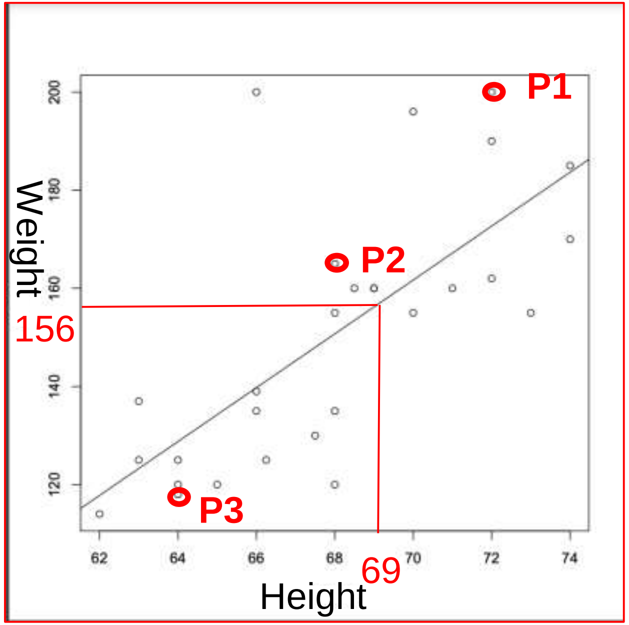
……



1. Two Quantitative Variables (motivation)

Example. Consider a *sample* of 28 people from our whole *population* and two variables of the population: *height* and *weight*.

- Variables *height* and *weight* are **positively correlated**: as the height increases, the weight increases.
- We call *height* the **explanatory** variable, and *weight* the **response** variable.
- *Prediction*: by the trend line (black), a person (outside the sample) of 69 inches could have a weight of around 156 pounds.



1. Two Quantitative Variables (motivation)

From the correlation of the height and weight, we can predict the weight variable using the height variable. The latter (height) is called *the explanatory (or independent) variable* while the former called *response (or dependent) variable*.

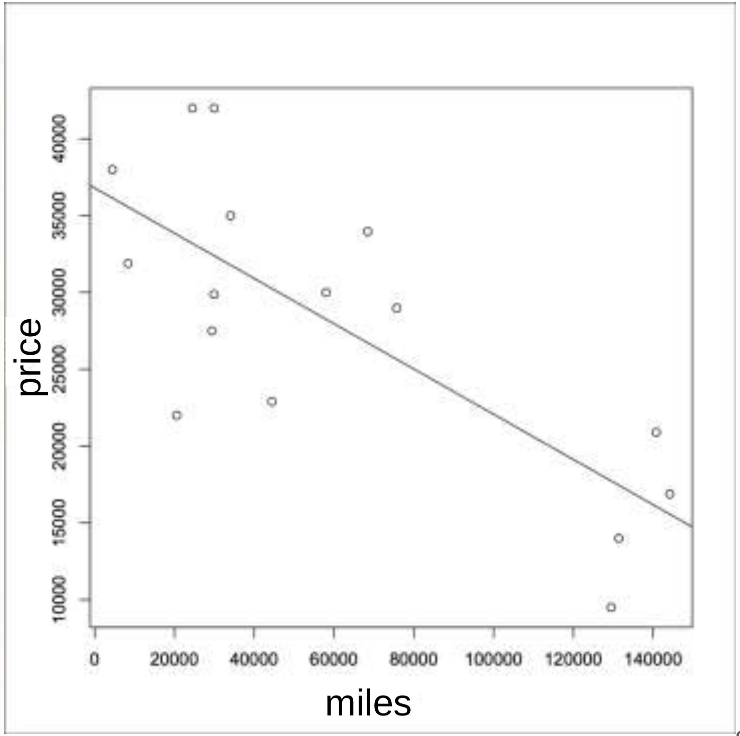
Please also note the use of the concept of *sample* and its relation to *population*.

- A *sample* is a subset of the *population*.
- We know the data of the *sample* only, but not the population.
- From the data of the sample, we can *predict* the data of individuals of the whole population.

1. Two Quantitative Variables (motivation)

Example. Consider the *sample* of 16 used cars which has two variables, price and miles driven. (The population here is all used cars).

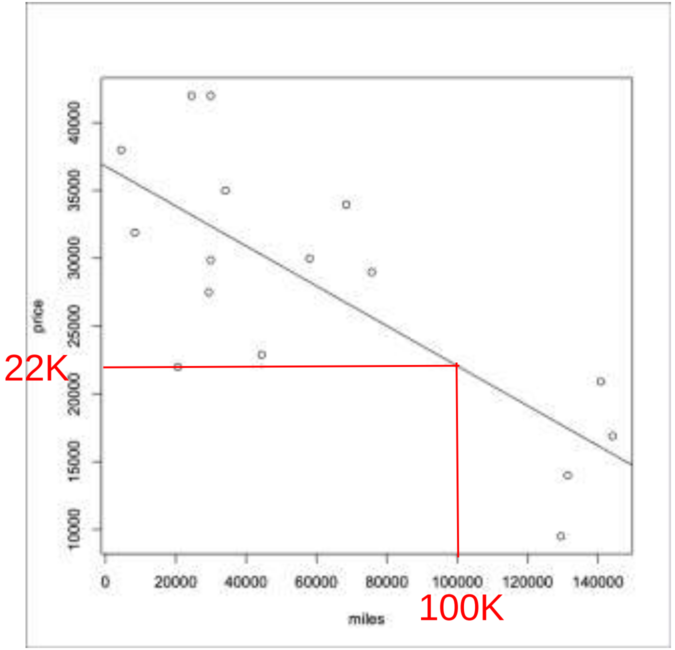
Car	Miles	Price
C1	20583	21994
C2	**12984**	**9500**
C3	29932	29875
C4	**29953**	**41995**
C5	24495	41995
C6	**75678**	**28986**
C7	8359	31891



1. Two Quantitative Variables (motivation)

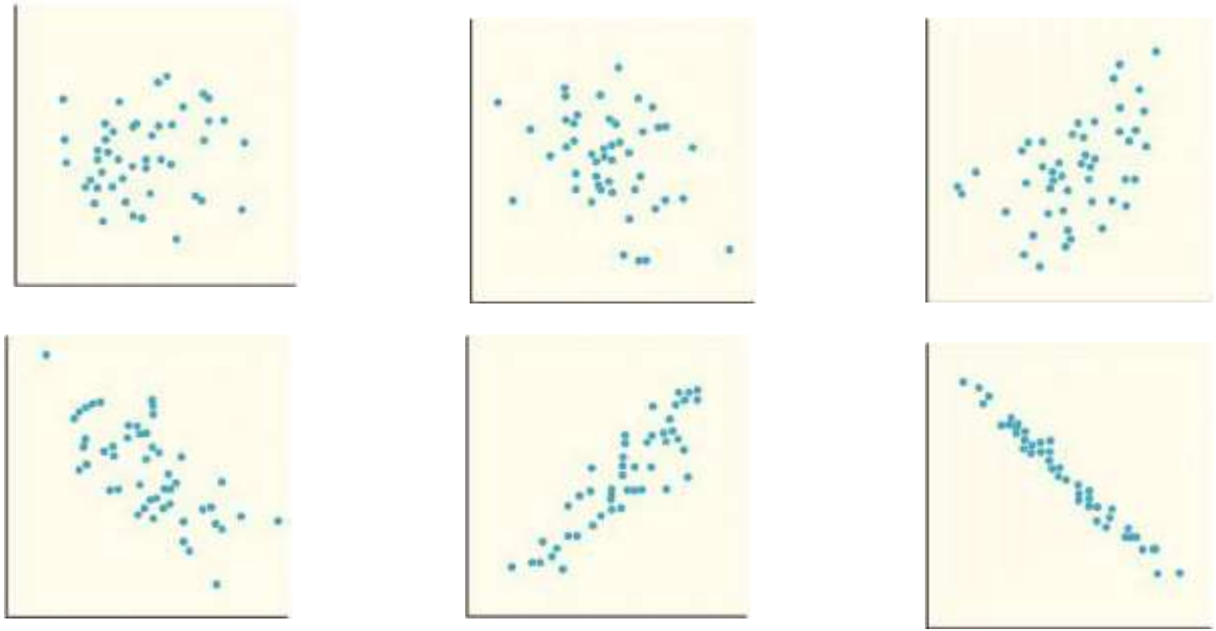
Example. Consider the *sample* of 16 used cars which have two variables, *miles* driven and *price*. (The population here is all used cars).

- The variables *miles* and *price* are **negatively correlated**: as the miles increase, the price decreases.
- Here *miles* is the **explanatory** variable, and *price* the **response** variable.
- *Prediction*: by the trend line, a car (outside the sample) driven 100K miles could have a price of $22K



Two Quantitative Variables (motivation)

Some messier data illustrating different "levels" of correlations:



1. Two Quantitative Variables (motivation)

In the rest of this lesson, we will

- study *scatterplots* of two variables of a sample - a visualization of data (variables) helping us to see the pattern (e.g., how variables are correlated).
- review of *linear functions* and their *graphs* as *lines* (linear function, coordinate systems, drawing basics)
- *predict* data (i.e., values of variables) using *linear function* obtained from data of a sample.
- study “some” *best line*, called *least-squares regression line*, that fits the data.

Two Quantitative Variables

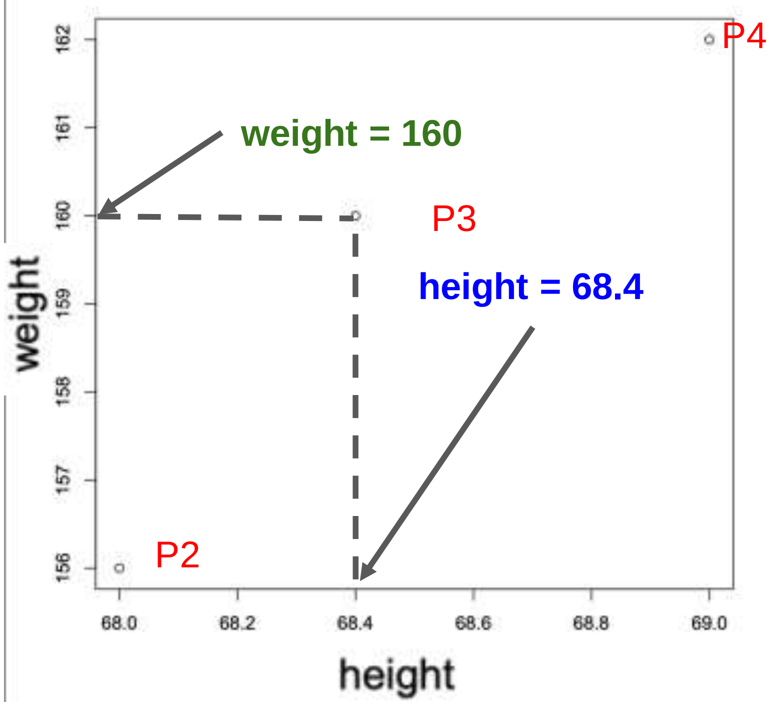
1. Two quantitative variables (Motivation)
2. Scatterplot of These Variables
3. Linear Functions and Their Lines
4. Regression Line and Prediction
5. Best Line to Describe the Relation of The Two Variables

2. Scatterplots of Two Quantitative Variables

Recall the sample of a group of people on their height and weight. To illustrate a *scatterplot* of variables, we consider a smaller sample with P2, P3, & P4.

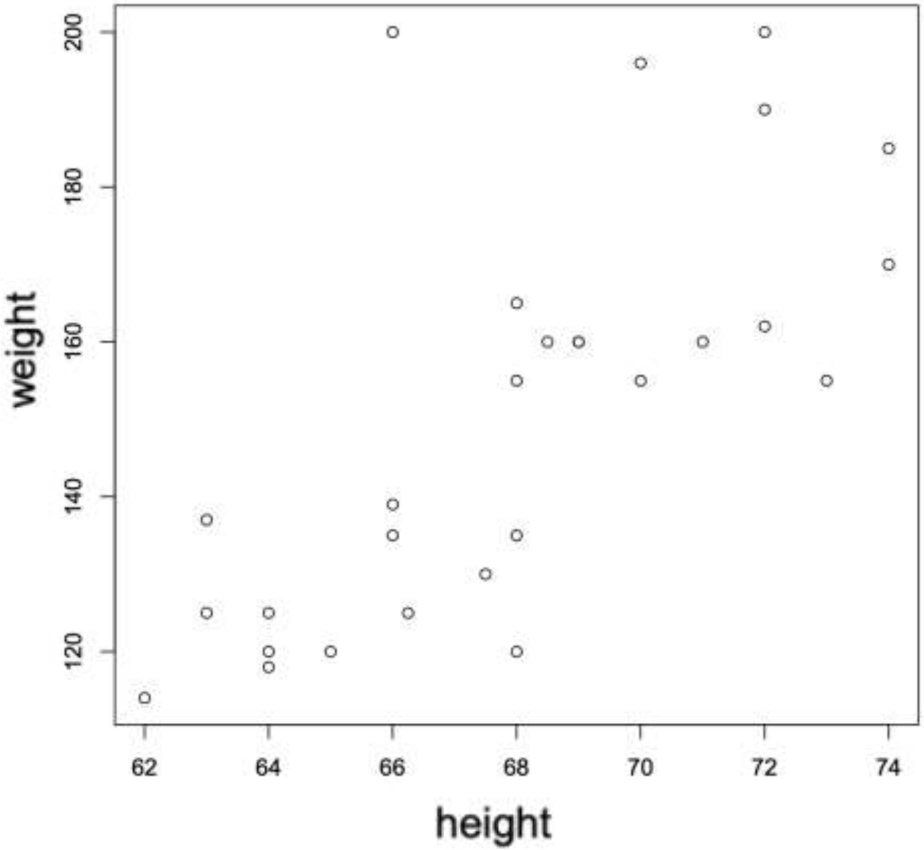
People	height	weight
P2	68	156
P3	**68.4**	**160**
P4	69	162

- *x-axis* is marked by the height values and labeled as "height"
- *y-axis* is marked by the weight values and labeled as "weight"
- For each person (e.g., **P3**), we draw a point in the plot: its *x* coordinate is *height* (**68.4)** and *y* coordinate is *weight* (**160**).



2. Scatterplots of Two Quantitative Variables

The *scatterplot* for the height and weight for the whole *sample* is to the right. Note that each point represents the height and weight of a person. (There could be several people with the same height and weight. In this case, there is still only one point in the plot for all these people.)



2. Scatterplots of Two Quantitative Variables

Let's draw the *scatterplot* using the R function: plot(*v1*, *v2*): its inputs are *v1* and *v2* that are two statistics variables of a sample represented as named vectors; its *side effect* is to draw and display the scatterplot for *v1* and *v2*.

Goto repl.it

- Create a file with name `w-h-plot.r` in your own repl.it
- Click the link: https://replit.com/@WendyStaffen/CStats#L6/w-h-plot.r
- Copy code there into your new file
- Follow instructions at ## T1 in your file, write the R expression there to draw the scatterplot
- (*to be continued* next page)

2. Scatterplots of Two Quantitative Variables

Let's draw the plot using the R function:

plot(*v1*, *v2*): its inputs are *v1* and *v2* that are two named vectors of the same sample; its *side effect* is to draw and display the scatterplot for the sample with values of *v1* forming *x-axis* and values of *v2* forming *y-axis*.

- (*continue* from the previous page)
- (Follows instructions at ## T2 in your file) Go to console, **do NOT run R**. Run your program with R:
  `r w-h-plot.r`
  the plot drawn by your program is "put" in a pdf file: `Rplots.pdf`
- click the file `Rplots.pdf`
- If you need to edit your program, click your file name `w-h-plot.r`
- Repeat the procedure above until you see the correct scatterplot.

Two Quantitative Variables

1. Two quantitative variables (Motivation)
2. Scatterplot of These Variables
3. Linear Functions and Their Lines
4. Regression Line and Prediction
5. Best Line to Describe the Relation of The Two Variables

3. Linear Functions and Their Lines

We will review *linear functions*: how to represent them using *symbolic form* and represent their graph in the *coordinate system*.

Remember, we use a table to represent the content of a function. A linear function usually maps real numbers to a real number. Since we have infinite real numbers, it would be hard to use table to represent such a function. Fortunately, we can use a *symbolic* way to represent a function.

3. Linear Functions and Their Lines

Motivation of Linear Function. Assume Peter works for a research project with an hourly rate of $5 per hour. The income of Peter from the project is a function of the number of hours he works. Assume the function name is *income*, we use

$$income(x) = 5 * x \qquad \text{--- (1)}$$

to represent the function from hours to income. Recall we use the expression *income(x)* to represent the income of working *x* hours. The statement (1) is **read** as the output (or the value) of function *income* with input *x*, or the output of *income(x)*, is 5 * *x*. For example, for *x* being 2 hours, $income(2) = 5 * 2 = 10$.

The function *income* is a *linear* function.

3. Linear Functions and Their Lines

Can we write Peter's income function as

$$income(y) = 5 * y?$$

$$income(hours) = 5*hours?$$

$$income(x) = 5 * hours?$$

3. Linear Functions and Their Lines

Can we write Peter's income function as

*income(y) = 5 * y?* Yes!

*income(hours) = 5*hours?* Yes!

*income(x) = 5 * hours?* No!

Key: We can name the input using any variables, and the output of the function has to be based on the input variables!

- So, the first two are good and the second is preferred because the input variable name gives a good idea.
- But the 3rd one is not correct, because the input is named as *x*, but the output is using *hours* and no one knows what *hours* means.

3. Linear Functions and Their Lines

Motivation of Linear Function. Now we extend the example of Peter. Since Peter needs to perform his research in a university, the project team also pays Peter $10 for his travel cost to school every day. Now, the income of Peter for a day would be $10 plus the income for hours he works on. Assume the function name is *tIncome*, using a symbolic representation, the function *tIncome* is:

$$tIncome(x) = 10 + 5 * x \quad \text{--- (2)}$$

Statement (2) is **read** as the output (or the value) of function *tIncome* with input *x*, or the output of *income(x)*, is *10+5 * x*. It is also read as *given x*, the *tIncome* is *10 + 5 * x*.

The function *tIncome* is a *linear* function.

3. Linear Functions and Their Lines

Definition (Linear Function). A **linear function** with name f is of the form

$f(x) = a + b*x$ (or we omit "*" in the expression: $f(x) = a + bx$)

where a and b refers to some real numbers, and x is a variable.

Example.

$income(x) = 5*x$ is a linear function where $a = 0$ and $b = 5$.

$tIncome(x) = 10 + 5*x$ is a linear function where $a = 10$ and $b = 5$.

Note x in the expression above represents the input of the function, and thus its value can change all the time. However, a and b are fixed.

3. Linear Functions and Their Lines

Graph Representation of Functions. You have learned that a function can be represented by a graph. To represent it as a graph:

First we need a *coordinate system*: *x-axis*, *y-axis* and values on these axis'.

Second, the input and corresponding output of the function can be taken as a pair, and thus the function is the set of all such pairs. For example, the symbolic representation $income(x) = 5 * x$ means the set of pairs $\{(1, 5), (2, 10), \ldots\}$.

Third, each pair can be drawn as a point in the coordinate system

Finally, all the pairs of the function *income(...)* form a line

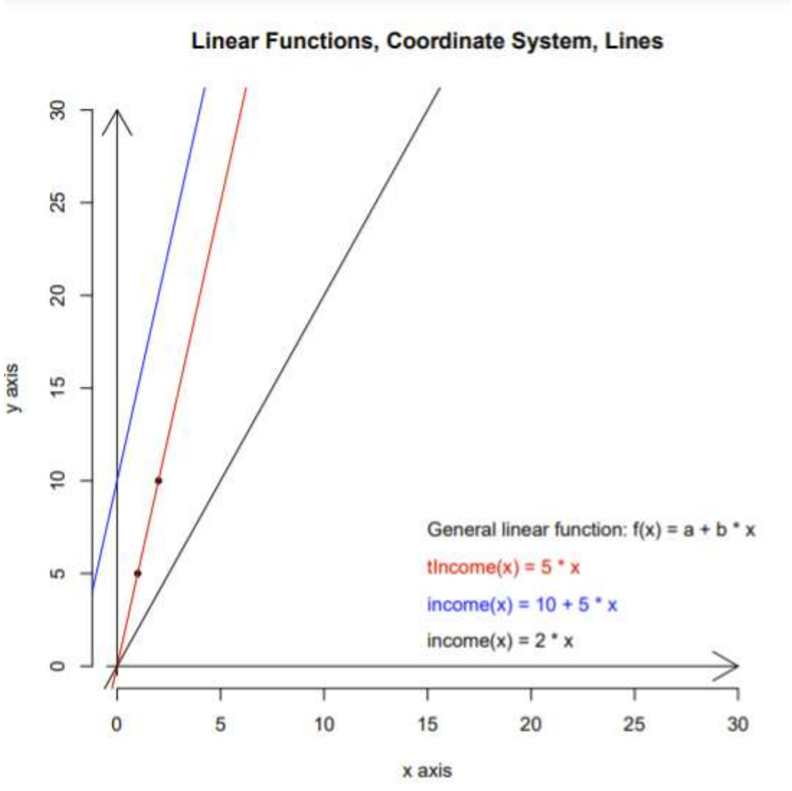
3. Linear Functions and Their Lines

Graph Representation of Functions.

In the right figure, the graph represent of each of the following functions is a line

- *tincome(x) = 5 * x*
- *Income(x) = 10 + 5*x*

In the graph of any function *f(x)* = *a* + *bx* is a line. Hence, *f(x)* = *a* + *bx* is called *linear function*.



3. Linear Functions and Their Lines

Linear Function: *y*-intercept, slope.

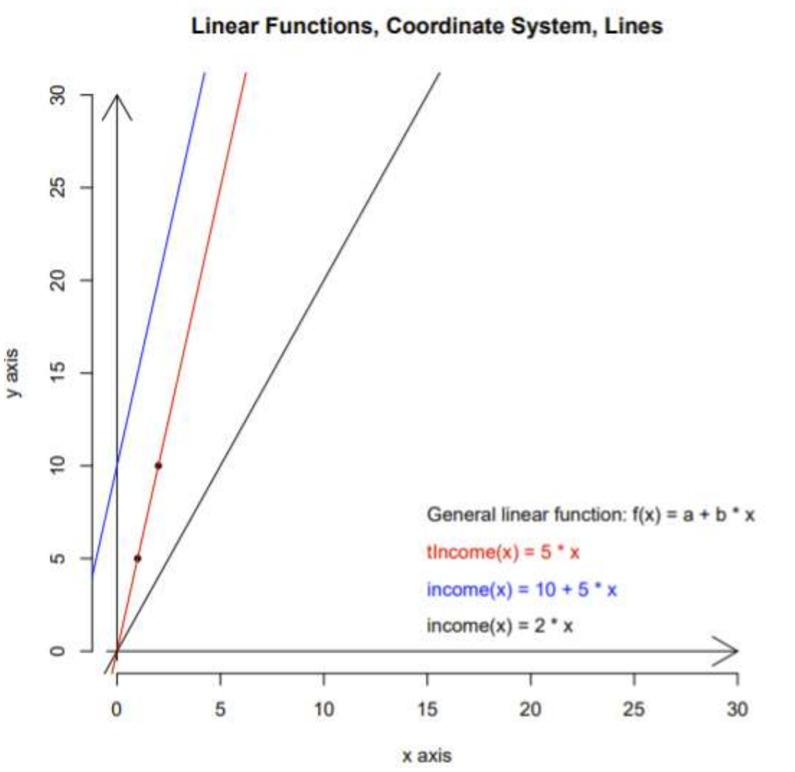
Recall a linear function *f* is of the form

$$f(x) = a + bx.$$

a is called the **<u>y-intercept</u>** of *f*, *b* is called the **<u>slope</u>** of *f*.

y-intercept is where the line intersects *y-axis.*
slope is the "slope" of the line.

- When it is positive, the larger it is, the steeper the line (the red line with *b =5* is steeper than the balck line with *b = 2)*



3. Linear Functions and Their Lines

Linear Function: *y*-intercept, slope.

Recall a linear function *f* is of the form

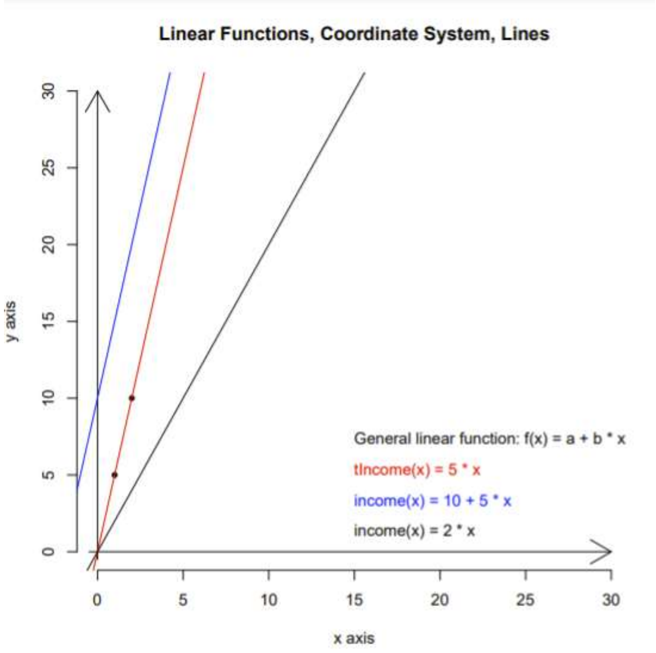
$$f(x) = a + bx$$

a is called **y-intercept** of *f*, *b* is called **slope** of *f*.

Note: slope is the number multiplying the input variable

What is the *y-intercept* and *slope* of each of the following linear functions?

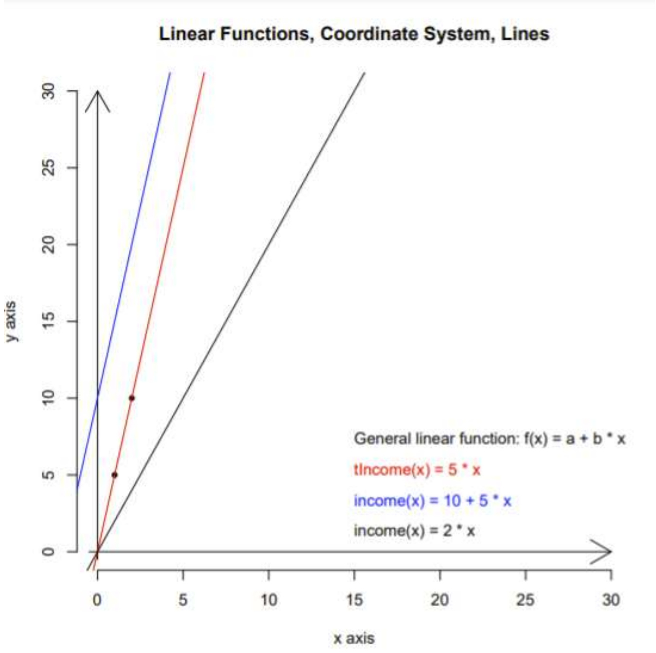
- $f_1(x) = 2x + 3$
- $f_2(x) = -2x - 10$
- $f_3(age) = 10 - 3*age$
- $f_4(height) = 3*height - 8$



3. Linear Functions and Their Lines

What is the *y-intercept* and *slope* of each of the following linear functions?

	slope	y-intercept
• $f_1(x) = 2x + 3$		2
3		
• $f_2(x) = -2x - 10$		-2
-10		
• $f_3(age) = 10 - 3*age$		-3
10		



3. Linear Functions and Their Lines

Linear Function: *y*-intercept, slope.

Recall a linear function f is of the form

$$f(x) = a + bx$$

a is called **<u>y-intercept</u>** of f, b is called **<u>slope</u>** of f.

Write the linear function for each of the following pair of *y-intercept* and *slope*?

- y-intercept: -10 and slope: -2
- y-intercept: 3 and slope: 2
- y-intercept: 10 and slope: 3
- y-intercept: 10 and slope: -3

3. Linear Functions and Their Lines

Linear Function: *y*-intercept, slope.

Recall a linear function f is of the form

$$f(x) = a + bx$$

a is called **y-intercept** of f, b is called **slope** of f.

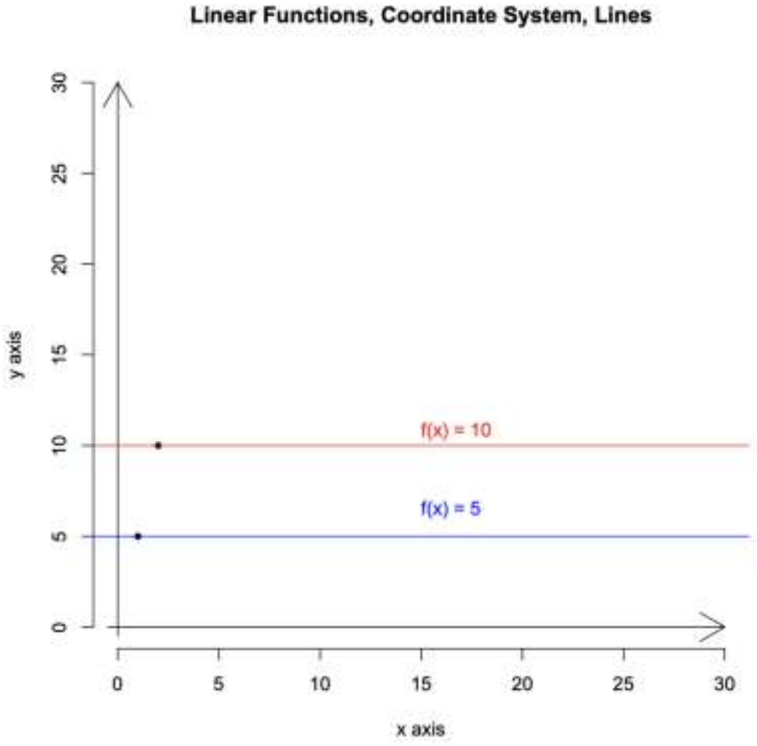
Write the linear function for each of the following pair of *y-intercept* and *slope*:

- y-intercept: -10 and slope: -2 . The function is $f(x) = -10 - 2x$
- y-intercept: 3 and slope: 2 . The function is $f_1(x) = 2x + 3$
- y-intercept: 10 and slope: 3 . The function is $f_3(height) = 10 + 3*height$
- y-intercept: 10 and slope: -3 . The function is $f_4(age) = 10 - 3*age$

3. Linear Functions and Their Lines

A Special Linear Function:

- $f(x) = a$ (when b=0 in $a + bx$)
- The drawing of this function is a *horizontal* line



How to read the function $f(x) = a$?

3. Linear Functions and Their Lines

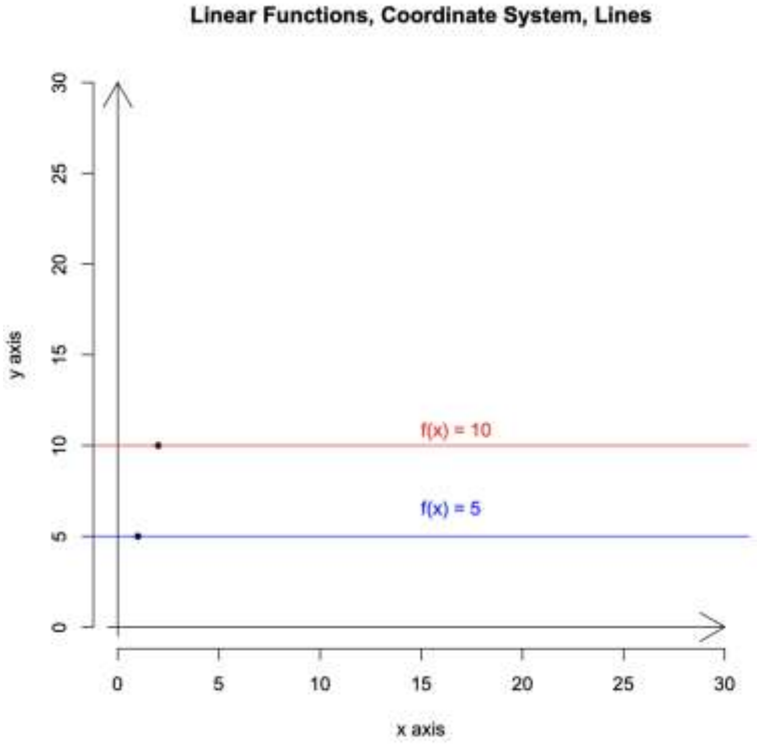
A Special Linear Function:

- $f(x) = a$ (when b=0 in $a + bx$)
- The drawing of this function is a *horizontal* line

How to read the function $f(x) = a$?

The output of the function f with input x is a.

(note the output is not relevant to input x).



3. Linear Functions and Their Lines

A Special Linear Function:

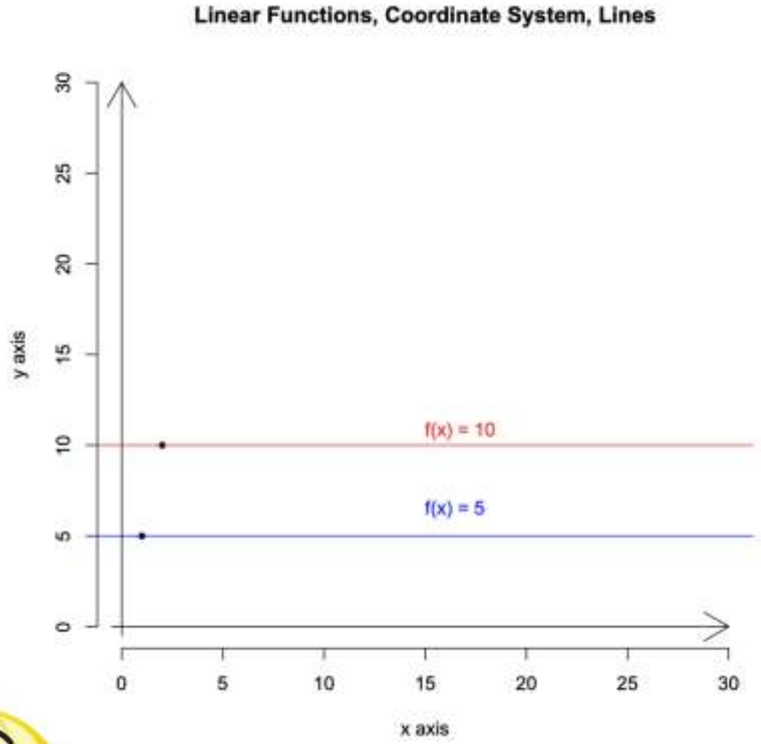
- $f(x) = a$
- The drawing of this function is a *horizontal* line

Example.

Assume we have a linear function:

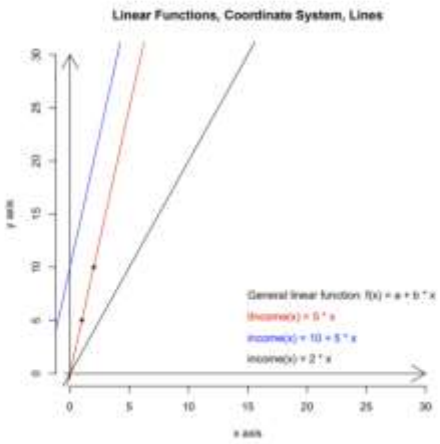
$f(x) = 10$

What is the value of $f(100)$? $f(5)$?



3. Linear Functions and Their Lines

Draw Linear Functions. R provides a function `abline(...)` to draw the line for a linear function



inputs of function `abline`

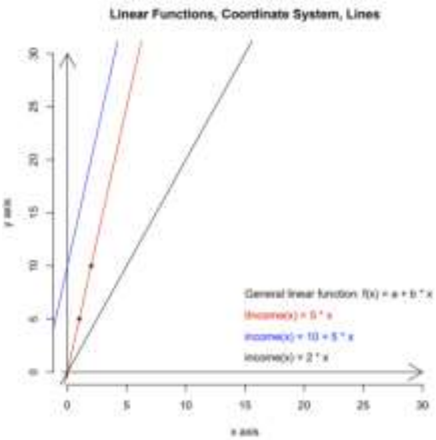
- *a*: the y-intercept of the function to be drawn
- *b*: the slope of the linear function to be drawn
- *col*: the color to use in the drawing

output: not interesting

side effect: a line for the linear function $f(x) = a + bx$ will be drawn in color *col*.

We will study a new way to use/call a function: associate an argument to a parameter using =

3. Linear Functions and Their Lines



inputs of function **abline**

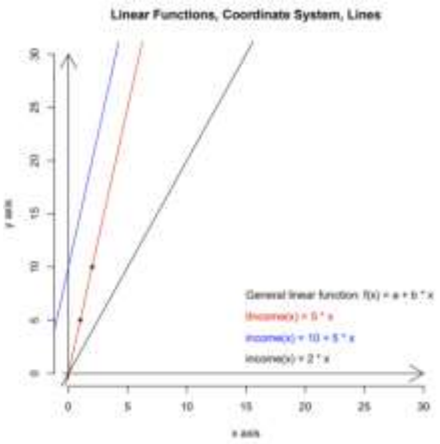
- *a*: the y-intercept of the function to be drawn
- *b*: the slope of the linear function to be drawn
- *col*: the color to use in the drawing. (this is optional)

output: not interesting

side effect: a line for the linear function $f(x) = a + bx$ will be drawn in color *col*.

A new way to use/call a function: *associate an argument to a parameter using* =. To draw line $income(x) = 10 + 5x$, we know 10 is the *y-intercept* and 5 *the slope*, and thus the arguments to **abline**: a = 10 and b = 5. So, R expression `abline(a = 10, b = 5)` will do. Since we give the parameter names, the argument order does not matter now. So `abline(b = 5, a = 10)` will draw the same line.

3. Linear Functions and Their Lines



inputs of function `abline`

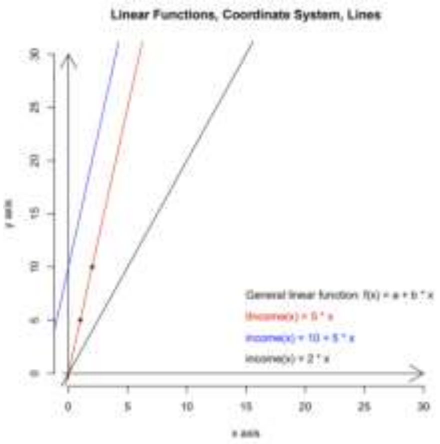
- *a*: the y-intercept of the function to be drawn
- *b*: the slope of the linear function to be drawn
- *col*: the color to use in the drawing. (this is optional)

output: not interesting

side effect: a line for the linear function $f(x) = a + bx$ will be drawn in color *col*.

If you want to draw $income(x) = 10 + 5x$ using blue color, you have to use *col* parameter. *col* has to be a vector of color names as values. The R expression to draw the line is:

3. Linear Functions and Their Lines



inputs of function **`abline`**

- *a*: the y-intercept of the function to be drawn
- *b*: the slope of the linear function to be drawn
- *col*: the color to use in the drawing. (this is optional)

output: not interesting

side effect: a line for the linear function $f(x) = a + bx$ will be drawn in color *col*.

If you want to draw $income(x) = 10 + 5x$ using blue color, you have to use *col* parameter. *col* has to be a vector of color names as values. The R expression to draw the line is:

```
abline(a = 10, b = 5, col = c(“blue”))
```

3. Linear Functions and Their Lines

Practice: Draw Lines for Linear Functions

To draw a graph, we need the preparation (e.g., start plotting, drawing the *x-axis* and *y-axis* of the coordinate system) and finally the drawing. By reading through the sample drawing program, you will see all details.

- click **`L6-drawLines.r`** at link https://replit.com/@yuanlinzhangTTU/L6-Lecture-drawLines#L6-drawLines.r
- Follow instruction behind ## T in file `L6-drawLines.r`

(Tasks inside the program:

```
## T4
##   1) type (in your own file) R expression(s) using abline(...) in lecture notes,
##      to draw the linear function f(x) = 8 - 3x using red color
##      then run your program and see the drawing from the pdf file.
##   2) type in your own file R expression(s) to draw a linear function
##      with slope of 2 and y-intercept 12
```

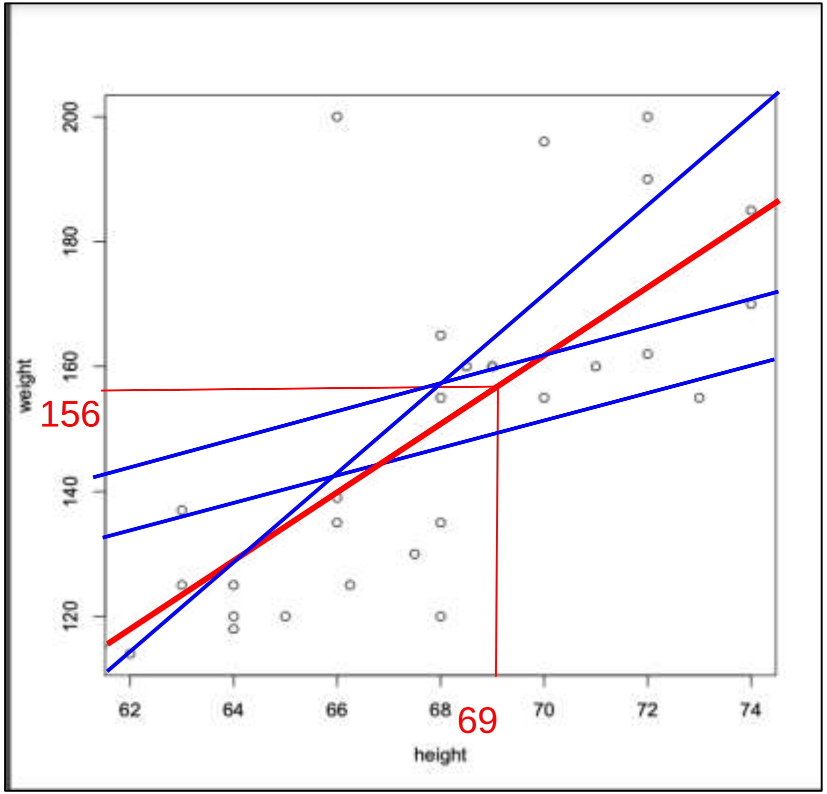
)

Two Quantitative Variables

1. Two quantitative variables (Motivation)
2. Scatterplot of These Variables
3. Linear Functions and Their Lines
4. Regression Line and Prediction
5. Best Line to Describe the Relation of The Two Variables

4. Regression Line and Prediction

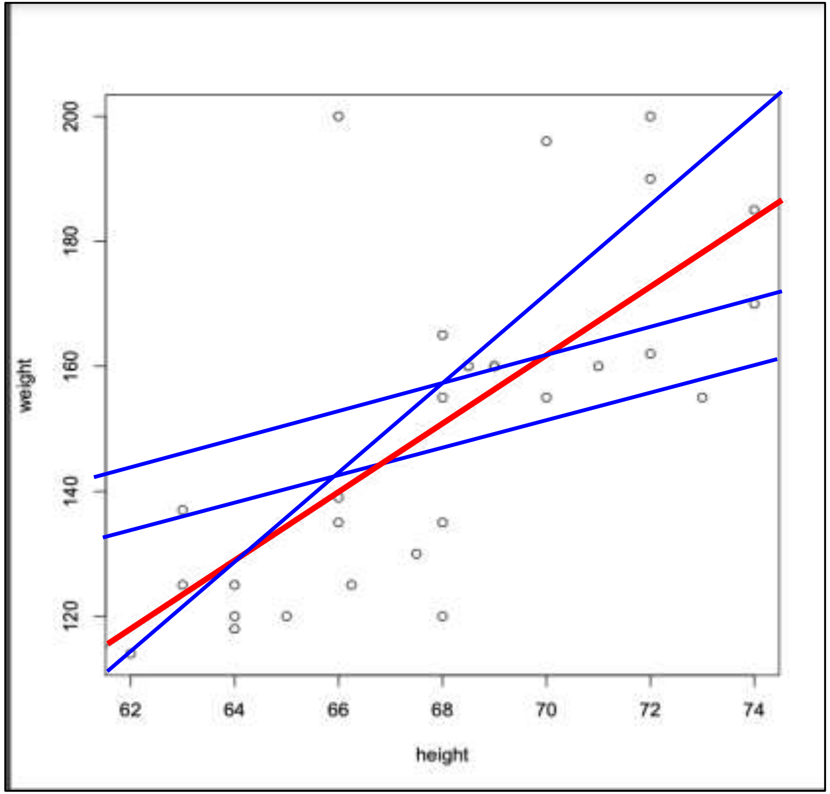
Motivation: Given the two variables of a sufficiently large sample, if they are strongly correlated, we would like to find the best line to "fit" the data or to relate the two variable so that we can use the line to predict the **dependent** variable from the **explainable** variable. But there are so many possibly lines (see blue lines in the height/weight variables, the red line was the one we used before), which one do we choose?



4. Regression Line and Prediction

Definition (Regression Line). Any possible line we can draw in the scatterplot for two variables is called a **<u>regression line</u>**.

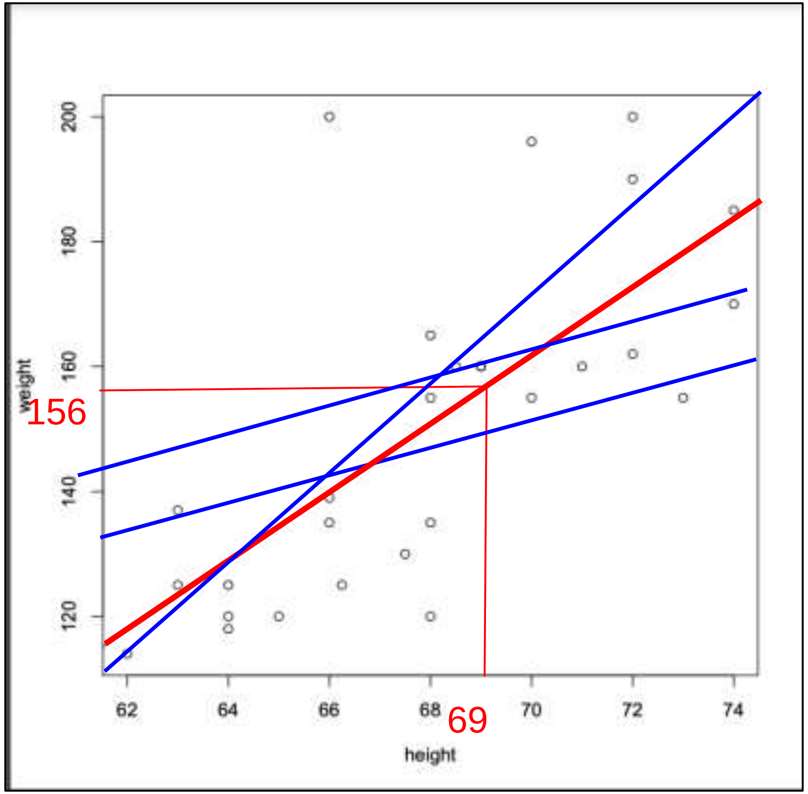
Example. Any line in the right picture is a regression line.



4. Regression Line and Prediction

Prediction.
Any regression line for a *explanatory* variable and a response variable can **predict** the value of the *response* variable from a value of the explanatory variable.

Example. Using the red regression line, given the value of 69 inches of the height variable, we can predict the value of the weight variable: 156.

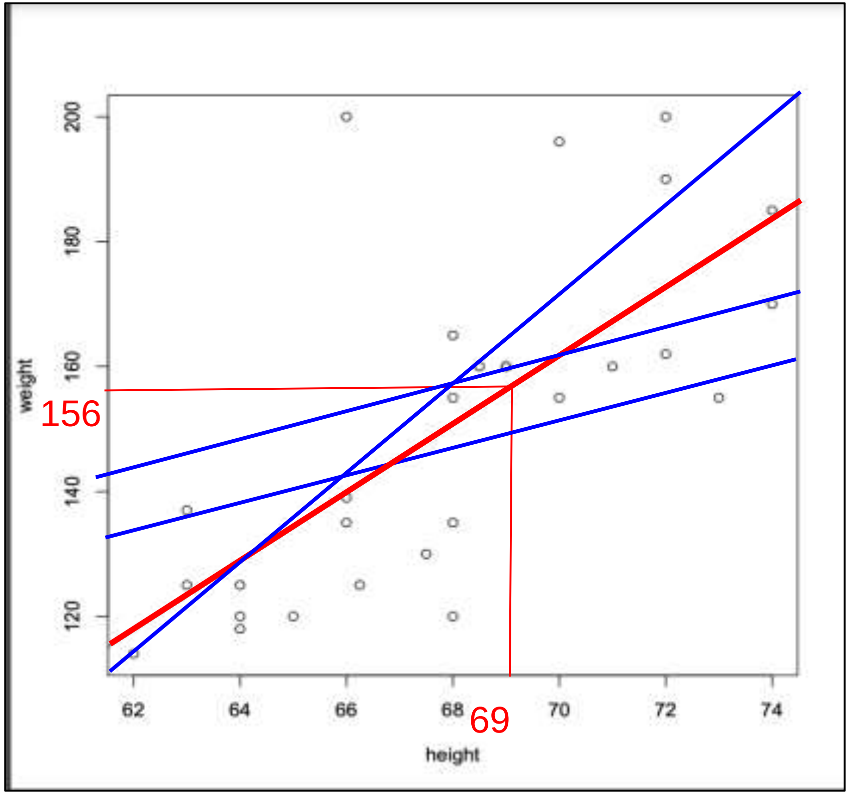


4. Regression Line and Prediction

Prediction.

Example. In fact, the *slope* of the red regression line (for the height and weight variables) is 5.488204 and its *y-intercept* is -222.479443.

Write the linear function for the red line:



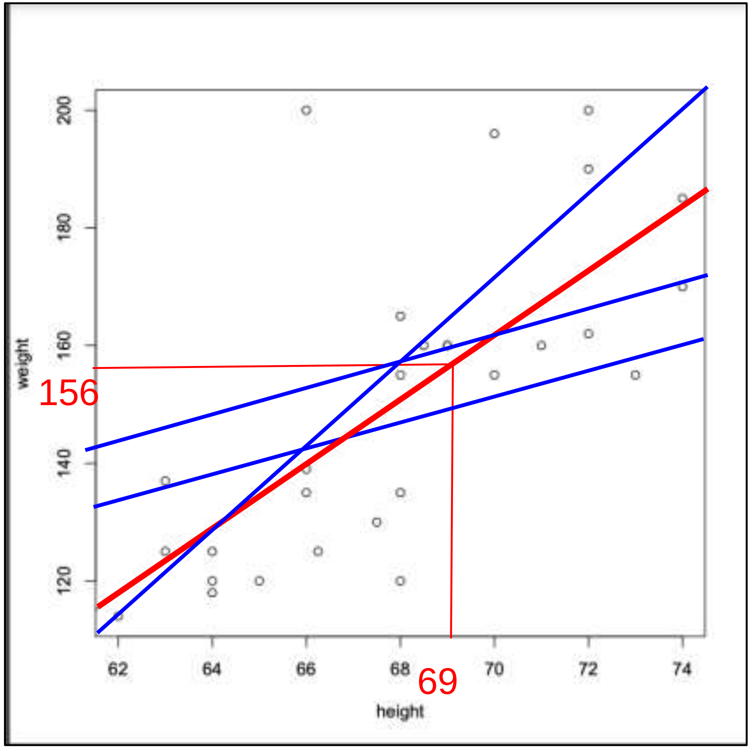
4. Regression Line and Prediction

Predication.

Example. In fact, the *slope* of the red regression line (for the height and weight variables) is 5.488204 and its *y-intercept* is -222.479443. Write the linear function for a line with that slope and y-intercept:

*weight_predict(x) = -222.479443 + 5.488204*x*

For a height of 71.5 inches, write an R-expression to find the value predicted by the regression line:



4. Regression Line and Prediction

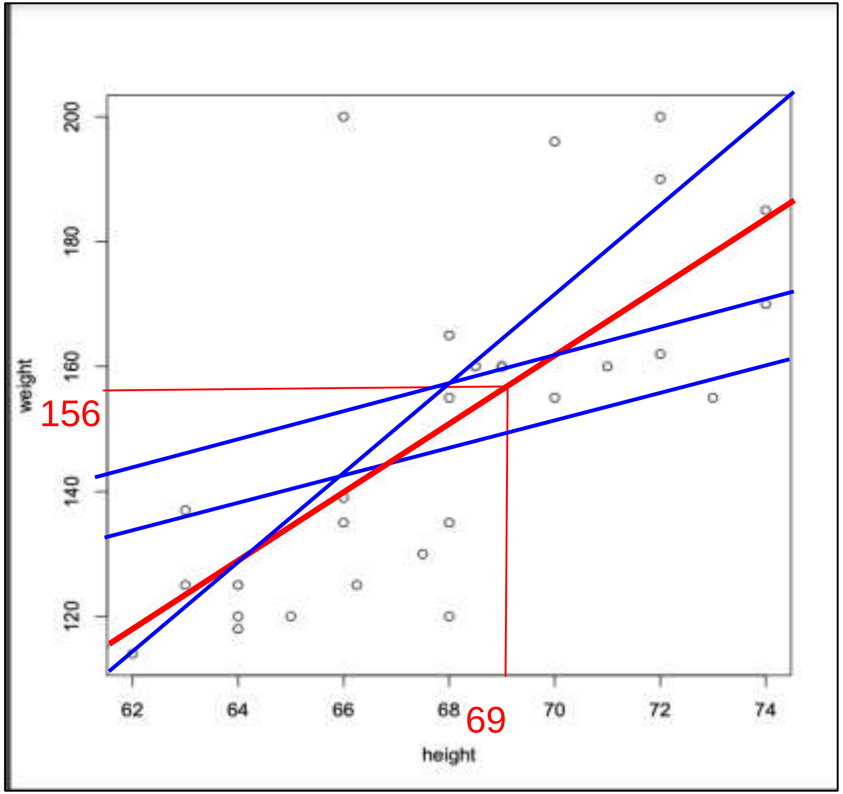
Predication.

Example. Write the corresponding linear function to predict weight?

*weigh_predict(x) = -222.479443 + 5.488204*x*

For a height of 71.5 inches, write an R-expression to find the value predicted by the regression line:

```
-222.479443 + 5.488204*71.5
```



Assignment

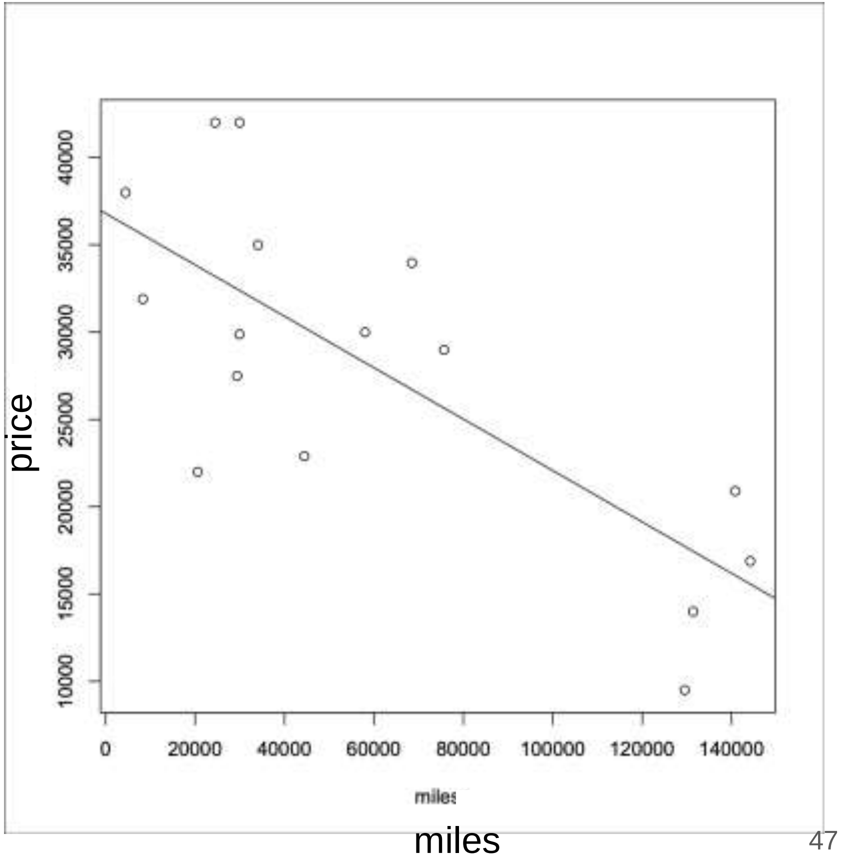
4. Regression Line and Prediction

Prediction (checking understanding)

Example. Recall the example of used cars and variable *miles* (explanatory) and *price* (response). The *y-intercept* of a regression line of these variables is *36792.5971*, and the slope of the regression line is *-0.1472*.

1) Write the linear function for the regression line:

1) Write R-expression to find the predicted price for a car with 58000 driven miles:



4. Regression Line and Prediction

Prediction (Programming)

Example. Recall the example of used cars and variable *miles* (explanatory) and *price* (response). The *y-intercept* of a regression line of these variables is *36792.5971*, and the slope of the regression line is *-0.1472*.

Write an R function to predict the price of a used car using its driven miles. Test your function.

- Recall method of writing first the intentional form of the function (function name, input and output of the function).
- Recall the method of translating your form into an R function skeleton, and complete the R function.
- Finally, test your function.

Review

4. Regression Line and Prediction

Predication (checking understanding)

Example. The *y-intercept* of a regression line for the *miles* (explanatory) variable and *price* (response) variables of used cars is *36792.5971*, and the slope of the regression line is *-0.1472*.

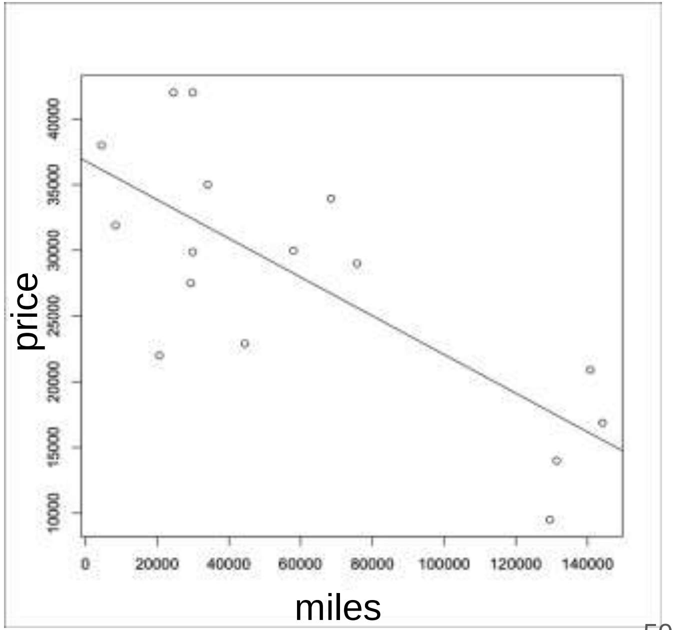
1) Write the corresponding linear function to predict the price of used car:

   *price_predict(m) = 36792.5971 - 0.1472 * m*

   note the function name and input name can be any names.
   But good names are preferred

1) Write R-expression to find the predicted price for a car with 58000 driven miles:

   36792.5971 - 0.1472 * 58000



4. Regression Line and Prediction

Prediction (Programming)

Example. The *y-intercept* of a regression line for the *miles* (explanatory) variable and *price* (response) variables of used cars is *36792.5971*, and the slope of the regression line is *-0.1472*. *Write an R function to predict the price of a used car using the driven miles of the car. Test your function.*

- What is the intentional form of the function?

Function name: price
input
m: miles driven
output
p: the predicted price of the car using using the given intercept 36792.5971 and slope *-0.1472*.

4. Regression Line and Prediction

Prediction (Programming)

Example. The *y-intercept* of a regression line for the *miles* (explanatory) variable and *price* (response) variables of used cars is *36792.5971*, and the slope of the regression line is -*0.1472*. *Write an R function to predict the price of a used car using the driven miles of the car. Test your function.*

- What is the intentional form of the function?
- Write R function in terms of your intentional form and test your function.

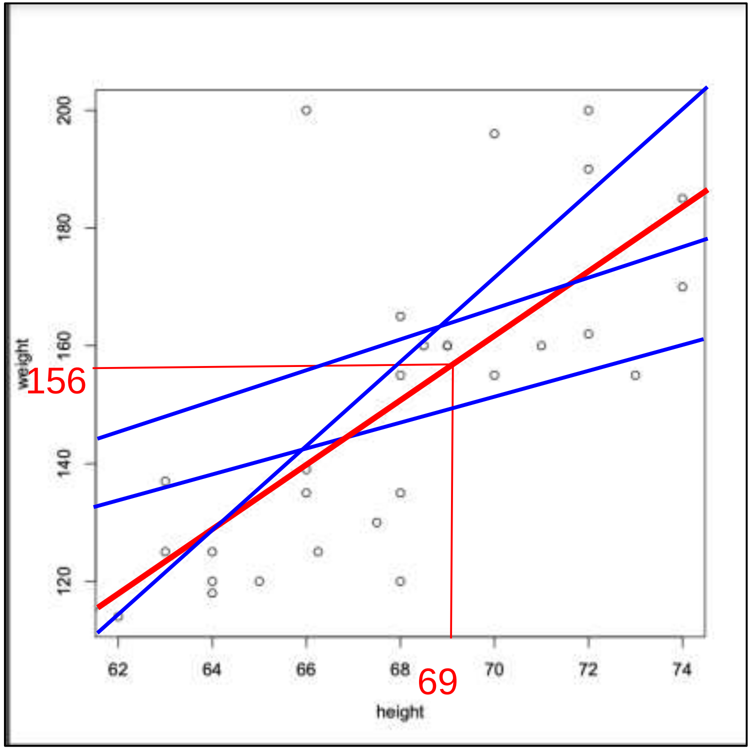
```
# price function
price <- function(m){
  # input
  #   m: miles driven
  # output
  #   p: the predicted price of
  #      the car using using
  #      given intercept and slope

  # get predicted price
   p <- 36792.5971 - 0.1472* m
   return(p) #output p
}
# copy and paste function to R console

# test the function in R console
# using driven miles of 100000
price(100000)
```

Two Quantitative Variables

1. Two quantitative variables (Motivation)
2. Scatterplot of These Variables
3. Linear Functions and Their Lines
4. Regression Line and Prediction
5. Best Line to Describe the Relation of The Two Variables

5. Best Line to Describe the Relation of The Two Variables

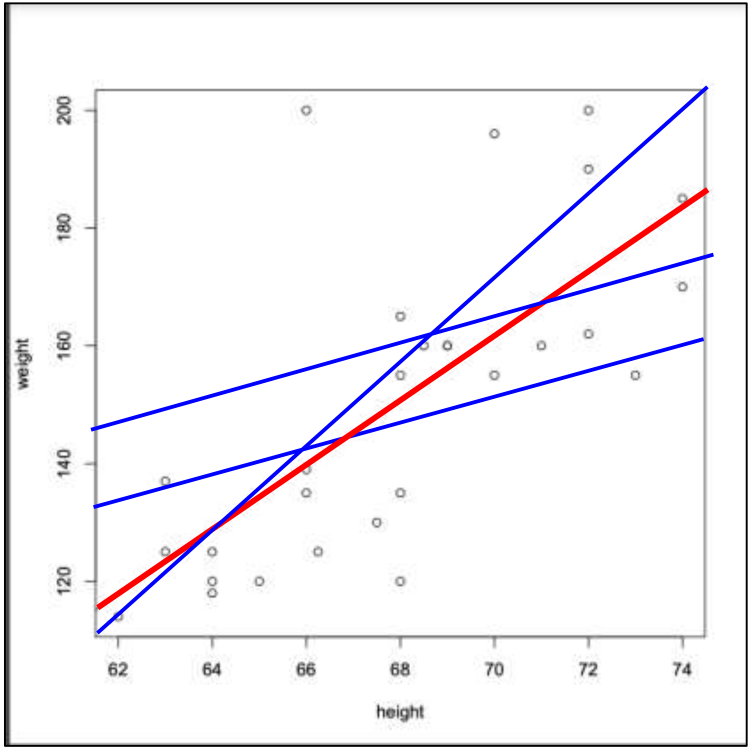
Motivation: As we have seen that regression line is useful in prediction. However, there are so many possible regression lines (see blue lines in the height/weight variables, the red line was the one we used before), which one do we choose?



5. Best Line to Describe the Relation of The Two Variables

Error of Regression Lines. As we can see from the diagram, for any regression line, it will miss some data points. We call that *error*. To figure out the error of a *regression line* for a sample, the idea is to find

- the error for each point
- summarize all errors of all points to form the error for the whole regression line.



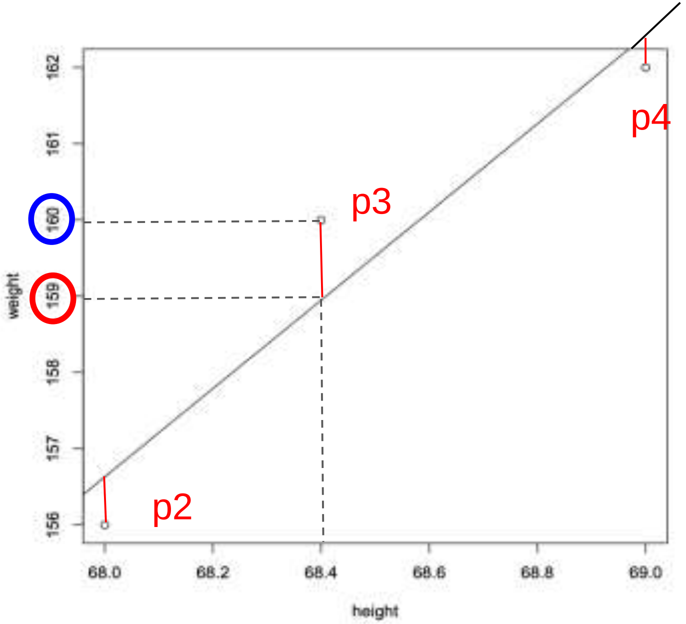
5. Best Line to Describe the Relation of The Two Variables

Error of Regression Lines for a Point.
Recall the height (explanatory) and weight variables of a sample. We define *the error* of a regression line *for a point* (from the sample) as the prediction error using height *h* of the point:

weight of the point - predicted weight using *h*

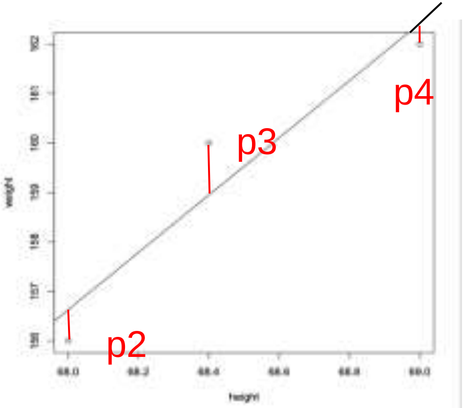
For example, for p3 in the picture, its height is 68.4, and the weight is 160. The predicted weight is 159. The error is 160-159 = 1.

This error is officially called *residual*.



People	height	weight
P3	68.4	160
......		

5. Best Line to Describe the Relation of The Two Variables



Error of Regression Lines for a Point.

Definition (Residual) Given two statistics variable x and y, let a sample individual has value s for x and value t for y. The **residual** of a regression line f on point (s, t) is

t - the predicted value by f using s,

i.e., $t - f(s)$.

5. Best Line to Describe the Relation of The Two Variables

Error of Regression Lines for a Point.

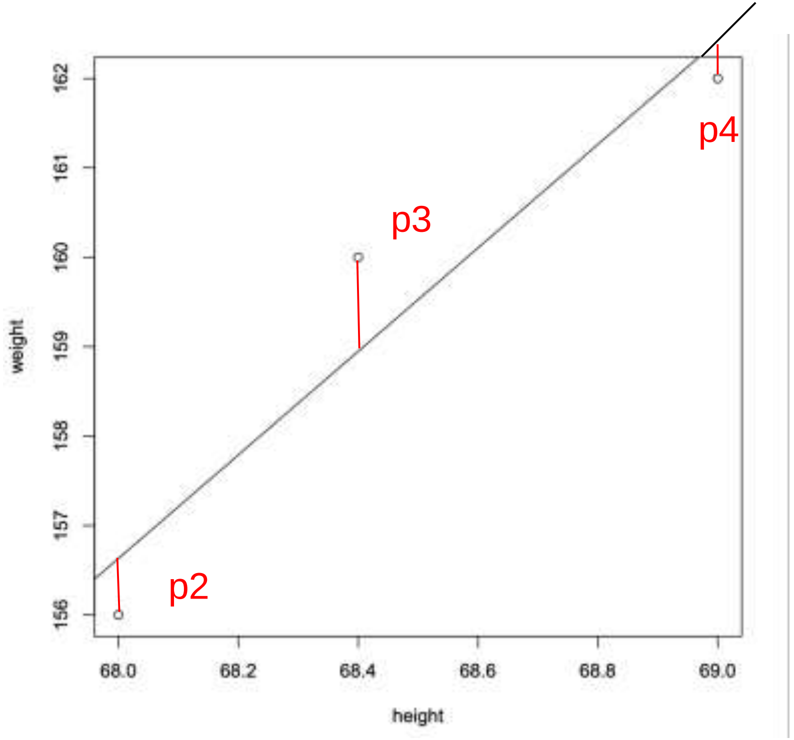
For sample in table below. Let regression line (from explanatory variable height to response variable weight) be *f*, fill expressions in the empty space below.

People	height	weight	predicted weight	residual
P2	68	156	$f(68)$	$156 - f(68)$
P3	**68.4**	**160**	**?**	**?**
P4	69	162	?	?

5. Best Line to Describe the Relation of The Two Variables

Error. (continued)

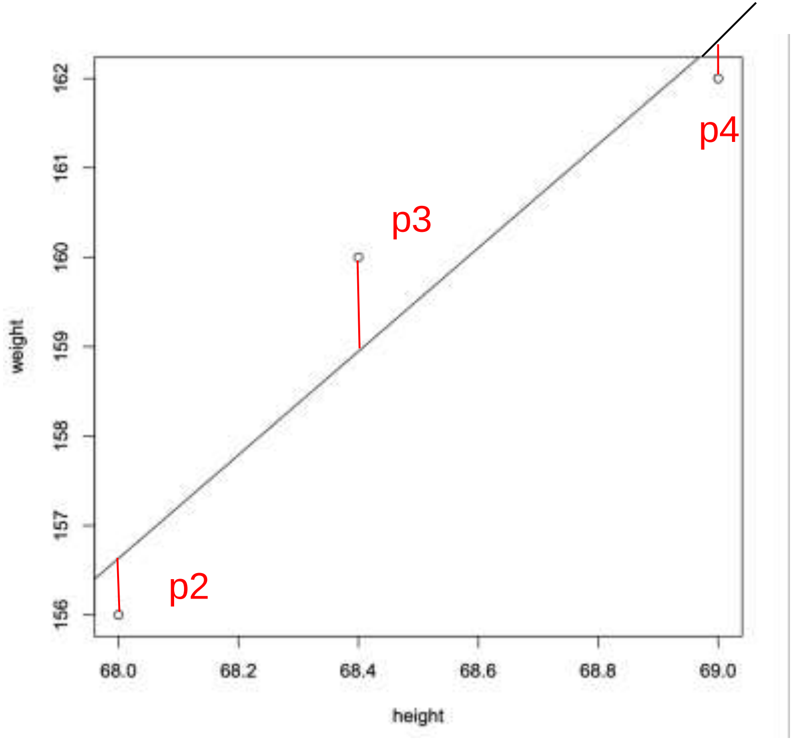
- How to summarize the errors (into one number) for all points?



5. Best Line to Describe the Relation of The Two Variables

Error. (continued)

- How to summarize the errors (into one number) for all points? If we simply sum all the residuals, some positive and negative residuals will cancel each other out. To avoid the cancelation, we square the residuals and sum the squares!



5. Best Line to Describe the Relation of The Two Variables

Define residuals using vectors:

Consider an explanatory variable *x* and a dependant variable *y* of a sample, and a regression line $f(h) = a + bh$ for these variables. Remember *x* and *y* can be represented as named vectors. Write an expression whose value is the vector of the **<u>residuals</u>** of the regression line for all individuals of the sample:

5. Best Line to Describe the Relation of The Two Variables

Define residuals using vectors:

Consider an explanatory variable x and a dependant variable y of a sample, and a regression line $f(h) = a + bh$ for these variables.. Remember x and y can be represented as named vectors. Write an expression whose value is the vector of the **residuals** of the regression line for all individuals of the sample:

```
r <− y - (a + b*x)
```

Write an expression whose value is the sum of the square of the residuals:

5. Best Line to Describe the Relation of The Two Variables

Define residuals using vectors:

Consider an explanatory variable *x* and a dependant variable *y* of a sample, and a regression line *f(h) = a + bh* for these variables.. Remember *x* and *y* can be represented as named vectors. Write an expression whose value is the vector of the **residuals** of the regression line for all individuals of the sample:

*r <− y - (a + b*x)*

Write an expression whose value is the sum of the square of the residuals:

*sum(r*r)*

5. Best Line to Describe the Relation of The Two Variables

Definition (Least-squares Regression Line). Given a explanatory variable *x* and a dependant variable *y* of a sample, **the least-squares regression line** for the variables is the line for which *the sum of the squared residuals* is the minimal.

5. Best Line to Describe the Relation of The Two Variables

R provides a function `lm(...)` to find the *the least-squares regression line* for an explanatory variable and a response variable.

- **input**
  y~x: y is the *response* variable and x is the *explanatory* variable
- **output**:
  an “object” containing the intercept and slope of the least squares regression line for y and x.

  Note `abline`(...) can directly accept this “object” as an argument and draw the line using the intercept and slope inside the “object”.

Example. Write an R expression to draw the least squares regression line for an explanatory variable x and response variable y:

5. Best Line to Describe the Relation of The Two Variables

R provides a function `lm(...)` to find the *the least-squares regression line* for an explanatory variable and a response variable.

- input
  y~*x*: *y* is the response variable and *x* the *explanatory* variable
- output: an "object" containing the *y-intercept* and *slope* of the *least squares regression line* for *y* and *x*.

Note abline(...) can directly accept the "object" and draw the line using the intercept and slope inside the "object"

Example. Write an R expression to draw the least squares regression line for an explanatory variable *x* and response variable *y*:

```
abline( lm(y ~ x))
```

5. Best Line to Describe the Relation of The Two Variables

Practice. Drawing scatterplot and least squares regression line.

Follow instructions in the file **`milesPriceUsedCars.r`** at link https://replit.com/@yuanlinzhangTTU/L6-usedCarsDemo#milesPriceUsedCars.r to draw the *scatterplot* and *least squares regression line* for the *miles* (explanatory variable) and *price* variables of a sample of used cars.

5. Best Line to Describe the Relation of The Two Variables

Other measures of errors of regression lines.

Definition (Standard deviation of the residuals) Given an explanatory variable *x* and a dependant variable *y* of a sample of size *n*, and a regression line, **<u>the standard deviation of the residuals</u>**, usually denoted by **s,** is the *sqrt*(*S / (n-2))* where *S* is the *sum of squared residuals* of the regression line.

Assignment

5. Best Line to Describe the Relation of The Two Variables

Practice (Programming).

Consider the *payroll* and *wins* of baseball teams. Follow the instructions in `baseBallTeams.r` at link https://replit.com/@yuanlinzhangTTU/baseBallTeam#baseBallTeams.r

Team	Payroll	Wins
Arizona Diamondbacks	103	69
Atlanta Braves	**122**	**68**
Baltimore Orioles	157	89
Boston Red Sox	**215**	93
Chicago Cubs	182	103
Chicago White Sox	**141**	**79**
Cincinnati Reds	114	**68**
Cleveland Indians	**114**	**94**

Review

5. Best Line to Describe the Relation of The Two Variables

Practice. Review

Consider the *payroll* and *wins* of baseball teams. Follow the instructions in `baseBallReview.r` at link https://replit.com/@yuanlinzhangTTU/L6-baseBallTeamsReview#baseBallReview.r

Team	Payroll	Wins
Arizona Diamondbacks	103	69
Atlanta Braves	**122**	**68**
Baltimore Orioles	157	89
Boston Red Sox	**215**	93
Chicago Cubs	182	103
Chicago White Sox	**141**	**79**
Cincinnati Reds	114	**68**
Cleveland Indians	**114**	**94**

Summary

- Two quantitative statistics variables
- Scatterplot of these variables
- Linear Functions and Their Lines
- Regression Line and Prediction
- Least squares regression line to fit the data (i.e., two variables)
- Errors of regression line: sum of squared residuals and standard deviation of residuals
- Computing: using R expressions to represent the information above