The Beginning

TIPS FOR SERVICE

Let's assume that you are a small restaurant owner or a very business minded server / waiter at a nice restaurant. Here in the U.S. "tips" are a very important part of a waiter's pay. Most of the time the dollar amount of the tip is related to the dollar amount of the total bill.

TIPS FOR SERVICE

Let's assume that you are a small restaurant owner or a very business minded server / waiter at a nice restaurant. Here in the U.S. "tips" are a very important part of a waiter's pay. Most of the time the dollar amount of the tip is related to the dollar amount of the total bill.

As the waiter or owner, you would like to develop a model that will allow you to make a prediction about what amount of tip to expect for any given bill amount. Therefore one evening, you collect data for six meals.

TIPS FOR SERVICE

Unfortunately when you begin to look at your data, you realize you only collected data for the tip amount and not the meal amount also! So this is the best data you have.

Meal #	Tip amount ($)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00

TIPS FOR SERVICE

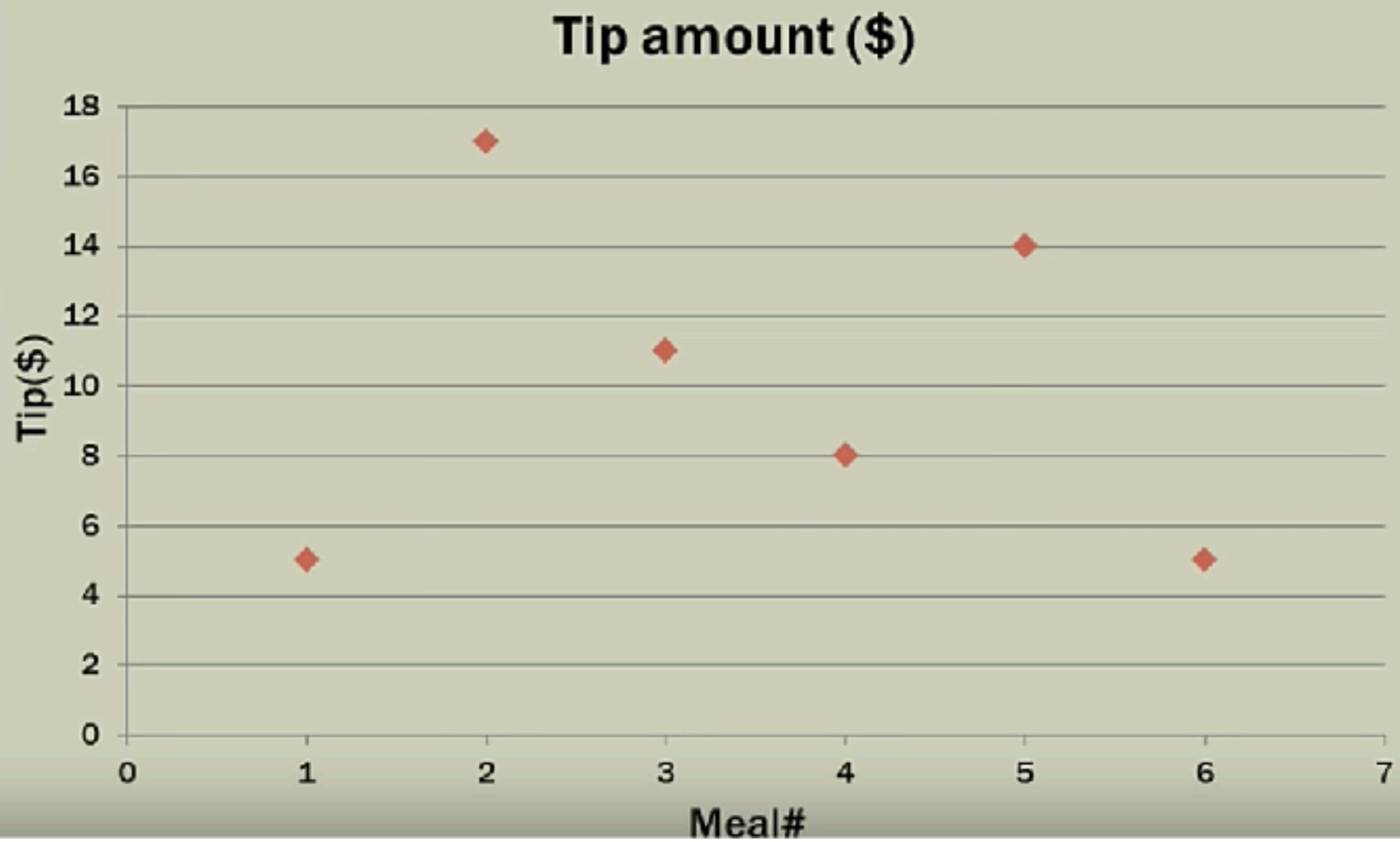
Unfortunately when you begin to look at your data, you realize you only collected data for the tip amount and not the meal amount also! So this is the best data you have.

How might you predict the tip amount for future meals using only this data?

Meal #	Tip amount ($)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00

TIPS FOR SERVICE

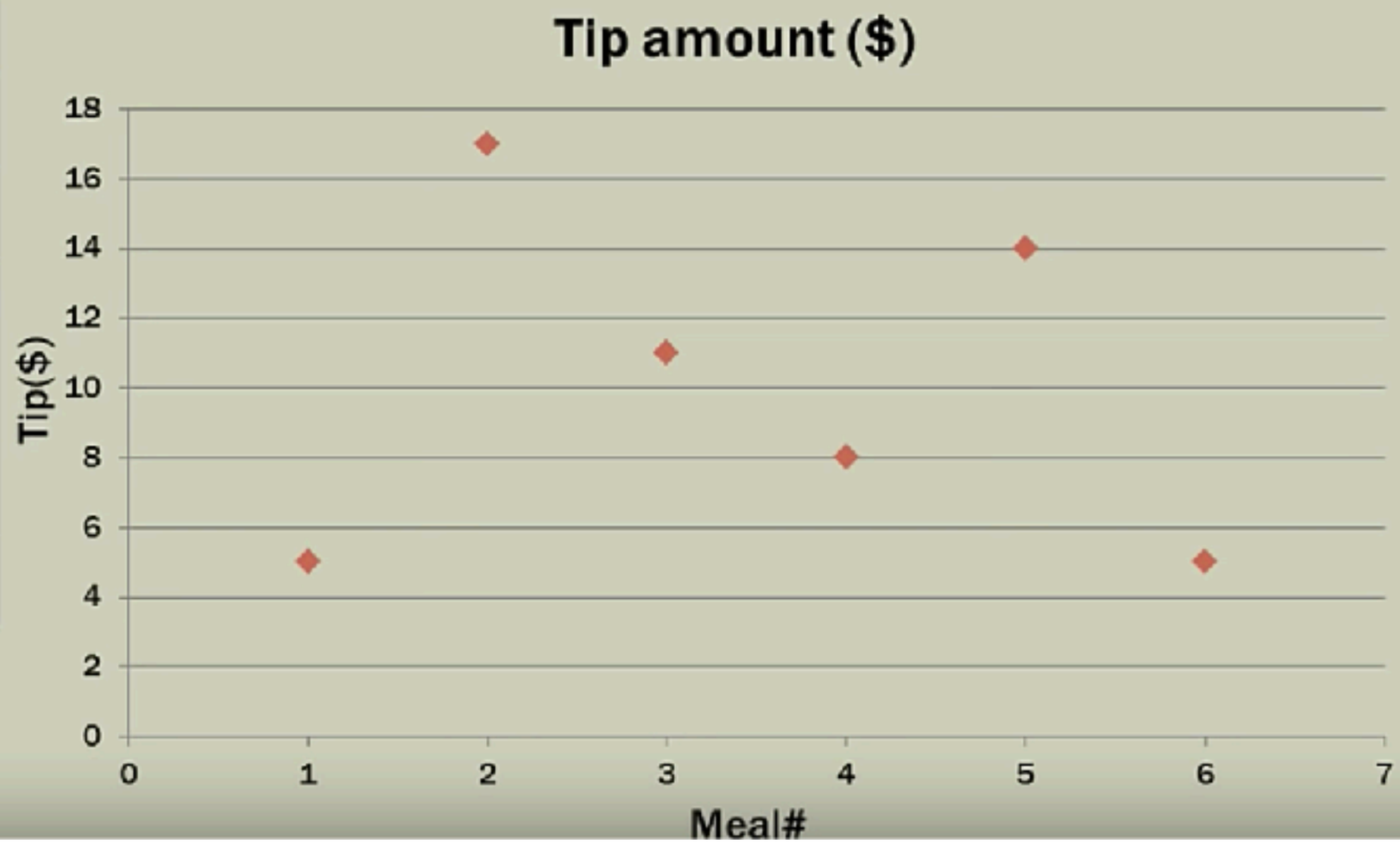
Meal#	Tip amount ($)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00



TIPS FOR SERVICE

Meal#	Tip amount ($)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00

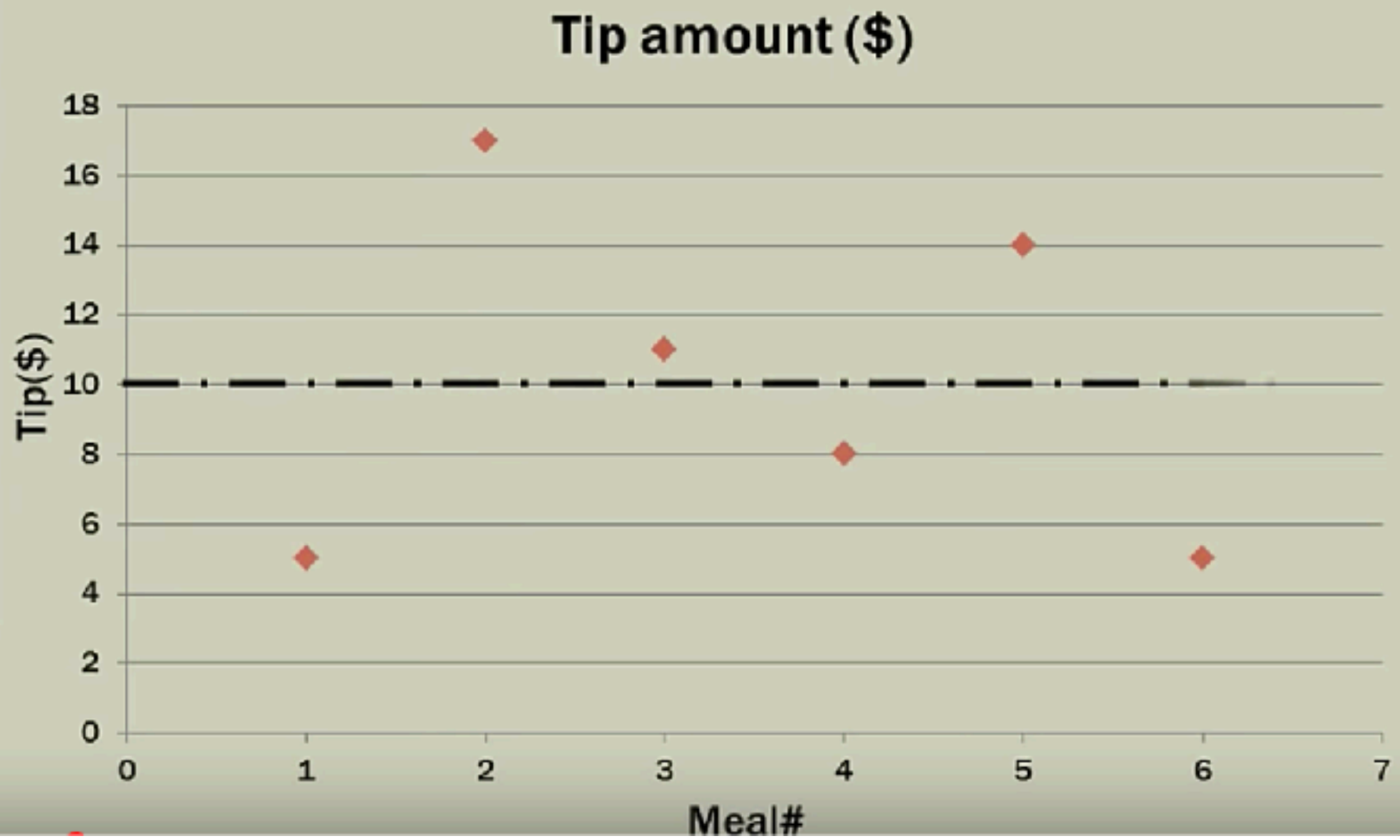
$$\bar{y} = \$10$$



TIPS FOR SERVICE

Meal#	Tip amount ($)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00

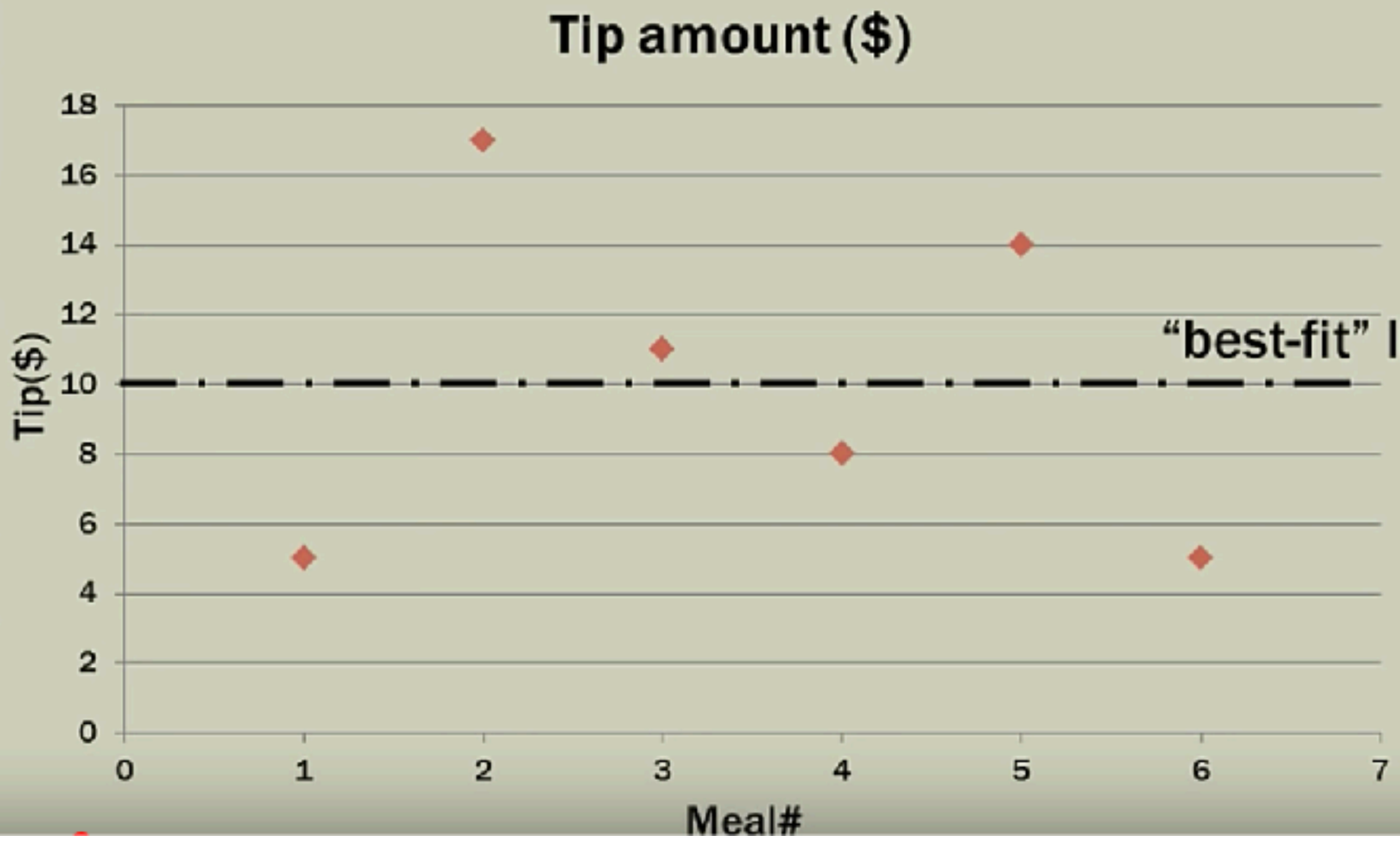
$$\bar{y} = \$10$$



TIPS FOR SERVICE

Meal#	Tip amount ($)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00

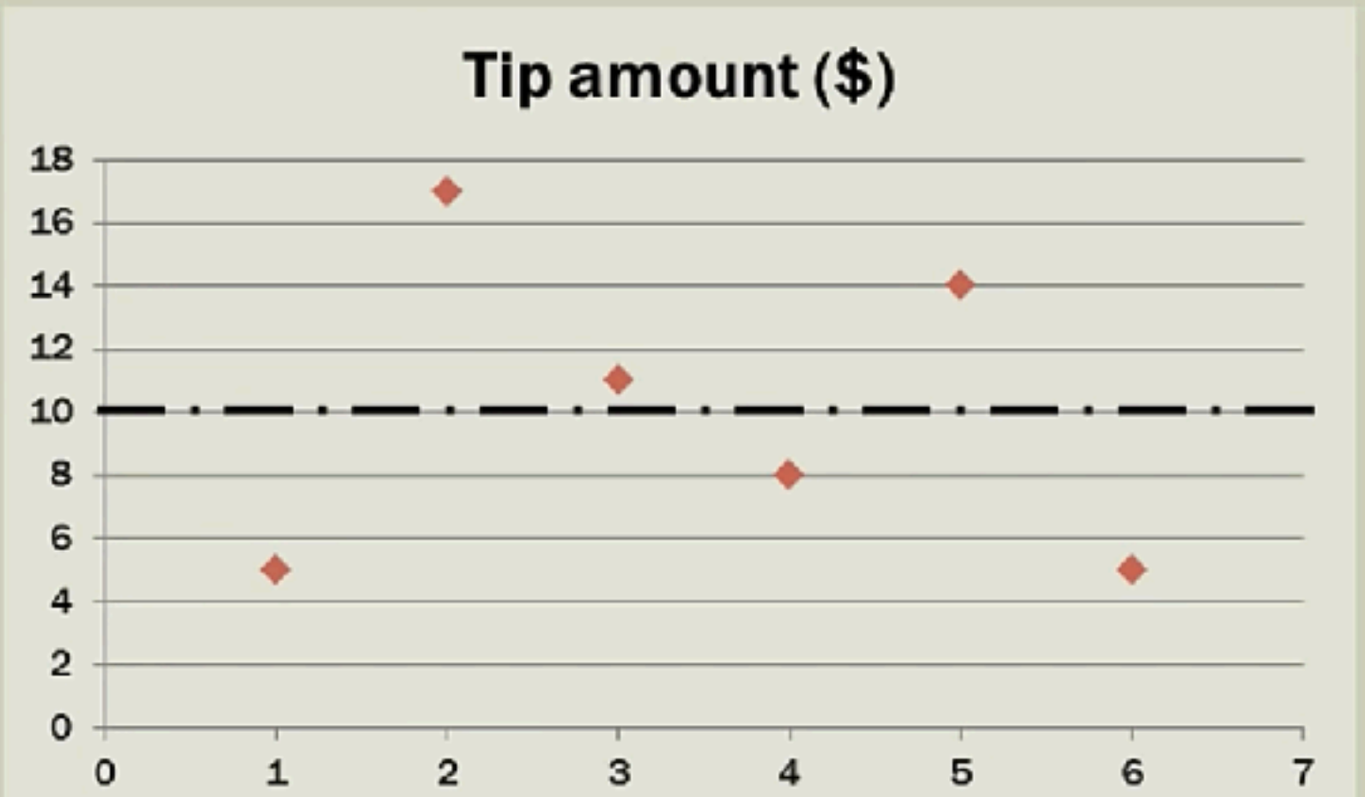
$$\bar{y} = \$10$$



TIPS FOR SERVICE

Meal#	Tip amount ($)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00

$$\bar{y} = \$10$$



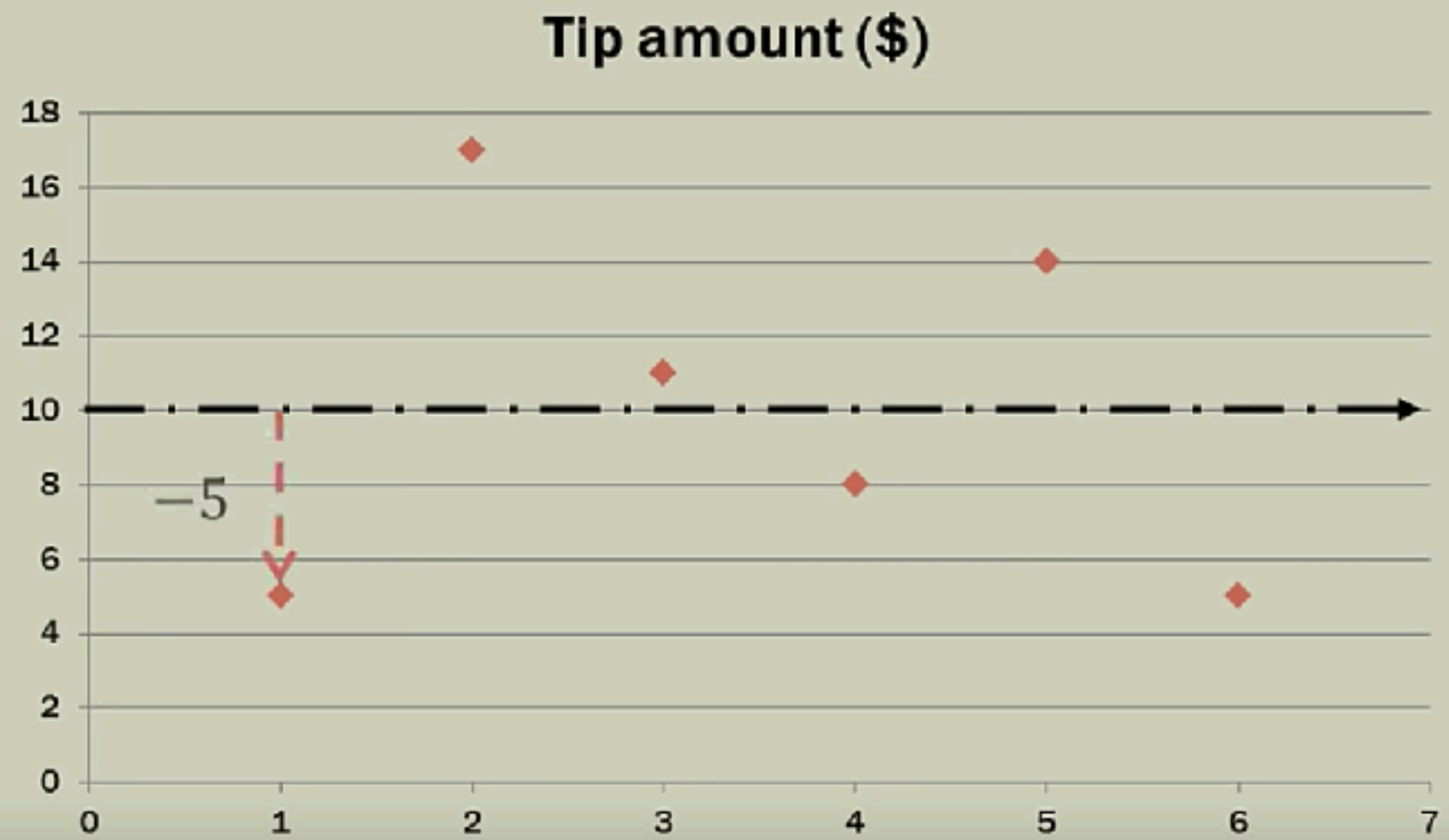
$$\bar{y} = \$10$$

With only one variable, and no other information, the best prediction for the next measurement is the mean of the sample itself. The variability in the tip amounts can only be explained by the tips themselves.

"GOODNESS OF FIT" FOR THE TIPS

Meal#	Tip amount ($)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00

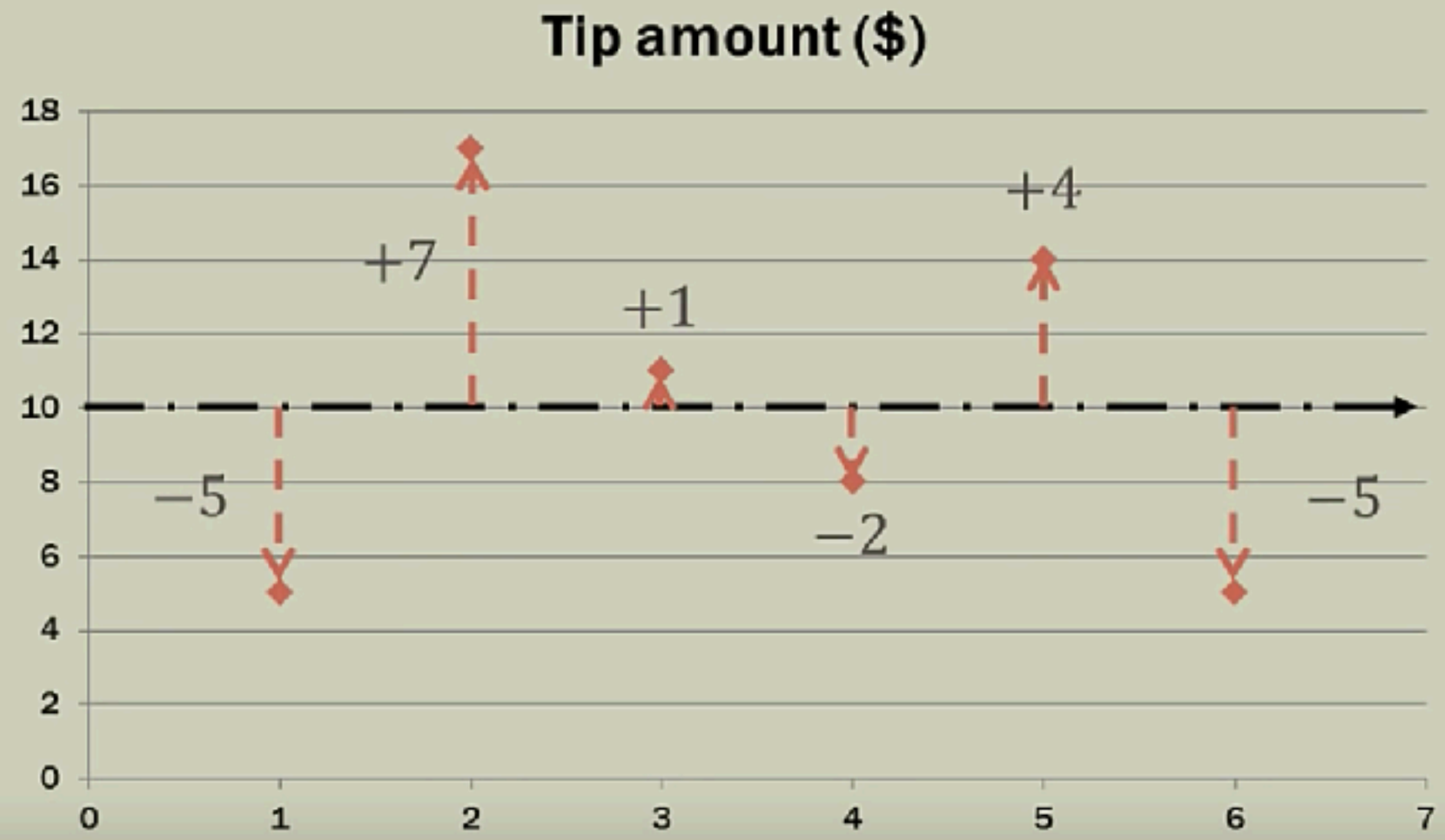
$$\bar{y} = \$10$$



"GOODNESS OF FIT" FOR THE TIPS

Meal#	Tip amount ($)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00

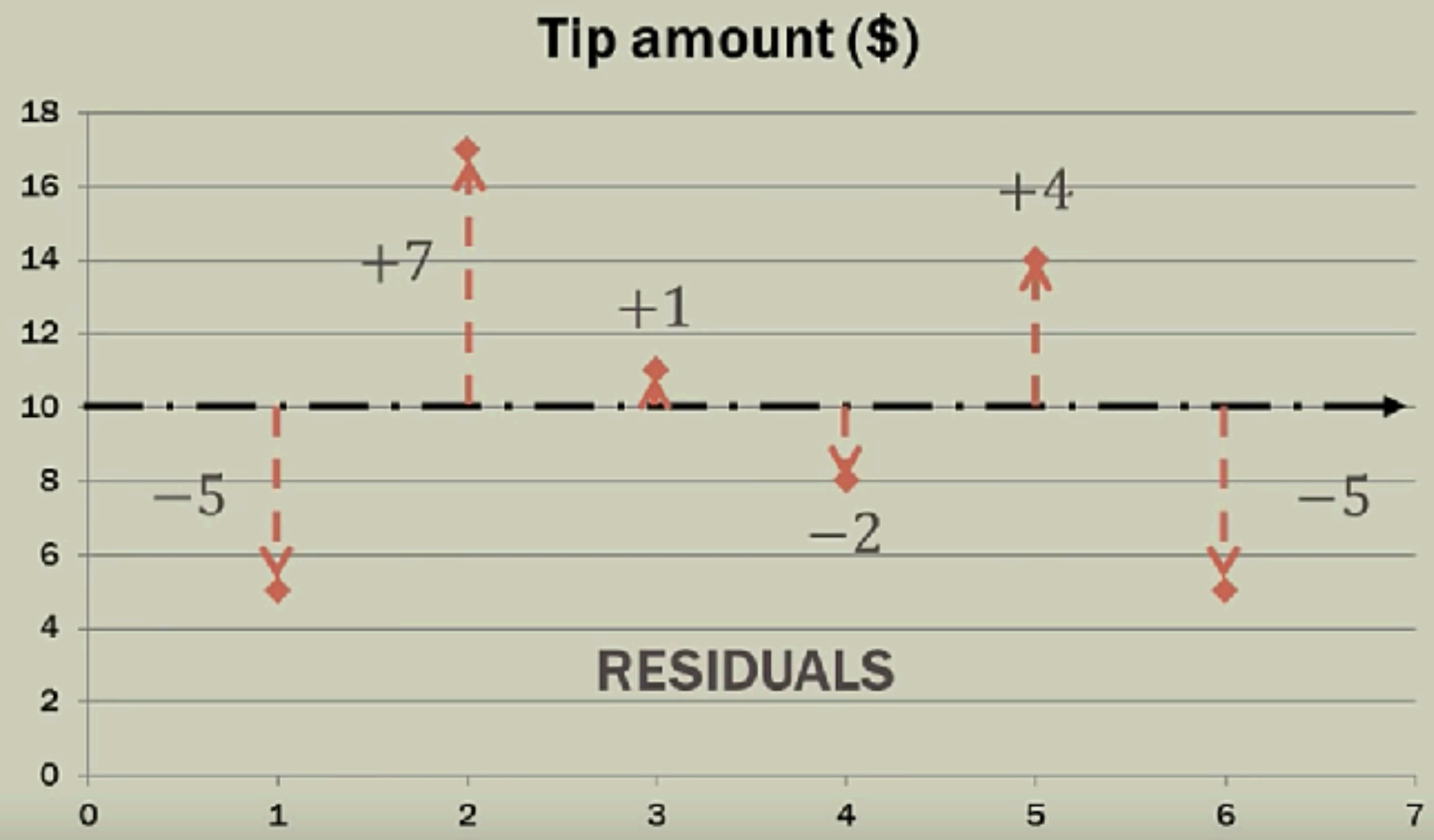
$$\bar{y} = \$10$$



"GOODNESS OF FIT" FOR THE TIPS

Meal#	Tip amount ($)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00

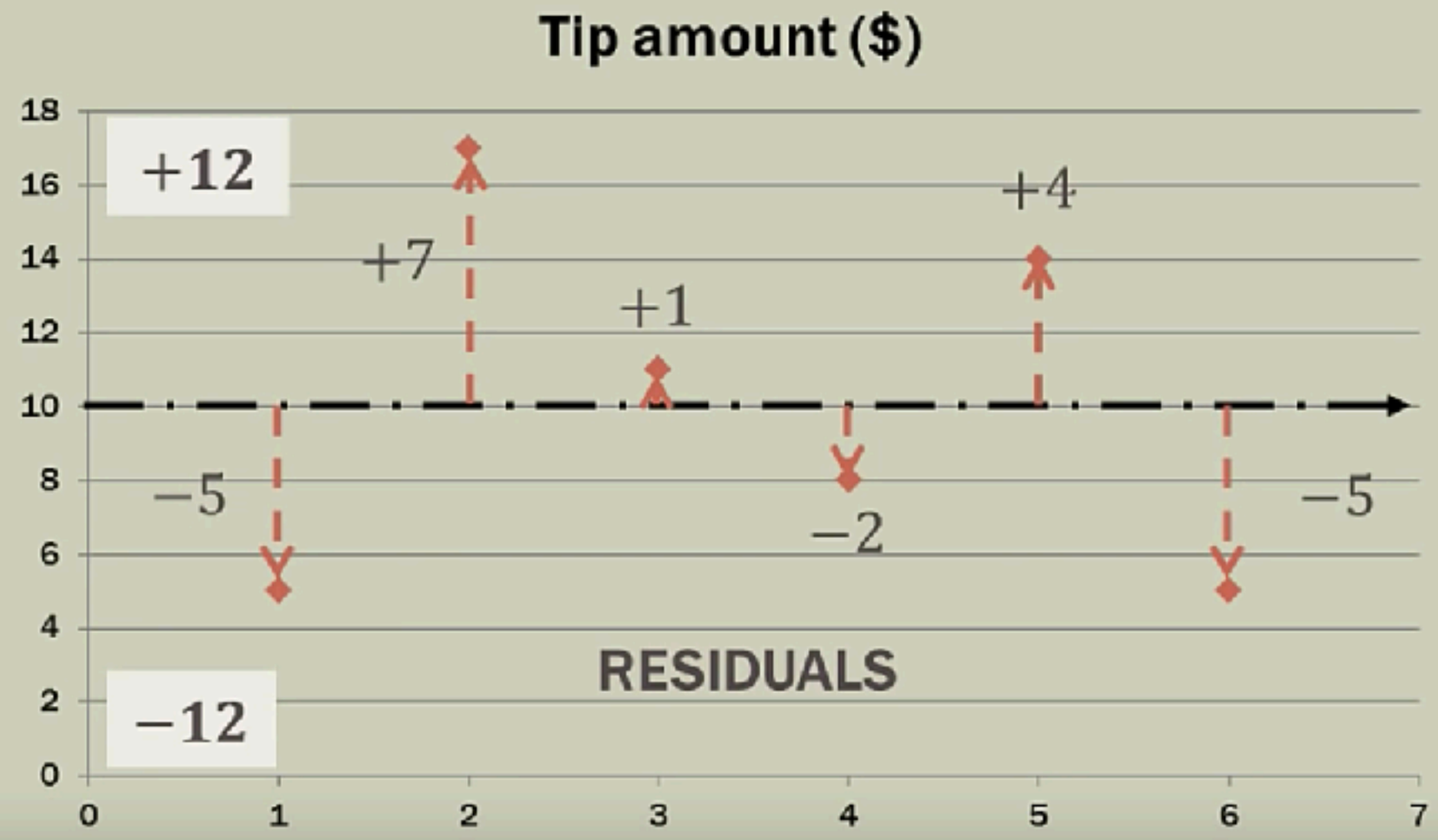
$$\bar{y} = \$10$$



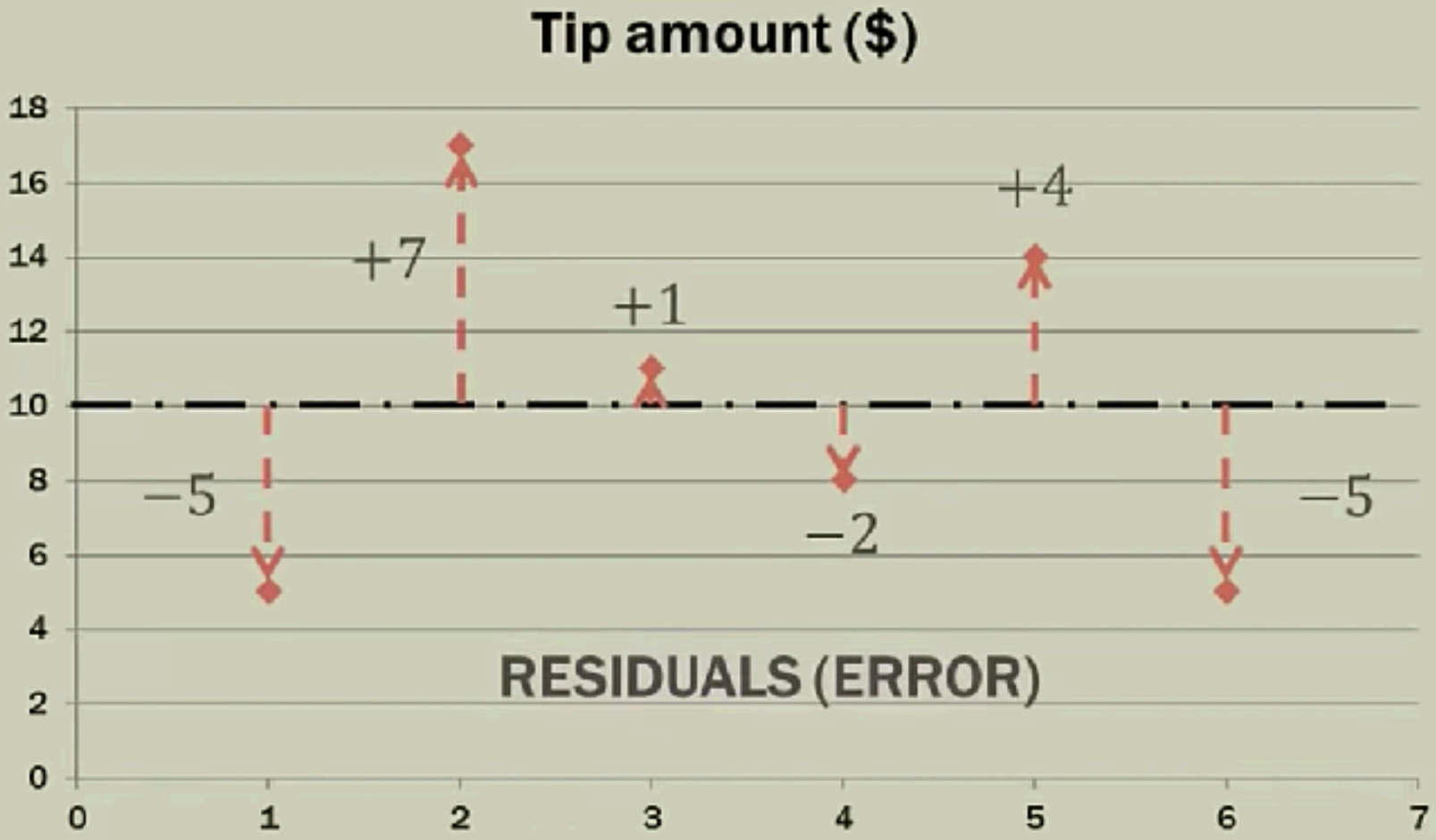
"GOODNESS OF FIT" FOR THE TIPS

Meal#	Tip amount ($)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00

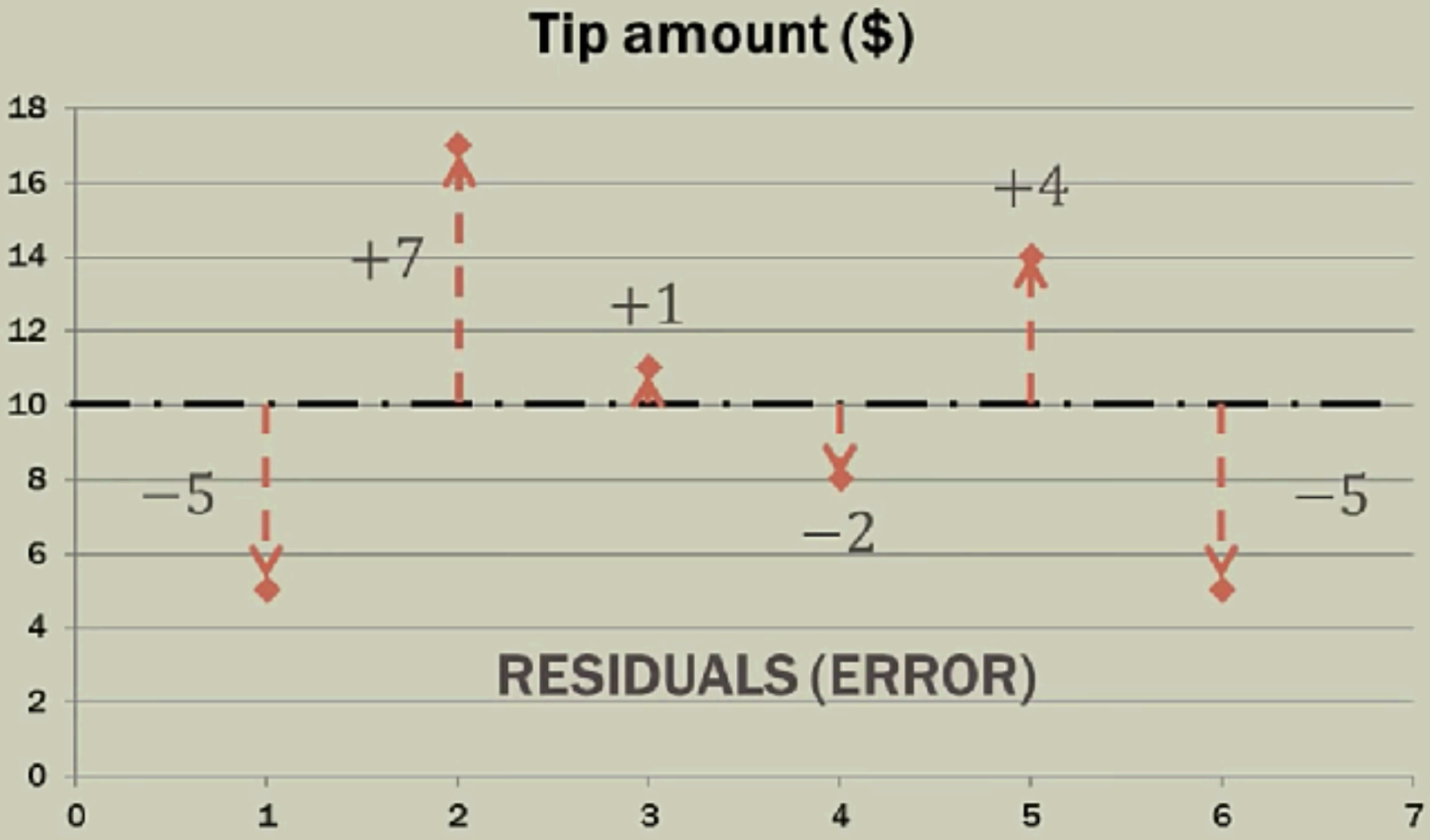
$$\bar{y} = \$10$$



SQUARING THE RESIDUALS (ERROR)

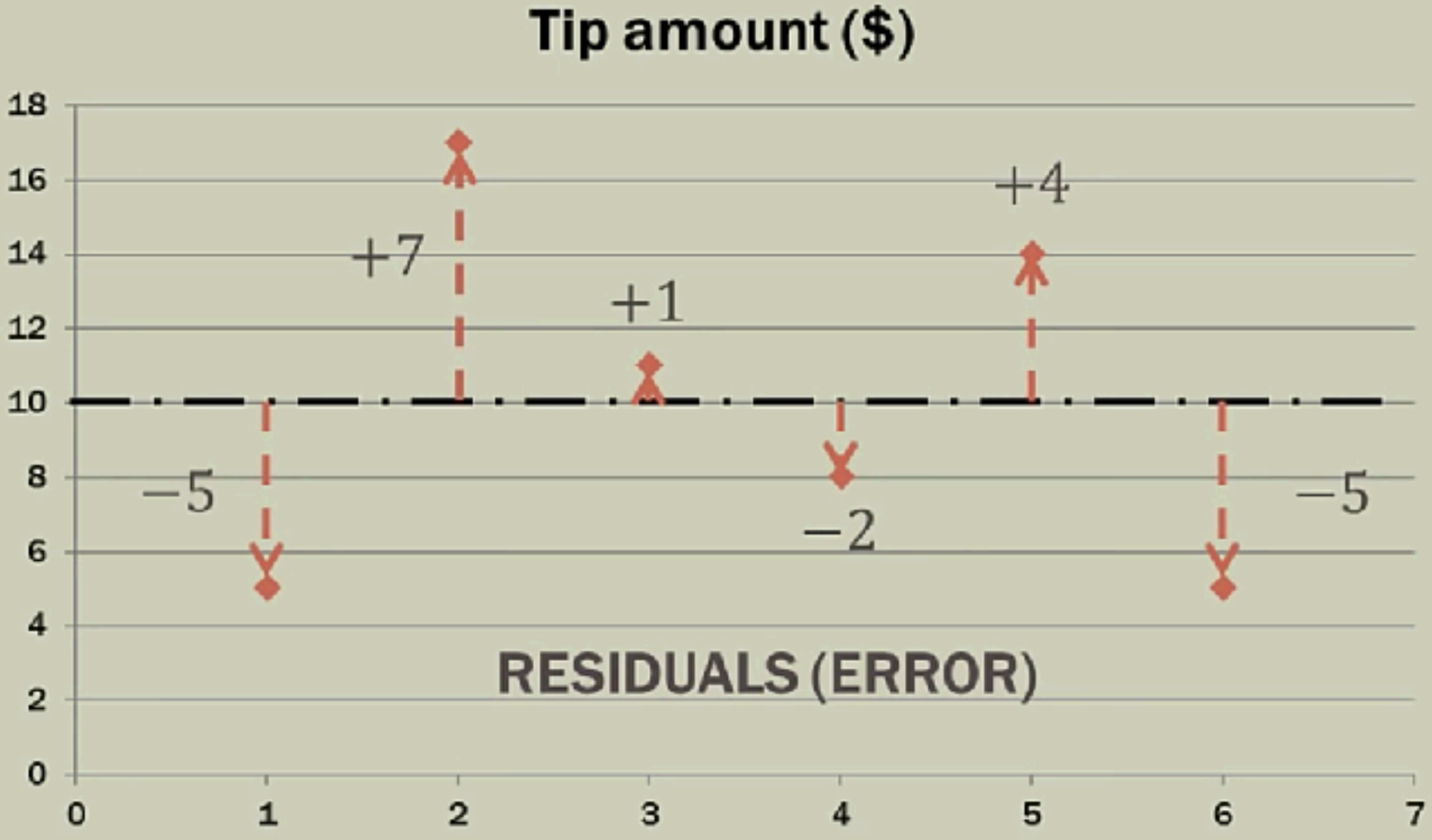


SQUARING THE RESIDUALS (ERROR)



Meal#	Residual	Residual^2
1	−5	25
2	+7	49
3	+1	1
4	−2	4
5	+4	16
6	−5	25

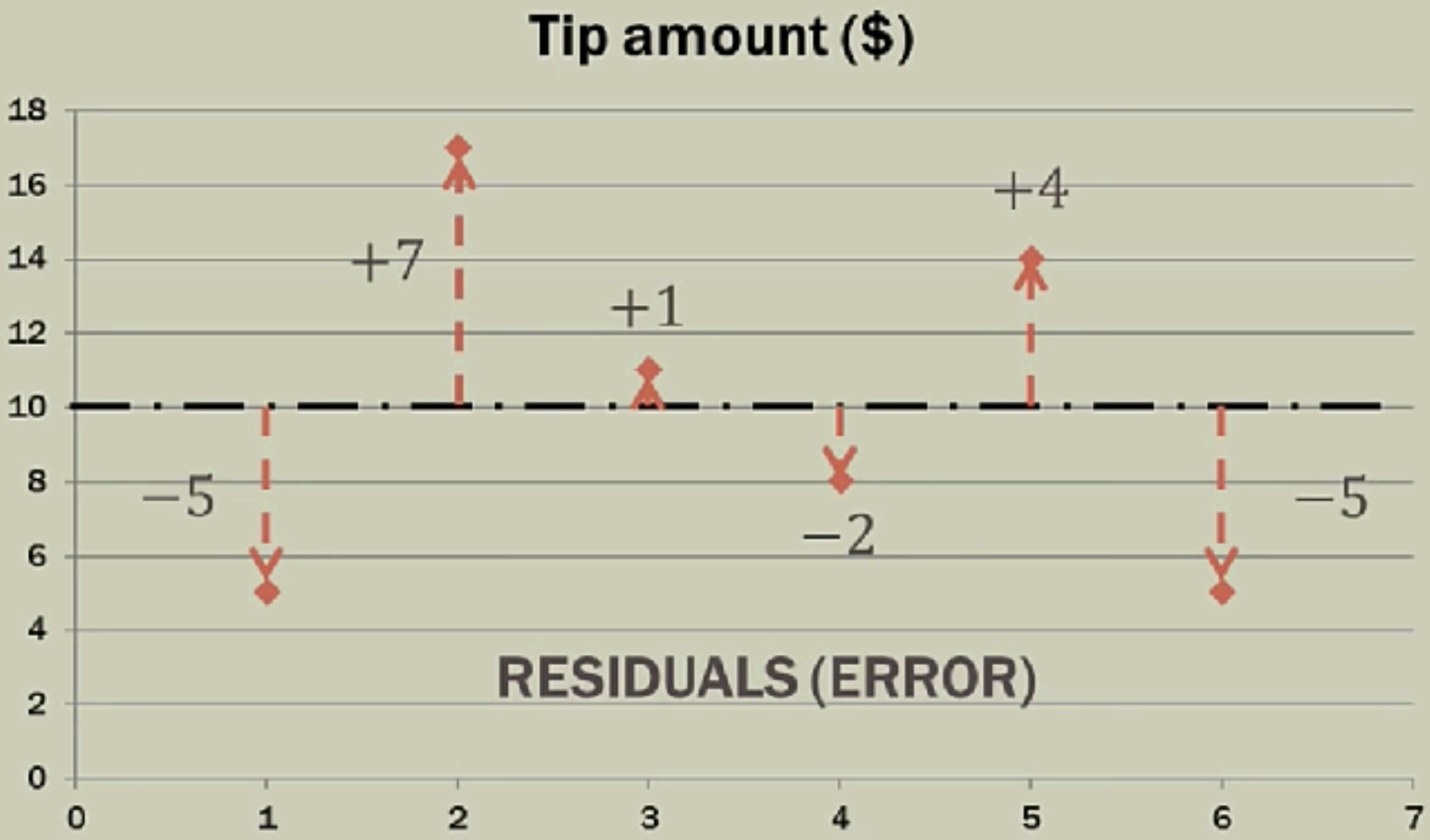
SQUARING THE RESIDUALS (ERROR)



Meal#	Residual	Residual^2
1	−5	25
2	+7	49
3	+1	1
4	−2	4
5	+4	16
6	−5	25

Why square the residuals? 1) makes them positive and 2) emphasizes larger deviations.

SQUARING THE RESIDUALS (ERROR)

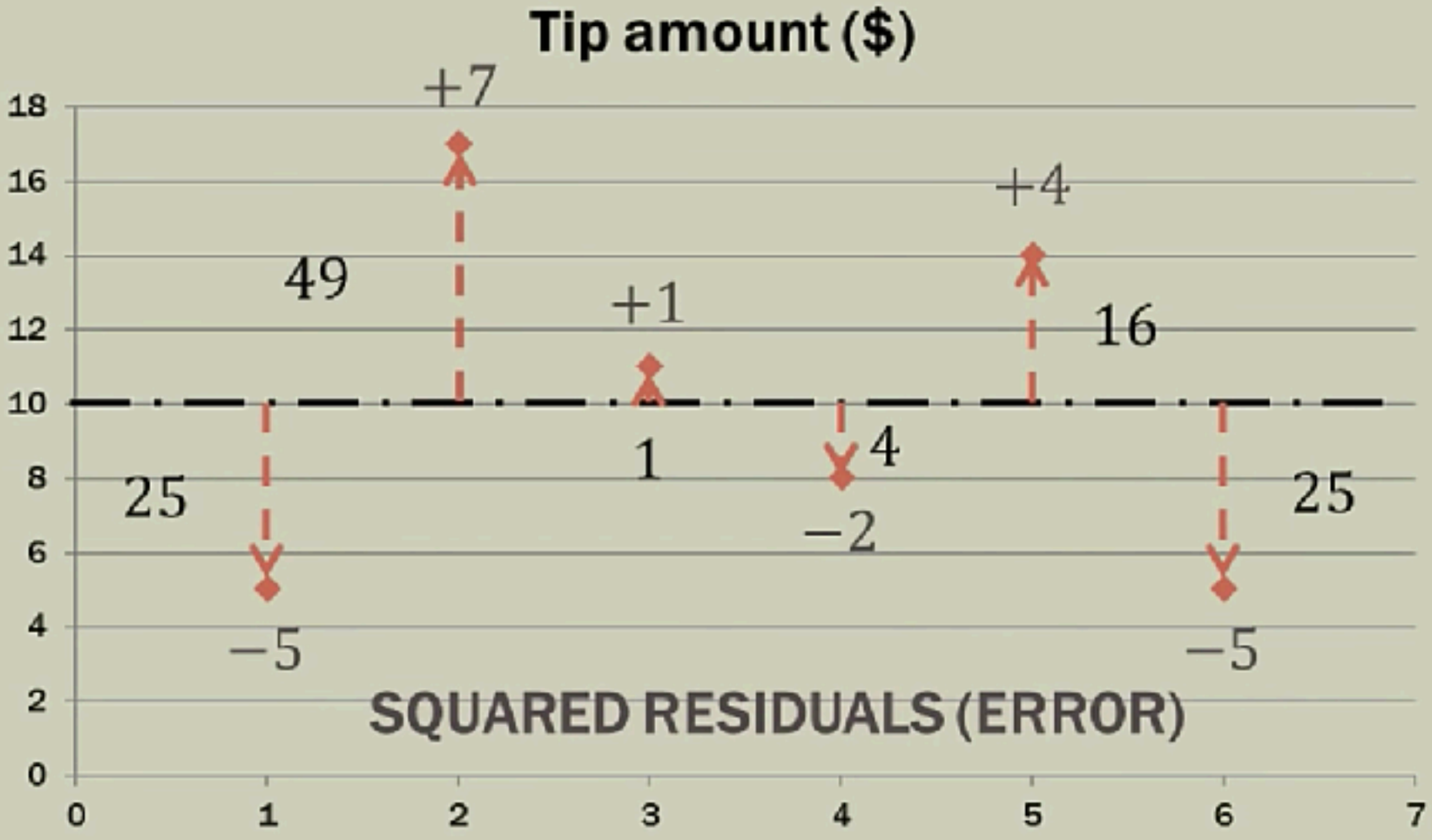


Meal#	Residual	Residual^2
1	−5	25
2	+7	49
3	+1	1
4	−2	4
5	+4	16
6	−5	25

Why square the residuals? 1) makes them positive and 2) emphasizes larger deviations.

$$Sum\ of\ squared\ errors\ (SSE) = 120$$

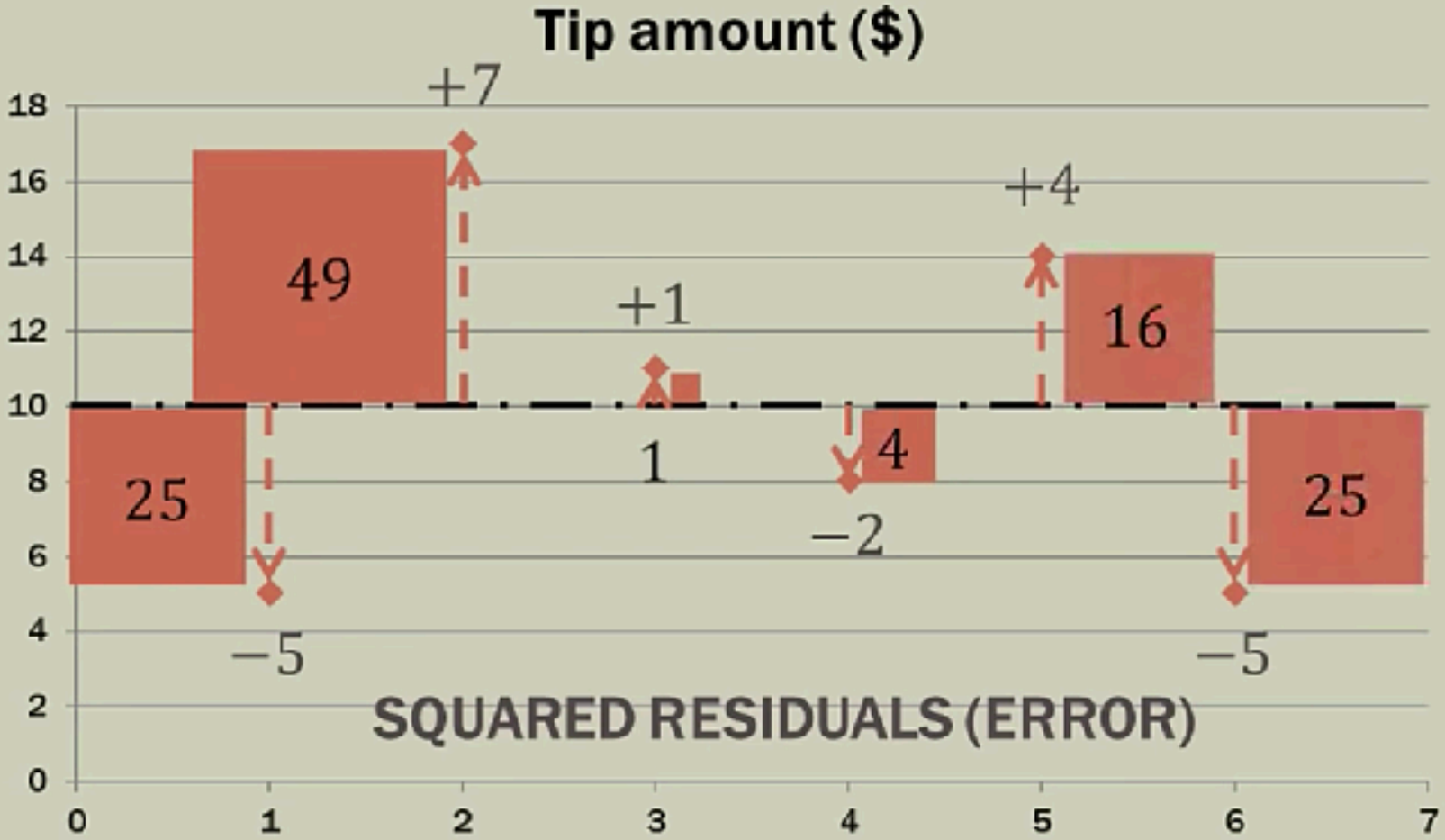
SQUARING THE RESIDUALS (ERROR)



Meal#	Residual	Residual²
1	−5	25
2	+7	49
3	+1	1
4	−2	4
5	+4	16
6	−5	25

$$Sum\ of\ squared\ errors\ (SSE) = 120$$

SQUARING THE RESIDUALS (ERROR)



Meal#	Residual	Residual^2
1	−5	25
2	+7	49
3	+1	1
4	−2	4
5	+4	16
6	−5	25

$$Sum\ of\ squared\ errors\ (SSE) = 120$$

SUM OF SQUARES

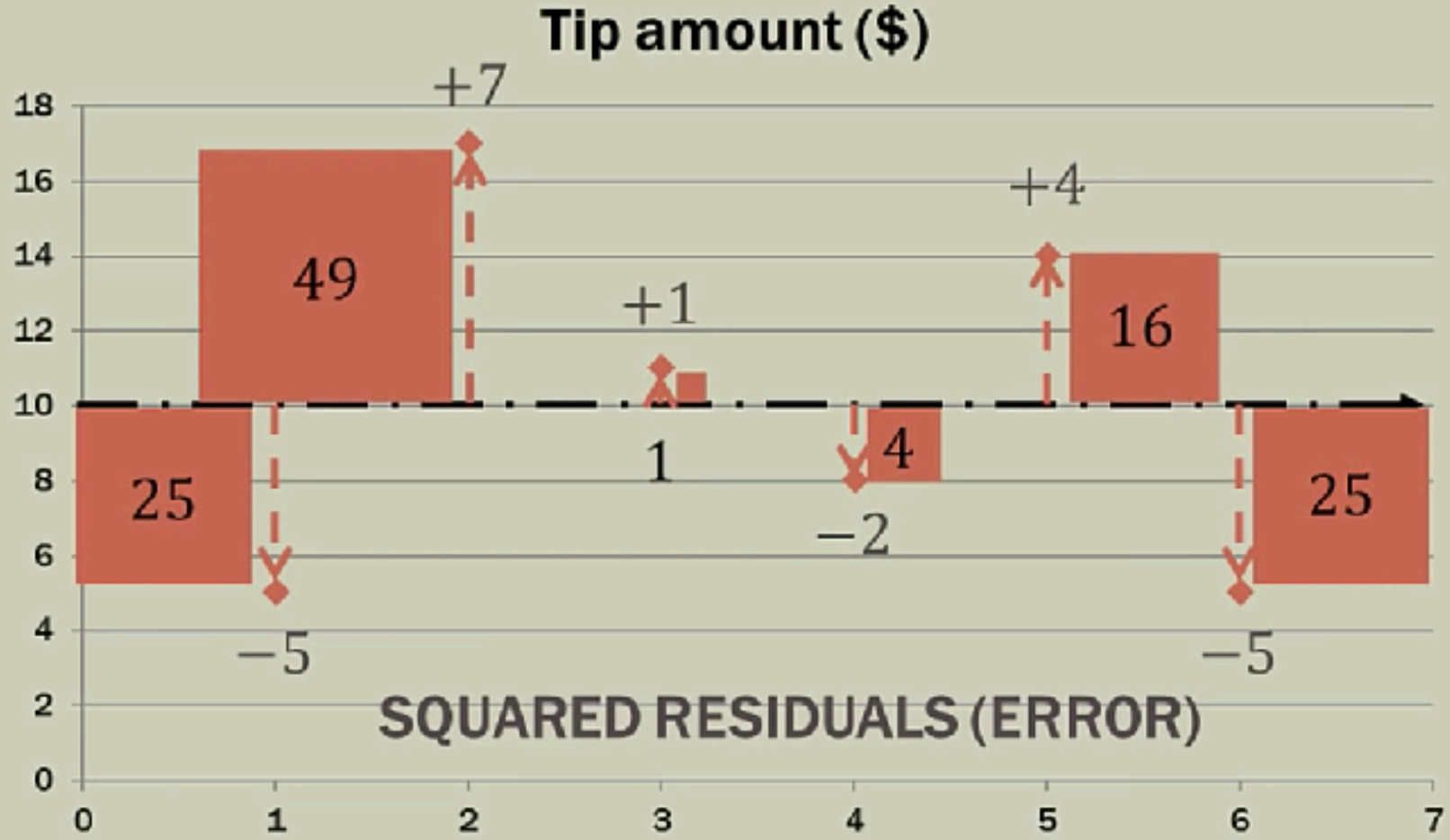
$$49 + 25 + 1 + 4 + 16 + 25 = 120$$

SUM OF SQUARES

The goal of simple linear regression is to create a linear model that minimizes the sum of squares of the residuals / error (SSE).

If our regression model is significant, it will "eat up" much of the raw SSE we had when we assumed (like this problem) that the independent variable did not even exist. The regression line will/should literally "fit" the data better. It will minimize the residuals.

VERY IMPORTANT



$$Sum\ of\ squared\ errors\ (SSE) = 120$$

When conducting simple linear regression with TWO variables, we will determine how good that line "fits" the data by comparing it to THIS TYPE; where we pretend the second variable does not even exist.

If a two-variable regression model looks like this example, what does the other variable do to help explain the dependent variable?

NOTHING.

QUICK REVIEW

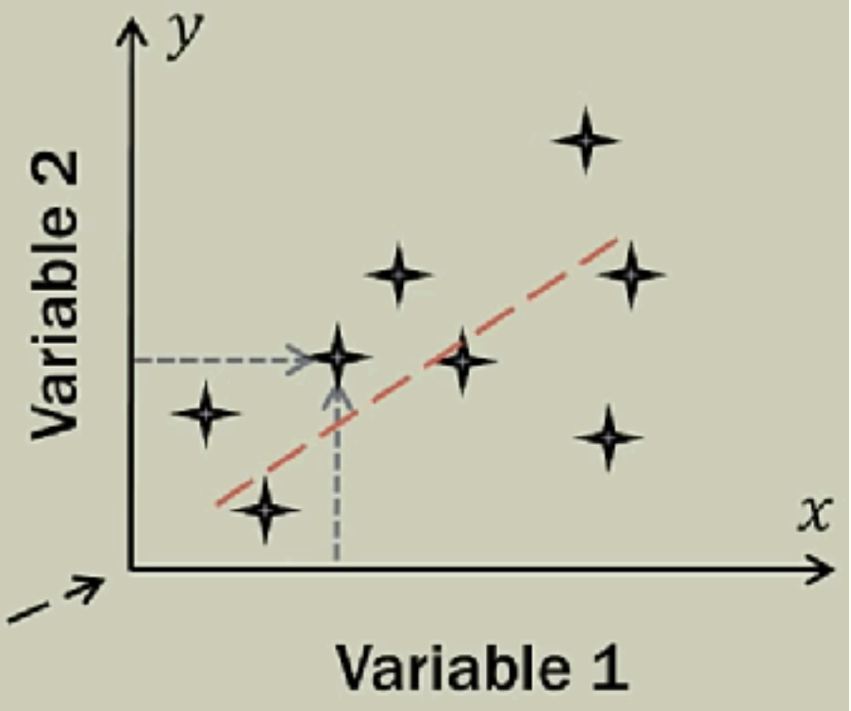
- Simple linear regression is really a **comparison of two models**
  - One is where the independent variable does not even exist
  - And the other uses the best fit regression line
- If there is only one variable, the best prediction for other values is the **mean of the "dependent" variable**
- The difference between the best-fit line and the observed value is called the **residual (or error)**
- The residuals are squared and then added together to generate **sum of squares (LITERALLY) residuals / error, SSE.**
- Simple linear regression is designed to find the best fitting line through the data that minimizes the SSE.

BIVARIATE STATISTICS

CORRELATION

ANOVA

LINEAR REGRESSION



The value of one variable, is a function of the other variable.

The value of y, is a function of x; $y = f(x)$.

The value of the dependent variable, is a function of the independent variable.

ALGEBRA REVIEW: LINES

slope−intercept form of a line

$$y = mx + b$$

x = random variable

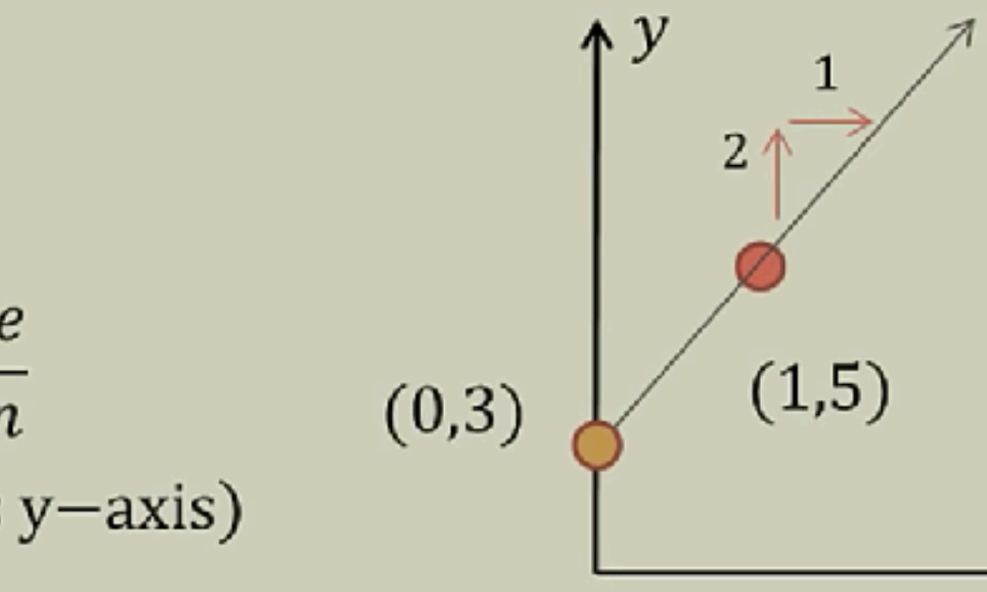
m = slope of the line $\frac{rise}{run}$

b = y−intercept (crosses y−axis)

y−intercept is where $x = 0$

Coordinate of $(0, y)$

$$y = 2x + 3$$



$$m = slope = \frac{2}{1}$$

$y = 2(0) + 3$

$y = 3$

(0,3)

where $x = 1$;

$y = 2(1) + 3$

$y = 5$

SIMPLE LINEAR REGRESSION MODEL

$$y = mx + b$$

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

β_0 = $y-$intercept **population** parameter

β_1 = slope **population** parameter

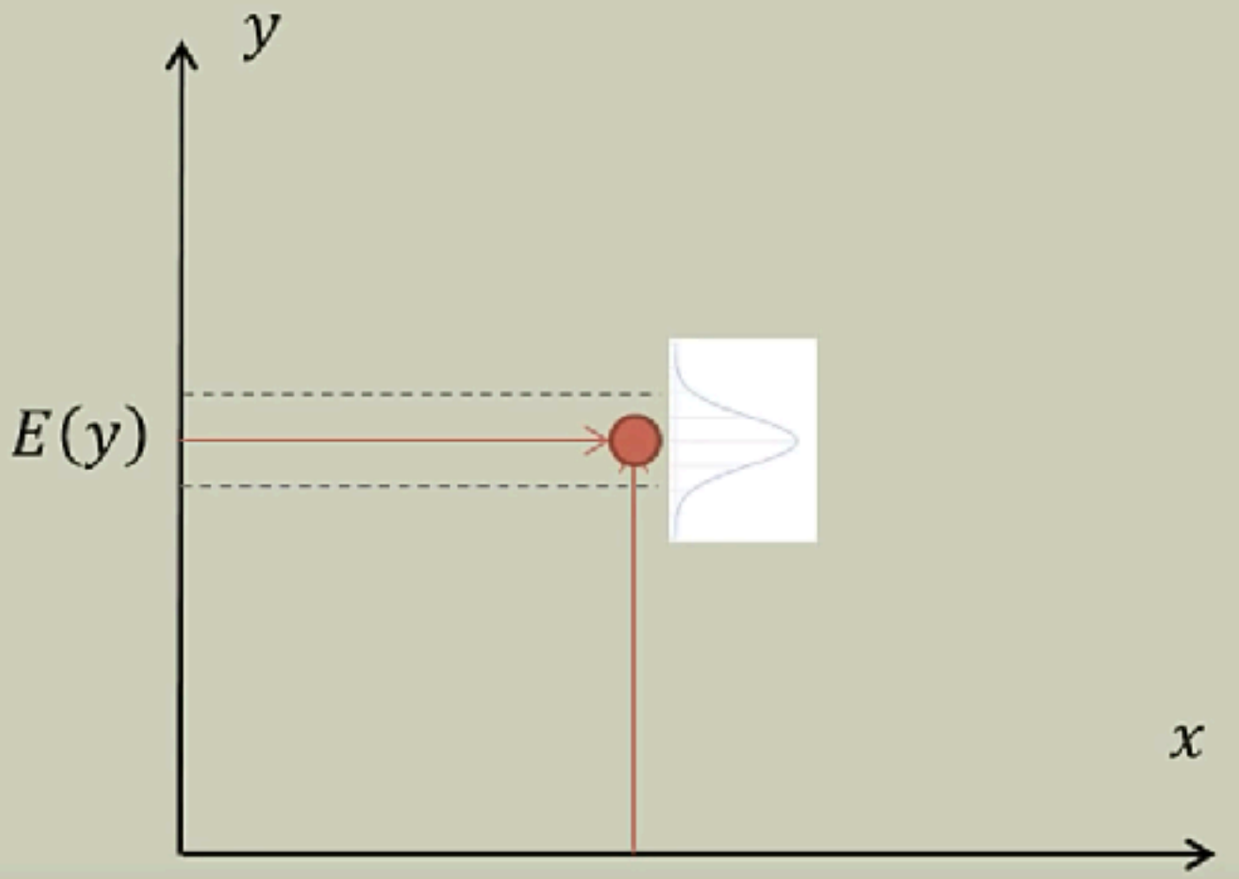
ϵ = error term, unexplained variation in y

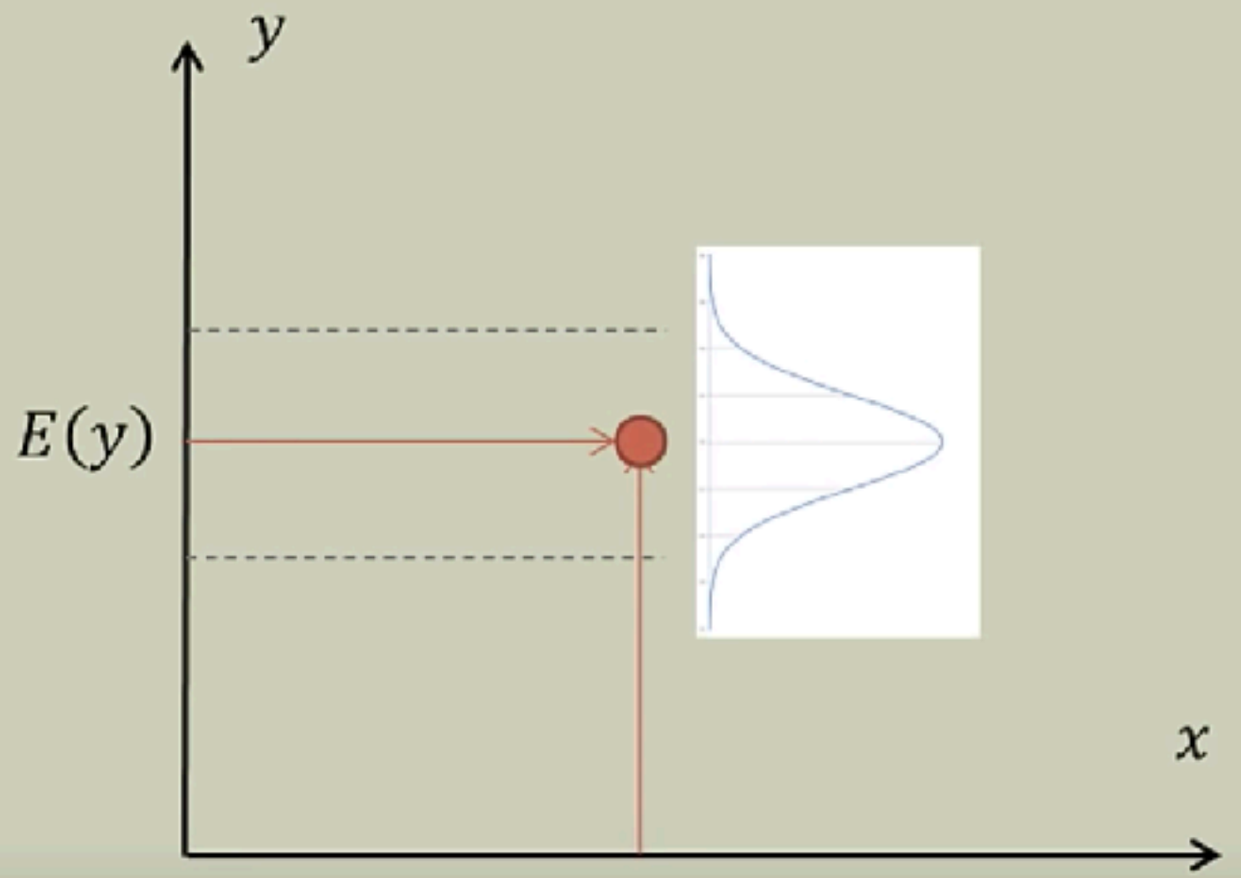
Simple Linear Regression Equation

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

$E(y)$ is the **mean or expected value** of y, for a given value of x

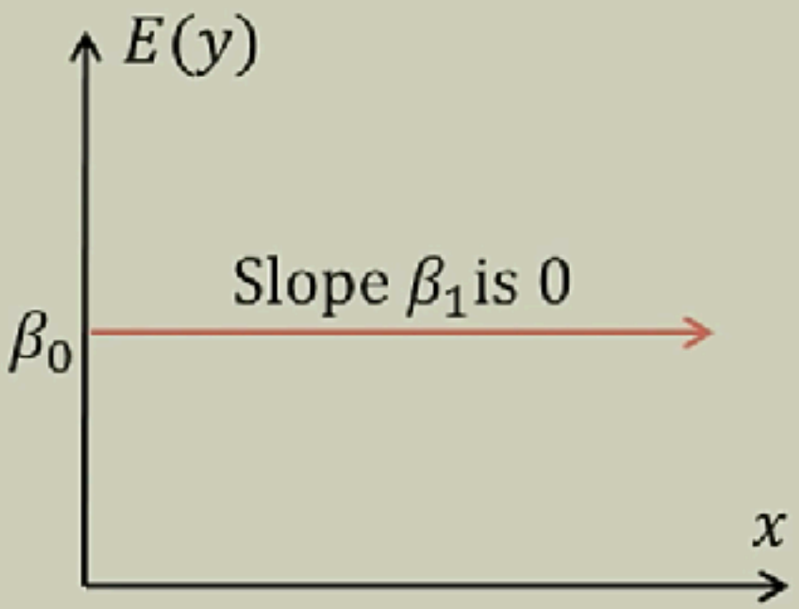
DISTRIBUTION OF y-VALUES



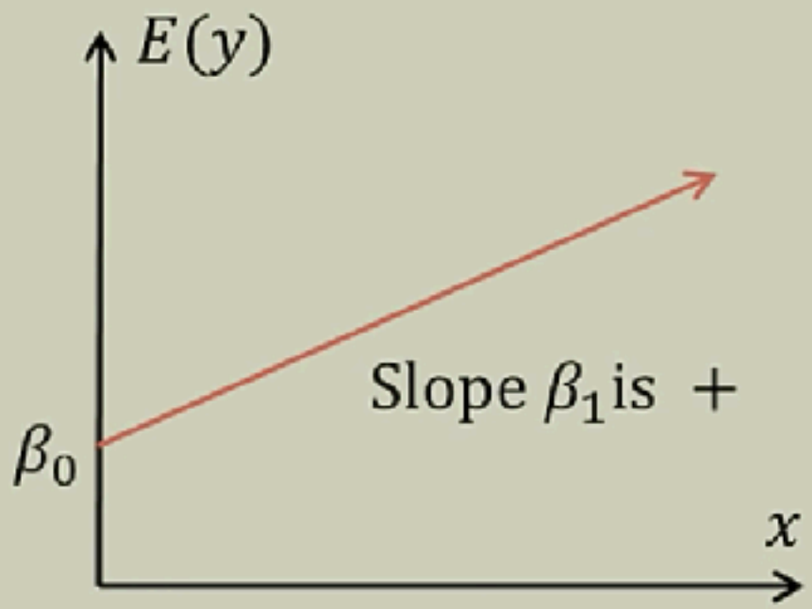


GENERAL REGRESSION LINES

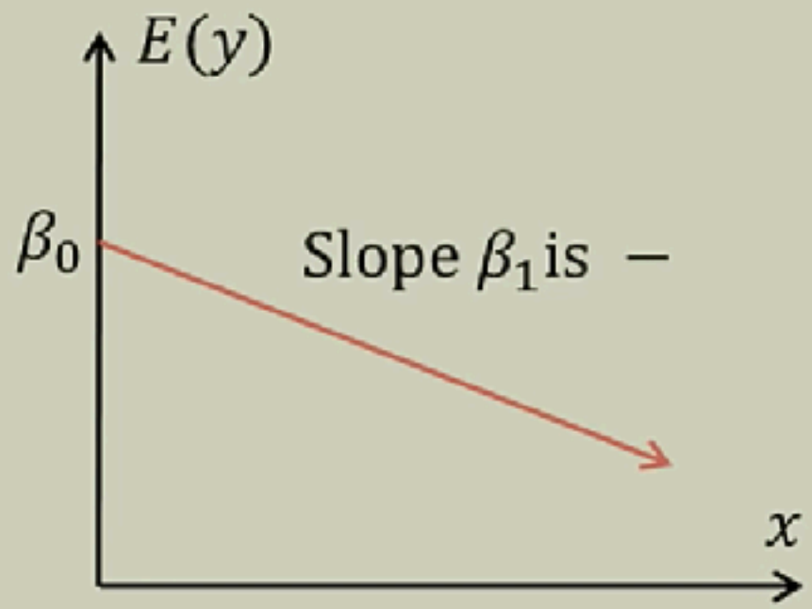
$$E(y) = \beta_0 + \beta_1 x$$



$$E(y) = \beta_0 + 0(x)$$



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



$$E(y) = \beta_0 - \beta_1 x$$

REGRESSION EQUATION WITH ESTIMATES

If we actually knew the population parameters, β_0 and β_1, we could use the Simple Linear Regression Equation.

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

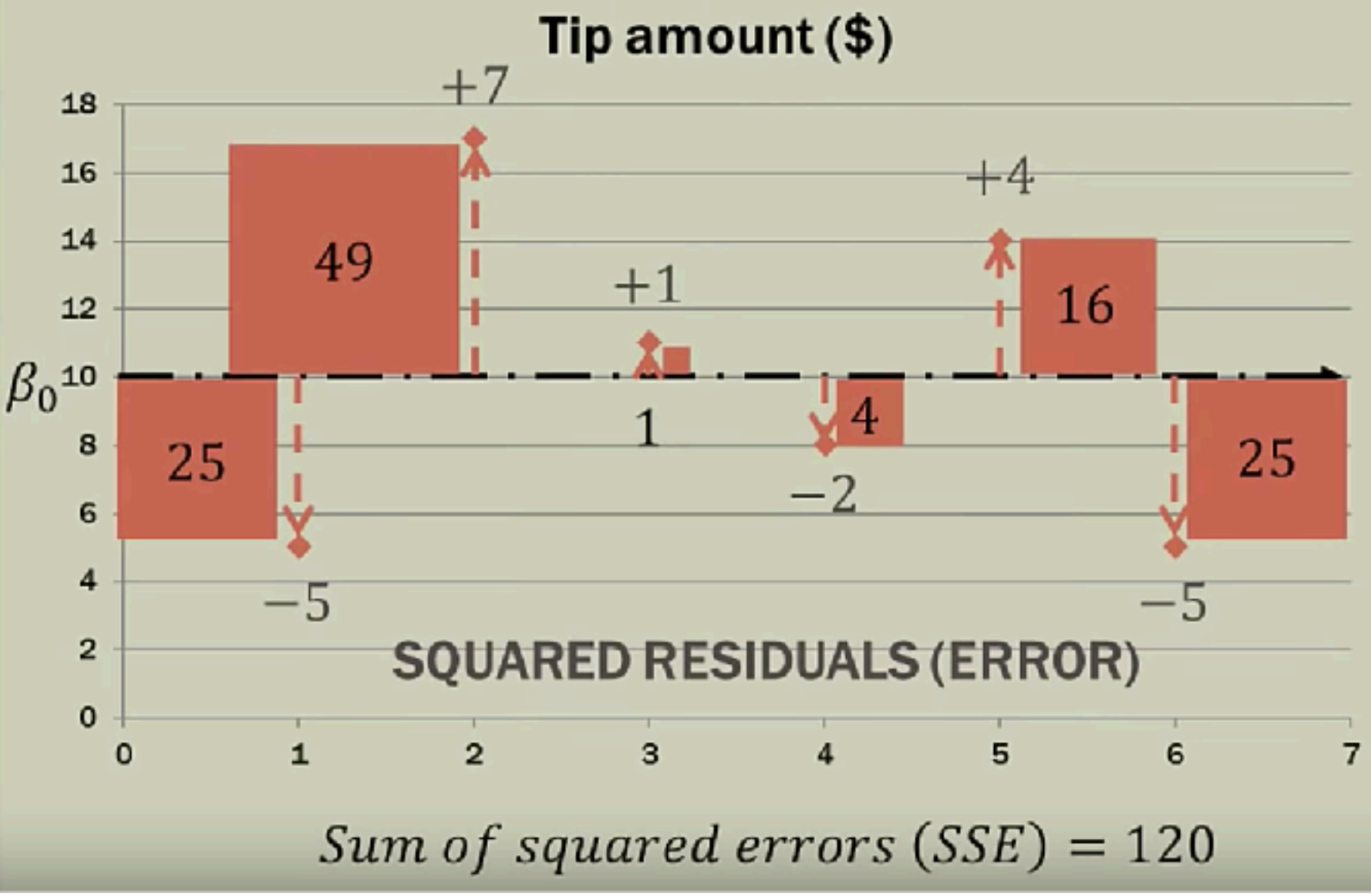
In reality we almost never have the population parameters. Therefore we will estimate them using sample data. When using sample data, we have to change our equation a little bit.

$\hat{y}$, pronounced "y-hat" is the point estimator of $E(y)$

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

$\hat{y}$, is the **mean value of** y for a given value of x.

WHEN THE SLOPE, $\beta_1 = 0$



$Sum\ of\ squared\ errors\ (SSE) = 120$

When conducting simple linear regression with TWO variables, we will determine how good the regression line "fits" the data by comparing it to THIS TYPE; where we pretend the second variable does not even exist; the slope, $\beta_1 = 0$.

In this situation, the value of $\hat{y}$ is 10 for every value of x.

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

$\hat{y} = b_0 + (0)x$

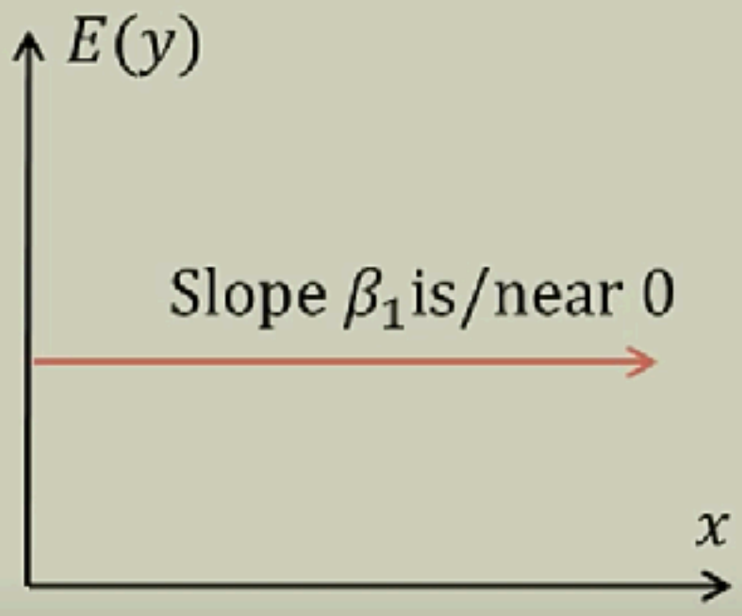
$\hat{y} = b_0$

$b_0 = 10$

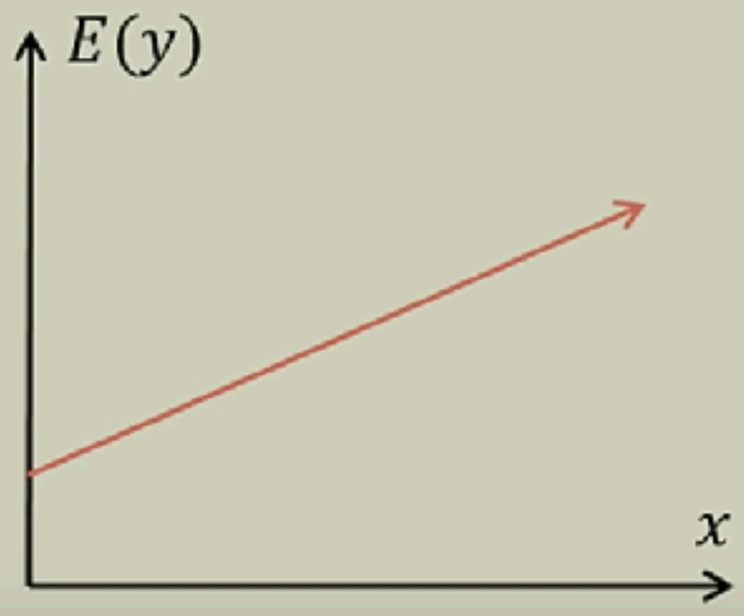
$\hat{y} = 10$

PATTERN MATCHING TO GENERAL REGRESSION MODELS

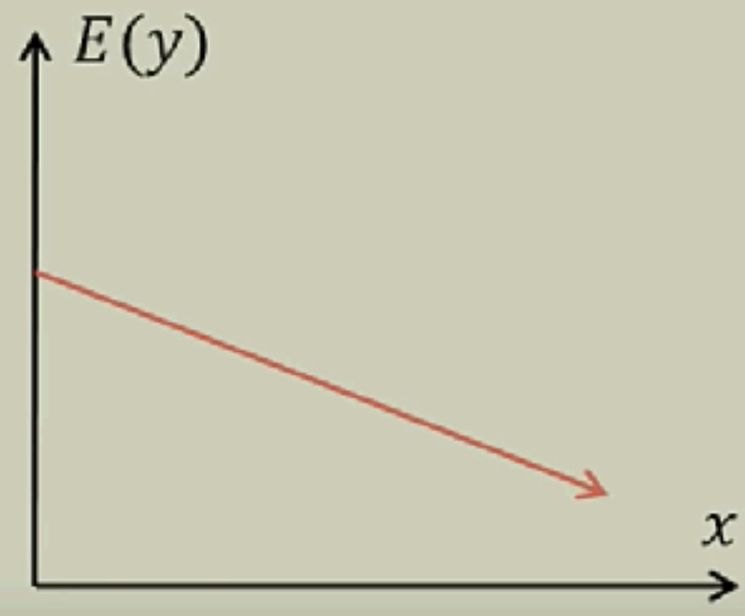
$\hat{y} = 0.3 - 3.3x$



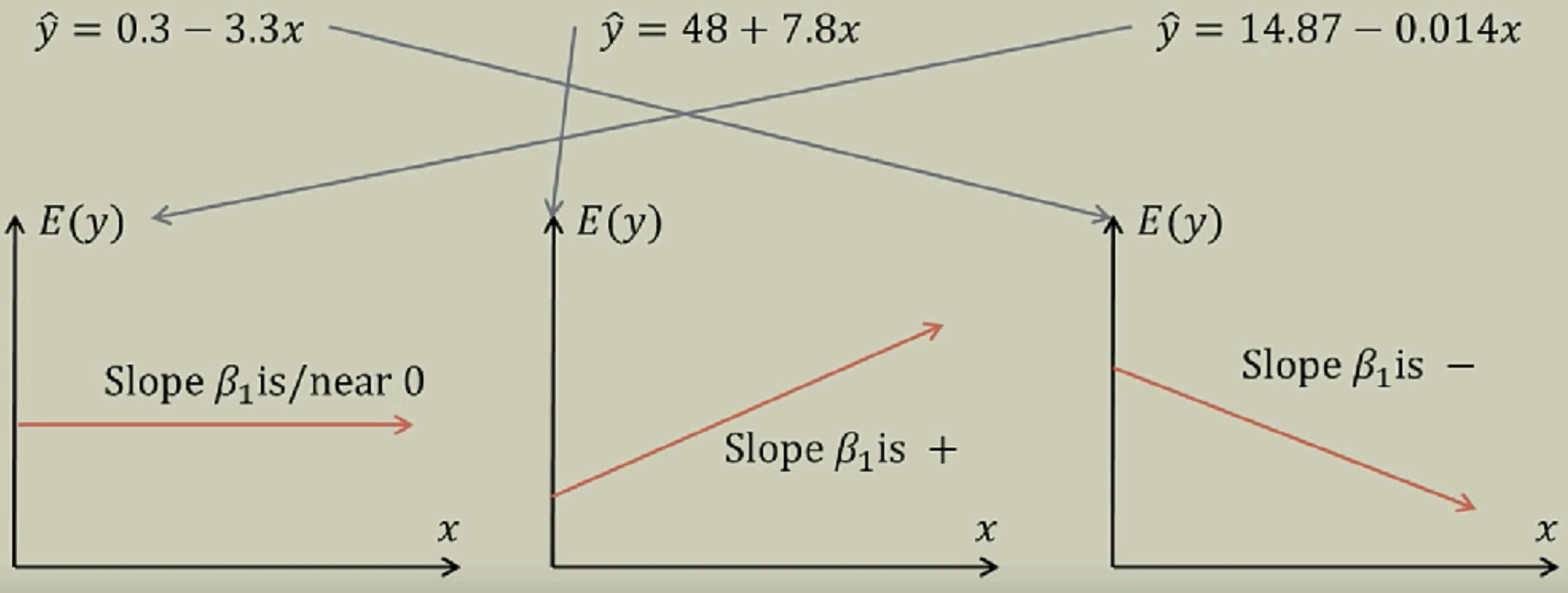
$\hat{y} = 48 + 7.8x$



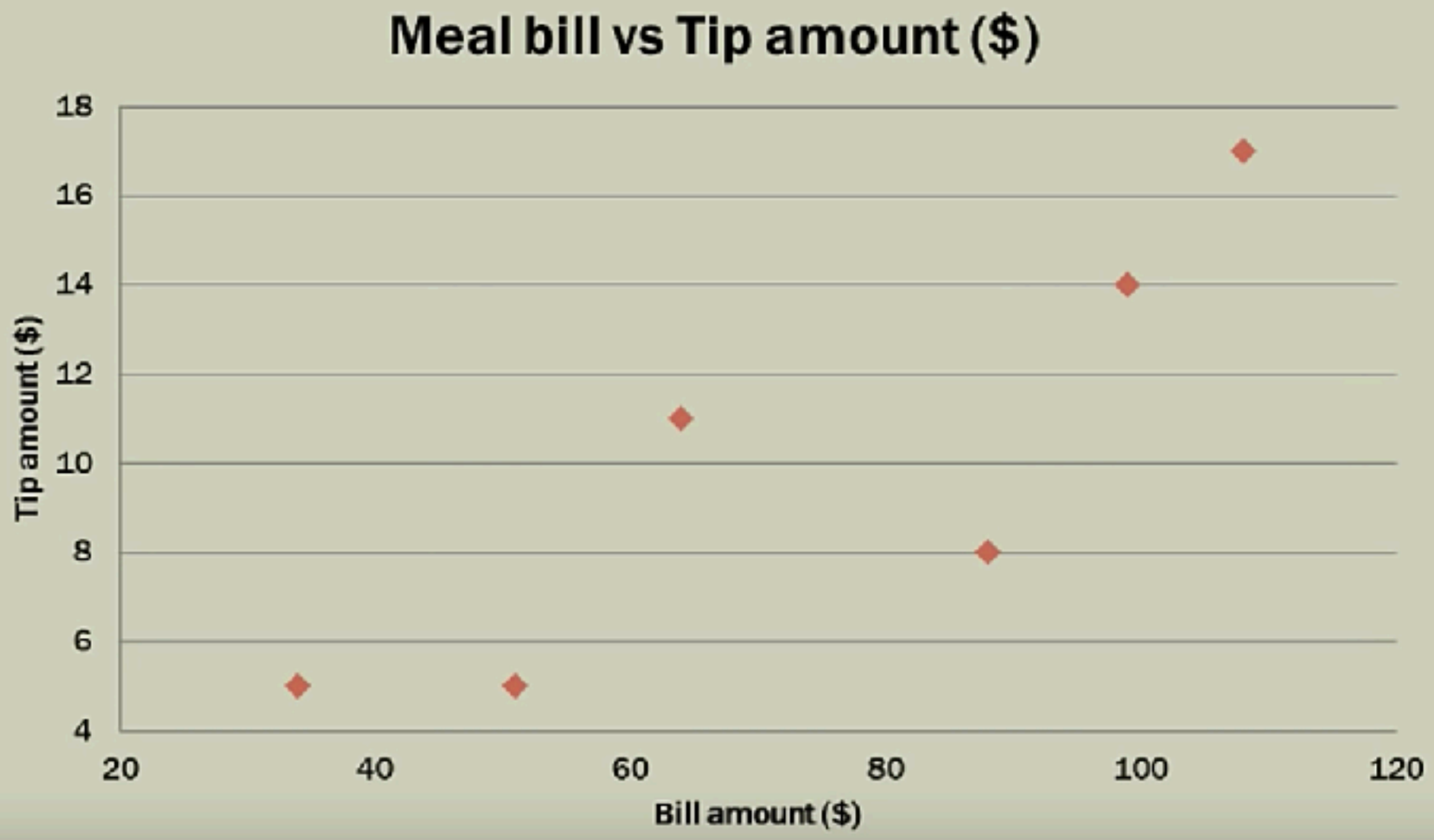
$\hat{y} = 14.87 - 0.014x$



PATTERN MATCHING TO GENERAL REGRESSION MODELS



GETTING READY FOR LEAST SQUARES



Bill ($)	Tip ($)
34.00	5.00
108.00	17.00
64.00	11.00
88.00	8.00
99.00	14.00
51.00	5.00

TIPS FOR SERVICE

Let's assume that you are a small restaurant owner or a very business minded server / waiter at a nice restaurant. Here in the U.S. "tips" are a very important part of a waiter's pay. Most of the time the dollar amount of tip is related to the dollar amount of the total bill.

TIPS FOR SERVICE

Let's assume that you are a small restaurant owner or a very business minded server / waiter at a nice restaurant. Here in the U.S. "tips" are a very important part of a waiter's pay. Most of the time the dollar amount of tip is related to the dollar amount of the total bill.

As the waiter or owner, you would like to develop a model that will allow you to make a prediction about what amount of tip to expect. Therefore one evening, you collect data for six meals.

LEAST SQUARES CRITERION

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

y_i = observed value of dependent variable (tip amount)

$\hat{y}_i$ = estimated(predicted)value of the dependent variable (predicted tip amount)

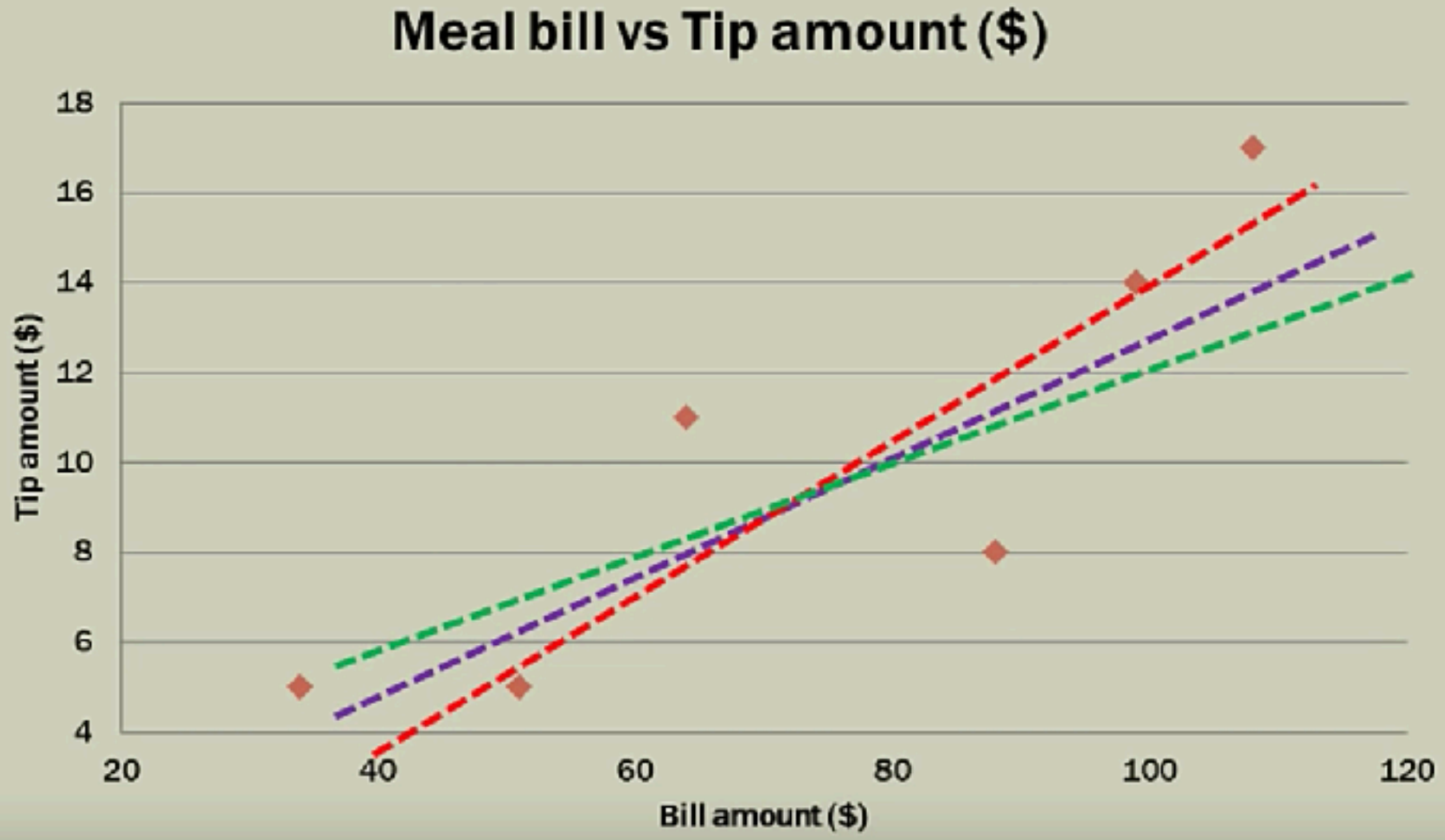
Plain English. The goal is to minimize the sum of the squared differences between the observed value for the dependent variable (y_i) and the estimated/predicted value of the dependent variable $(\hat{y}_i)$ that is provided by the regression line. Sum of the squared residuals.

Not only that, but the sum of the squared residuals should be much smaller than when we just used the dependent variable alone; $\beta_1 = 0$, $\hat{y} = 10$ for all values of x. That sum of squared residuals was 120.

STEP 1: SCATTER PLOT

Bill ($)	Tip ($)
34.00	5.00
108.00	17.00
64.00	11.00
88.00	8.00
99.00	14.00
51.00	5.00

STEP 2: LOOK FOR A VISUAL LINE

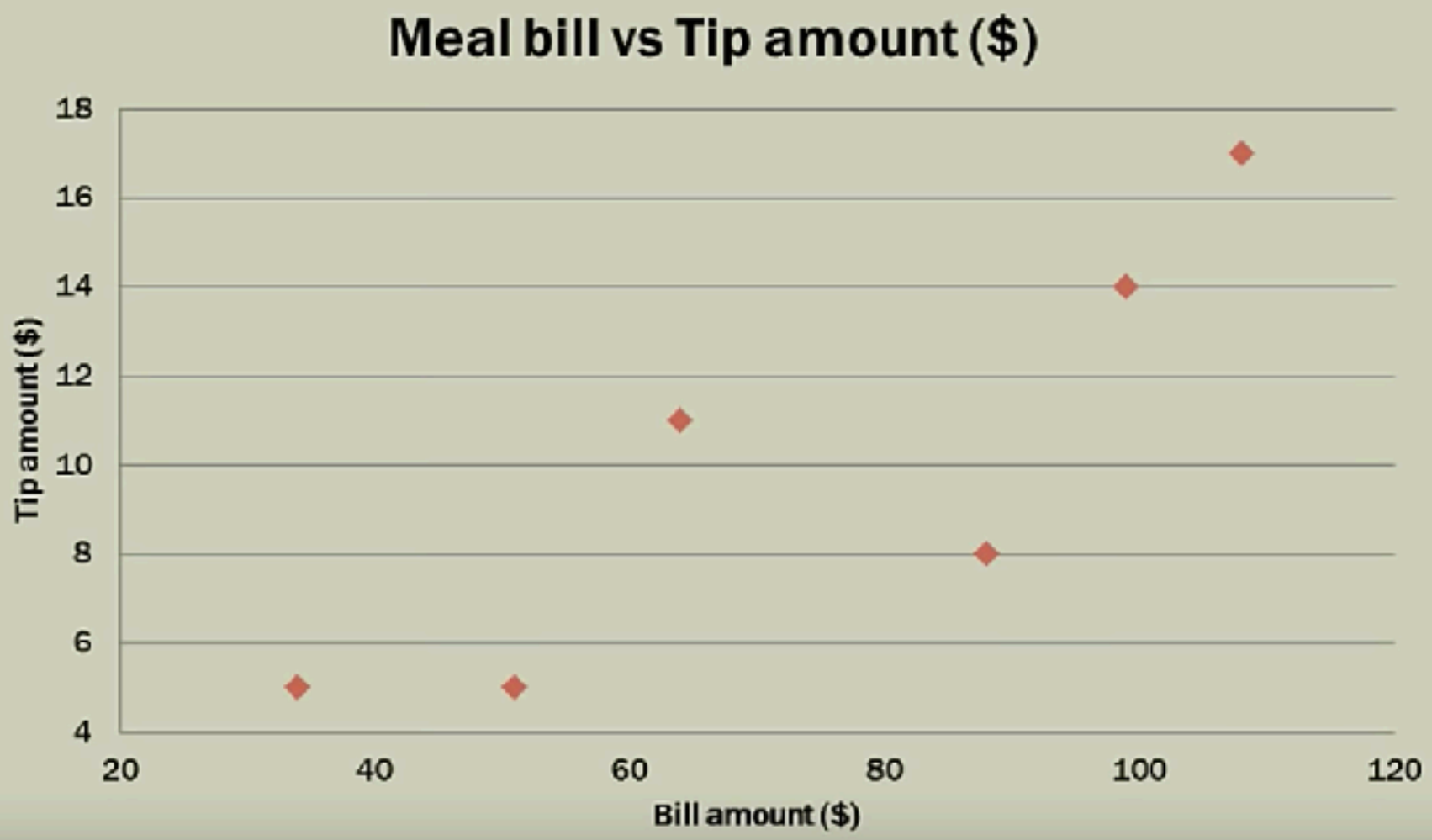


Does the data seem to fall along a line?

In this case, YES! Proceed.

If not...if it's a BLOB with no linear pattern, then stop.

STEP 3: CORRELATION (OPTIONAL)



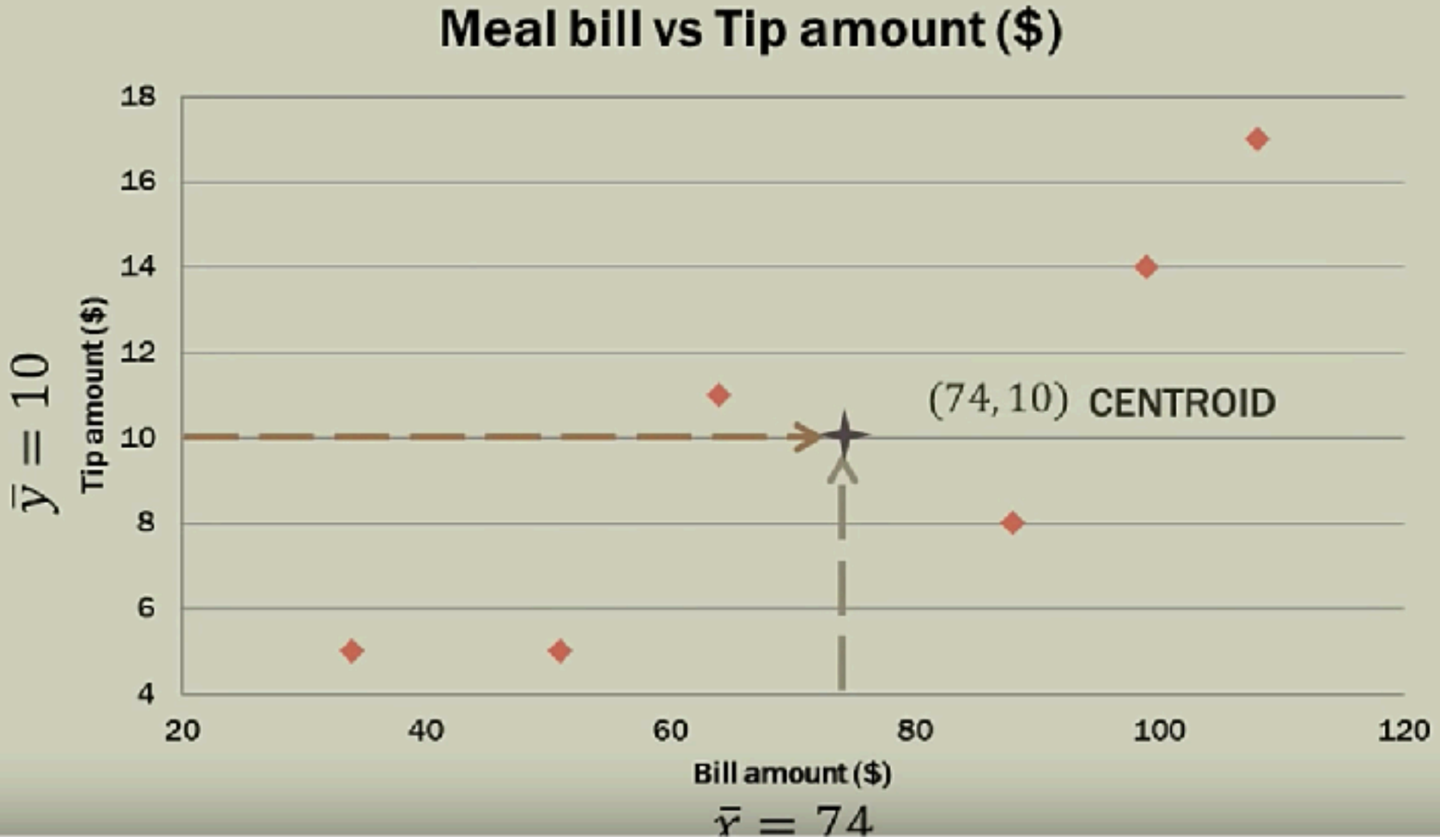
What is the correlation coefficient, r?

In this case, $r = 0.866$.

Is the relationship strong?

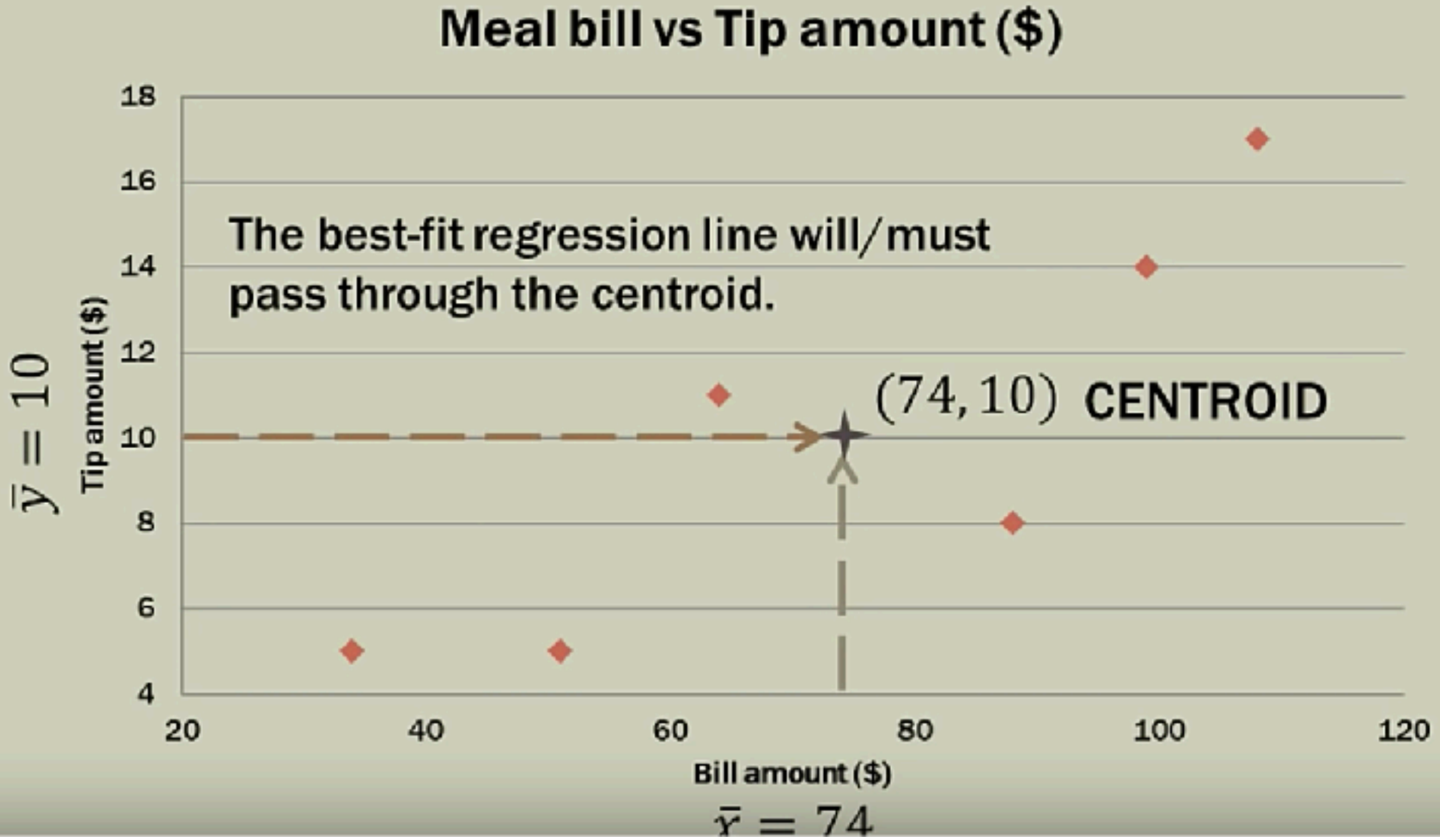
In this case, YES!

STEP 4: DESCRIPTIVE STATISTICS / CENTROID



Bill ($)	Tip ($)
34.00	5.00
108.00	17.00
64.00	11.00
88.00	8.00
99.00	14.00
51.00	5.00
$\bar{x} = 74$	$\bar{y} = 10$

STEP 4: DESCRIPTIVE STATISTICS / CENTROID



Bill ($)	Tip ($)
34.00	5.00
108.00	17.00
64.00	11.00
88.00	8.00
99.00	14.00
51.00	5.00
$\bar{x} = 74$	$\bar{y} = 10$

STEP 5: CALCULATIONS

Intercept $\hat{y}_i = b_0 + b_1 x_i$ **Slope**

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

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$\bar{x}$ = mean of the independent variable

$\bar{y}$ = mean of the dependent variable

x_i = value of independent variable

y_i = value of dependent variable

STEP 5: CALCULATIONS

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

1. For each data point.
2. Take the x-value and subtract the mean of x.
3. Take the y-value and subtract the mean of y.
4. Multiply Step 2 and Step 3
5. Add up all of the products.

---

1. For each data point.
2. Take the x-value and subtract the mean of x.
3. Square Step 2
4. Add up all the products.

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

STEP 5: CALCULATIONS

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

1. For each data point.
2. Take the x-value and subtract the mean of x.
3. Take the y-value and subtract the mean of y.
4. Multiply Step 2 and Step 3
5. Add up all of the products.

---

1. For each data point.
2. Take the x-value and subtract the mean of x.
3. Square Step 2
4. Add up all the products.

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

STEP 5: CALCULATIONS

Meal	Total bill ($)	Tip amount ($)	Bill deviation	Tip Deviations	Deviation Products	Bill Deviation Squared
	x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	34	5				
2	108	17				
3	64	11				
4	88	8				
5	99	14				
6	51	5				
	$\bar{x} = 74$	$\bar{y} = 10$				

STEP 5: CALCULATIONS

Meal	Total bill ($)	Tip amount ($)	Bill deviation	Tip Deviations	Deviation Products	Bill Deviation Squared
	x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	34	5	-40			
2	108	17	34			
3	64	11	-10			
4	88	8	14			
5	99	14	25			
6	51	5	-23			
	$\bar{x} = 74$	$\bar{y} = 10$				

STEP 5: CALCULATIONS

Meal	Total bill ($)	Tip amount ($)	Bill deviation	Tip Deviations	Deviation Products	Bill Deviation Squared
	x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	34	5	-40	-5		
2	108	17	34	7		
3	64	11	-10	1		
4	88	8	14	-2		
5	99	14	25	4		
6	51	5	-23	-5		
	$\bar{x} = 74$	$\bar{y} = 10$				

STEP 5: CALCULATIONS

Meal	Total bill ($)	Tip amount ($)	Bill deviation	Tip Deviations	Deviation Products	Bill Deviation Squared
	x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	34	5	-40	-5	200	1600
2	108	17	34	7	238	1156
3	64	11	-10	1	-10	100
4	88	8	14	-2	-28	196
5	99	14	25	4	100	625
6	51	5	-23	-5	115	529
	$\bar{x} = 74$	$\bar{y} = 10$			$\sum = 615$	$\sum = 4206$

b_1 CALCULATIONS (SLOPE)

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b_1 = \frac{615}{4206}$$

$$b_1 = 0.1462$$

Deviation Products	Bill Deviations Squared
$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
200	1600
238	1156
-10	100
-28	196
100	625
115	529
$\sum = 615$	$\sum = 4206$

b_0 CALCULATIONS(Y-INTERCEPT)

$$b_0 = \bar{y} - b_1\bar{x} \qquad b_1 = 0.1462$$

$$b_0 = 10 - 0.1462(74)$$

$$b_0 = 10 - 10.8188$$

$$b_0 = -0.8188$$

Total bill ($)	Tip amount ($)
x	y
34	5
108	17
64	11
88	8
99	14
51	5
$\bar{x} = 74$	$\bar{y} = 10$

YOUR REGRESSION LINE

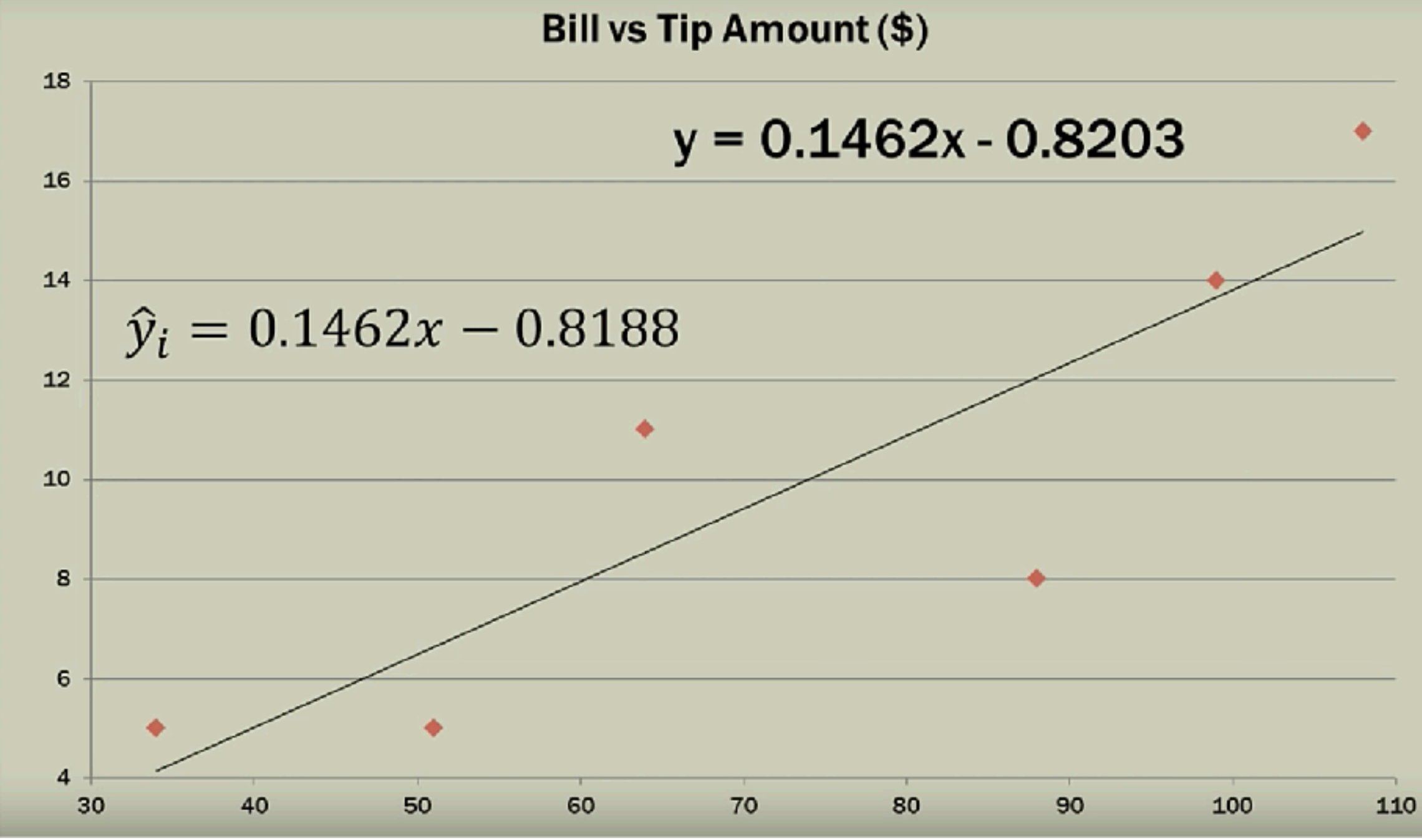
$\hat{y}_i = b_0 + b_1 x_i$ $\quad b_0 = -0.8188$ $\quad b_1 = 0.1462$

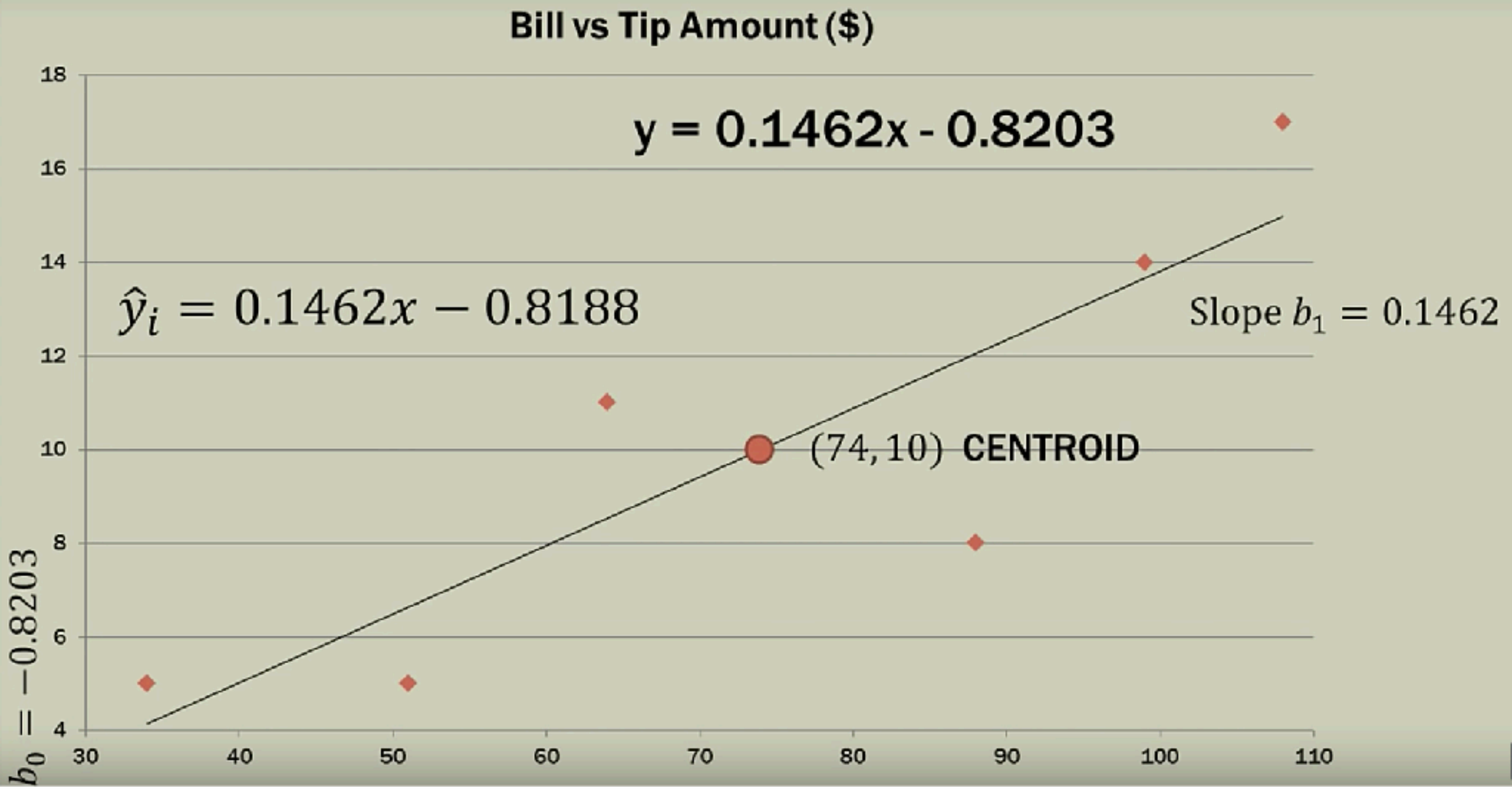
intercept slope

$$\hat{y}_i = -0.8188 + 0.1462x$$

OR

$$\hat{y}_i = 0.1462x - 0.8188$$


Bill vs Tip Amount ($)
y = 0.1462x - 0.8203
$\hat{y}_i = 0.1462x - 0.8188$
18
16
14
12
10
8
6
4
30
40
50
60
70
80
90
100
110


Bill vs Tip Amount ($)
y = 0.1462x - 0.8203
$\hat{y}_i = 0.1462x - 0.8188$
Slope $b_1 = 0.1462$
(74, 10) CENTROID
$b_0 = -0.8203$
18
16
14
12
10
8
6
4
30
40
50
60
70
80
90
100
110

QUICK INTERPRETATION

$$\hat{y}_i = 0.1462x - 0.8188$$

For every \$1 the bill amount (x) increases, we would expect the tip amount to increase by \$0.1462 or about 15-cents.

If the bill amount (x) is zero, then the expected/predicted tip amount is \$-0.8188 or negative 82-cents! Does this make sense? NO. The intercept may or may not make sense in the "real world."

BUT...

IS THIS REGRESSION LINE MODEL ANY GOOD?!?!?

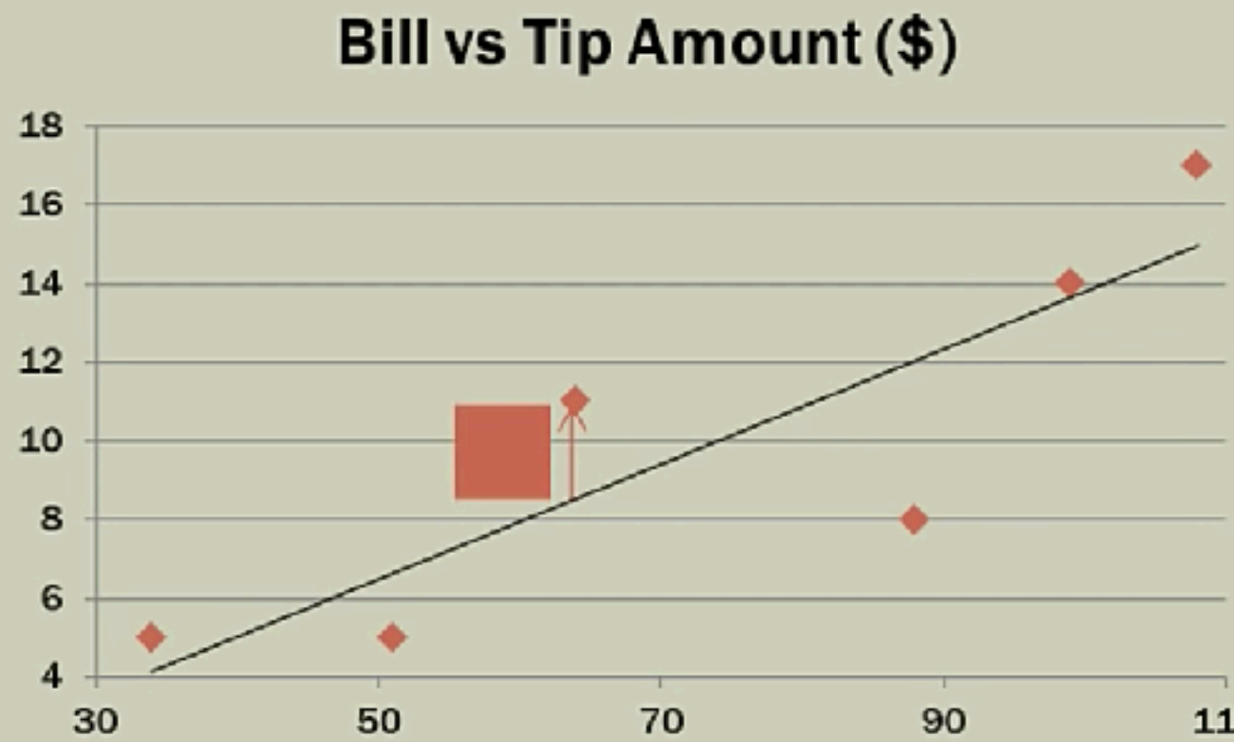
A TALE OF TWO LINES

$SSE = 120$

$SSE = SST$

$SST = 120$

With only the dependent variable, the only sum of squares is due to error. Therefore it is also the total, and **MAXIMUM** sum of squares for the data under analysis.



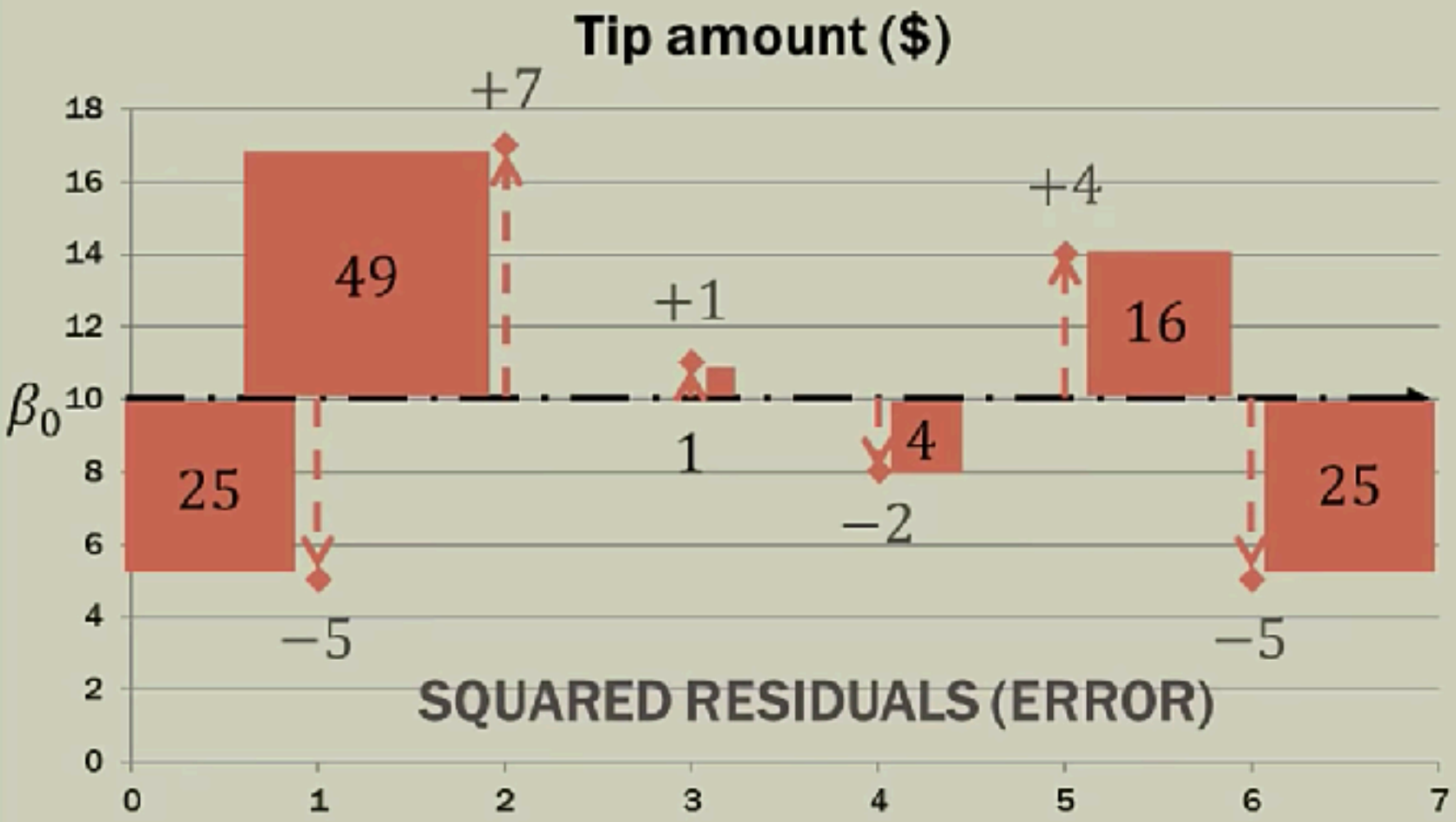
$SST = 120$

$SSE = ?$

$SST - SSE = SSR$

With both the IV and DV, the total sum of squares remains the same. But (ideally) the error sum of squares will be reduced significantly. The difference between SST and SSE is due to regression, SSR.

WHEN THE SLOPE, $\beta_1 = 0$



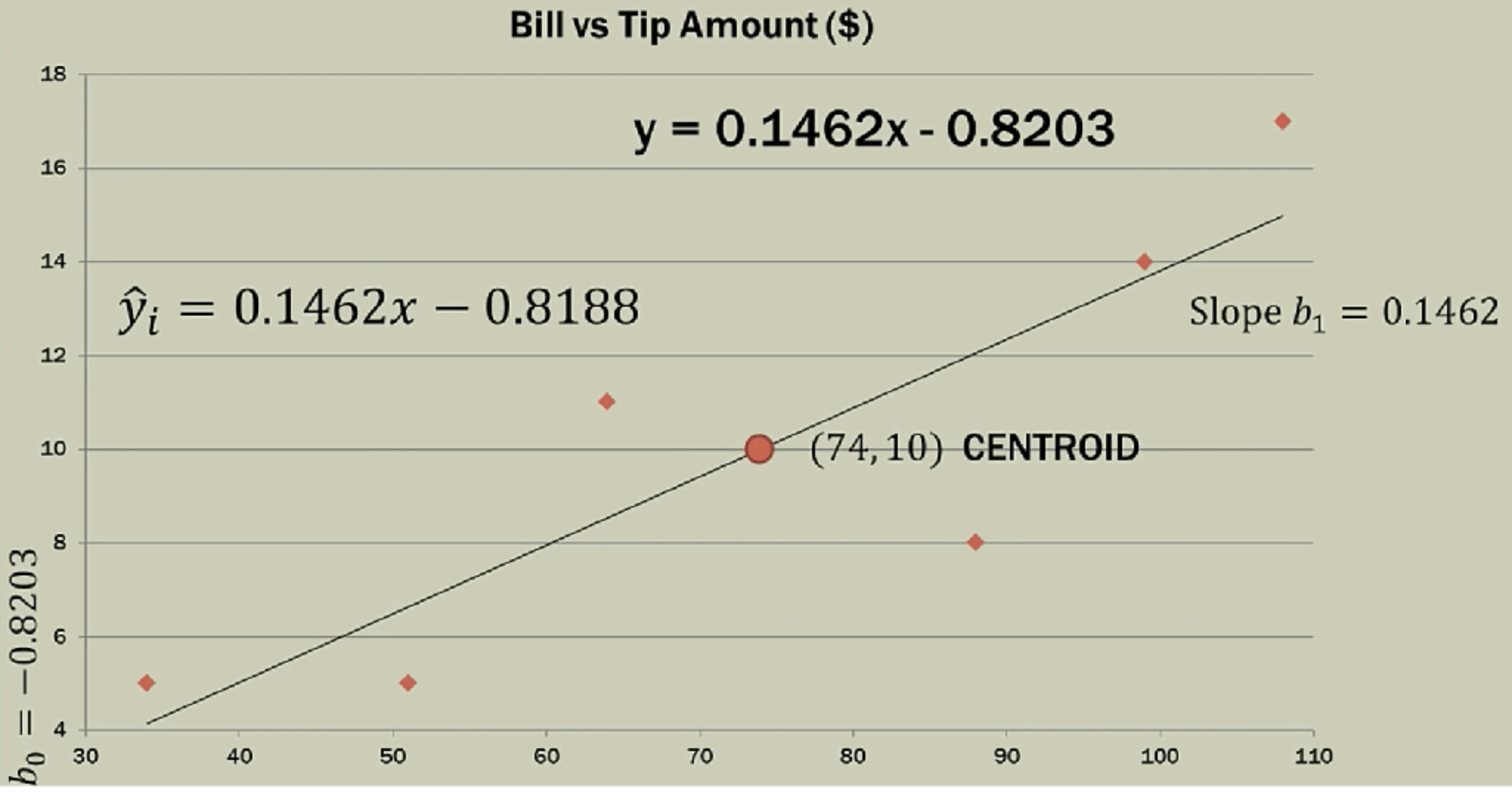
Having only the DV, the best prediction for the tip of the next meal is the mean of the tips.

$\hat{y} = 10$

Since the "mean" line is flat, its slope is zero.

$\beta_1 = 0$

$Sum\ of\ squared\ errors\ (SSE) = 120$


Bill vs Tip Amount ($)
y = 0.1462x - 0.8203
$\hat{y}_i = 0.1462x - 0.8188$
Slope $b_1 = 0.1462$
(74, 10) CENTROID
$b_0 = -0.8203$
18
16
14
12
10
8
6
4
30
40
50
60
70
80
90
100
110

ESTIMATED REGRESSION VALUES

Meal	Observed Total bill ($)	Observed Tip amount ($)	$\hat{y}_i = 0.1462x - 0.8188$	$\hat{y}_i$ (predicted tip amount)
	x	y		
1	34	5		
2	108	17		
3	64	11		
4	88	8		
5	99	14		
6	51	5		
	$\bar{x} = 74$	$\bar{y} = 10$		

ESTIMATED REGRESSION VALUES

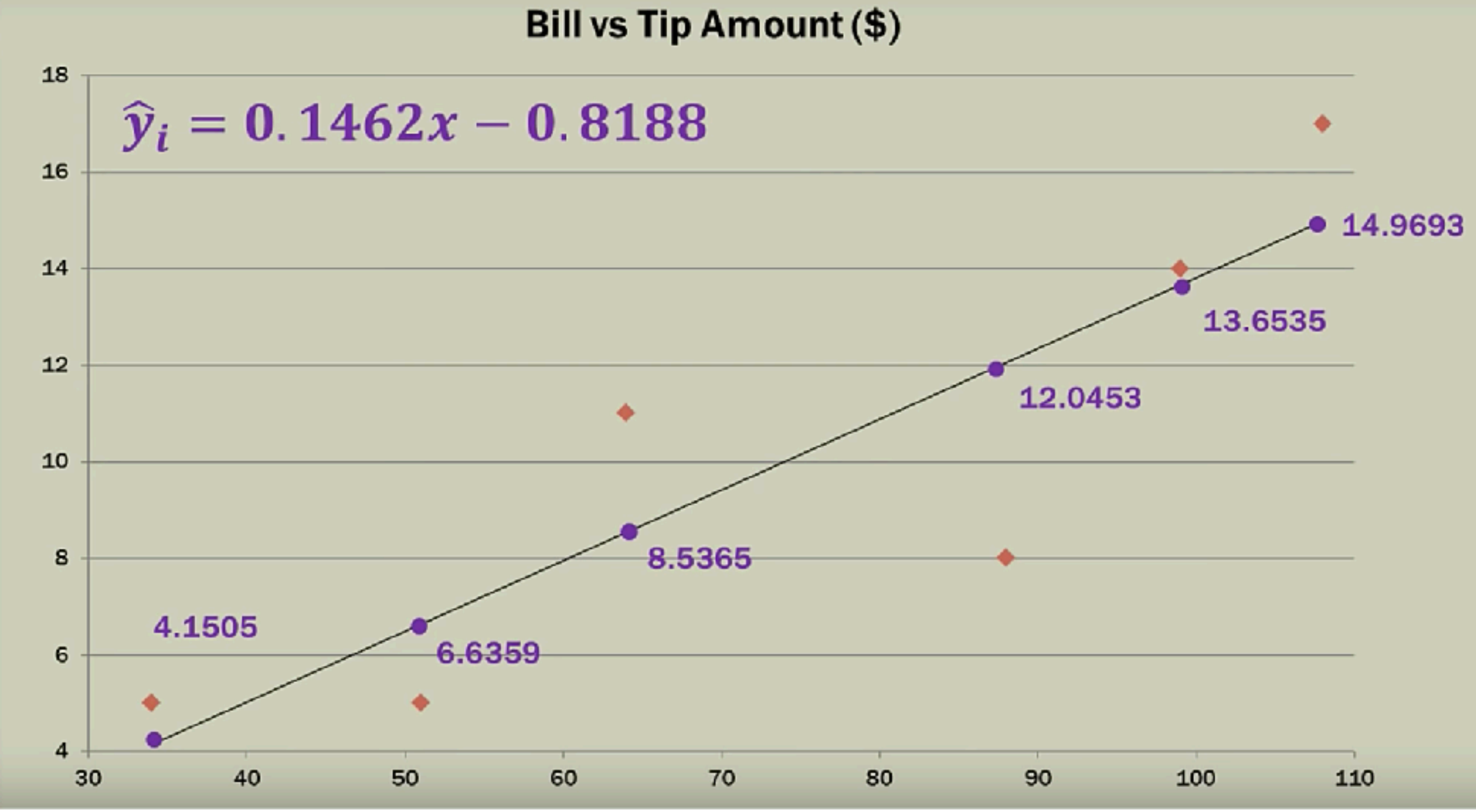
Meal	Total bill ($)	Tip amount ($)	$\hat{y}_i = 0.1462x - 0.8188$	$\hat{y}_i$ (predicted tip amount)
	x	y		
1	34	5	$\hat{y}_i = 0.1462(34) - 0.8188$	
2	108	17	$\hat{y}_i = 0.1462(108) - 0.8188$	
3	64	11	$\hat{y}_i = 0.1462(64) - 0.8188$	
4	88	8	$\hat{y}_i = 0.1462(88) - 0.8188$	
5	99	14	$\hat{y}_i = 0.1462(99) - 0.8188$	
6	51	5	$\hat{y}_i = 0.1462(51) - 0.8188$	
	$\bar{x} = 74$	$\bar{y} = 10$		

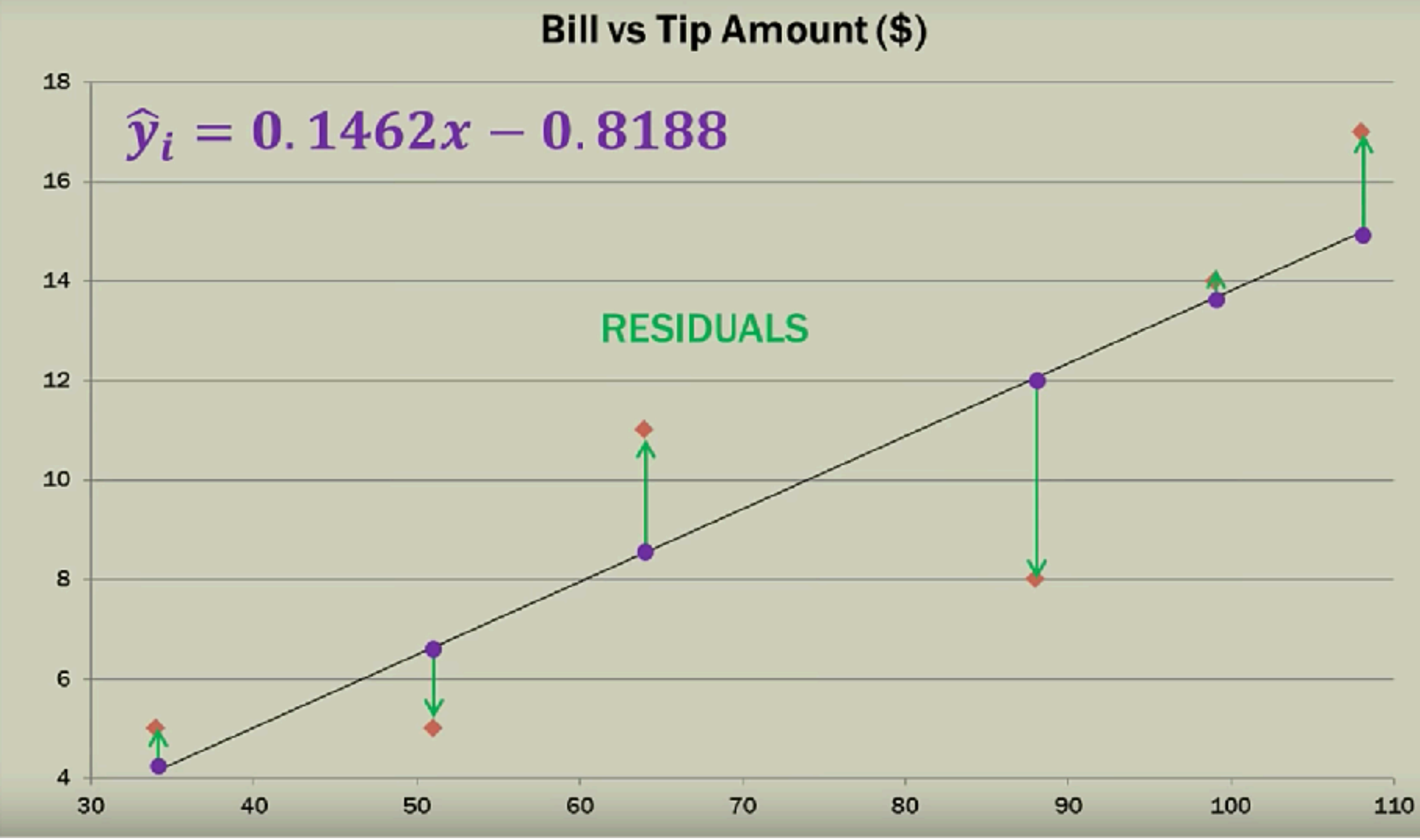
Observed vs. Predicted

ESTIMATED REGRESSION VALUES

Meal	Total bill ($)	Tip amount ($)	$\hat{y}_i = 0.1462x - 0.8188$	$\hat{y}_i$ (predicted tip amount)
	x	y		
1	34	5	$\hat{y}_i = 0.1462(34) - 0.8188$	4.1505
2	108	17	$\hat{y}_i = 0.1462(108) - 0.8188$	14.9693
3	64	11	$\hat{y}_i = 0.1462(64) - 0.8188$	8.5365
4	88	8	$\hat{y}_i = 0.1462(88) - 0.8188$	12.0453
5	99	14	$\hat{y}_i = 0.1462(99) - 0.8188$	13.6535
6	51	5	$\hat{y}_i = 0.1462(51) - 0.8188$	6.6359
	$\bar{x} = 74$	$\bar{y} = 10$		

Observed vs. Predicted


Bill vs Tip Amount ($)
$\hat{y}_i = 0.1462x - 0.8188$
18
16
14
12
10
8
6
4
30
40
50
60
70
80
90
100
110
14.9693
13.6535
12.0453
8.5365
6.6359
4.1505

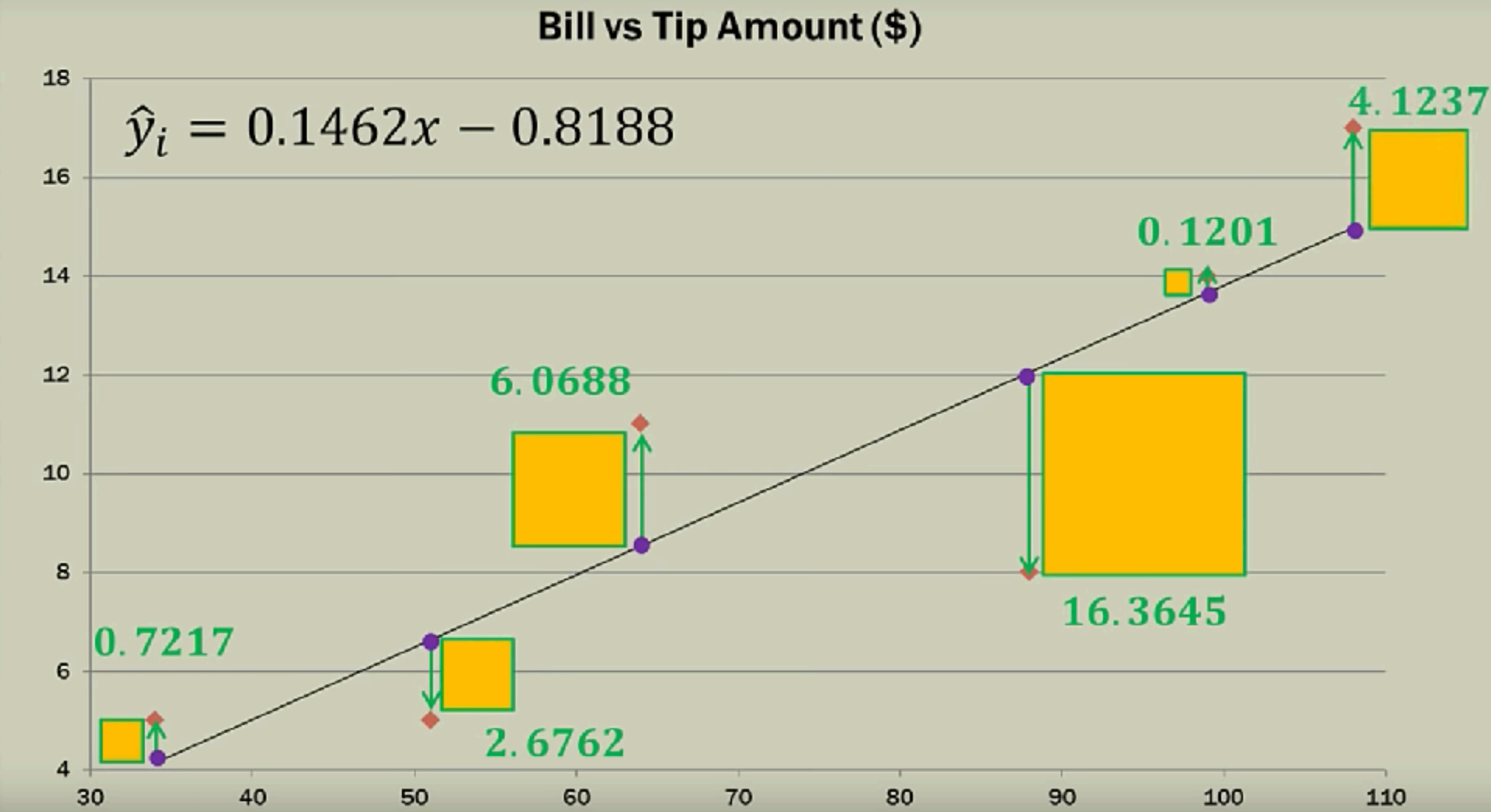

Bill vs Tip Amount ($)
$\hat{y}_i = 0.1462x - 0.8188$
RESIDUALS
18
16
14
12
10
8
6
4
30
40
50
60
70
80
90
100
110

REGRESSION ERROR (RESIDUALS)

Meal	Total bill ($)	Observed tip amount ($)	$\hat{y}_i$ (predicted tip amount)	Error: $(observed - predicted$ $(y - \hat{y}_i)$
	x	y		
1	34	5	4.1505	$5 - 4.1505 = 0.8495$
2	108	17	14.9693	$17 - 14.9693 = 2.0307$
3	64	11	8.5365	$11 - 8.5365 = 2.4635$
4	88	8	12.0453	$8 - 12.0453 = -4.0453$
5	99	14	13.6535	$14 - 13.6535 = 0.3465$
6	51	5	6.6359	$5 - 6.6359 = -1.6359$
	$\bar{x} = 74$	$\bar{y} = 10$		

REGRESSION SQUARED ERROR (RESIDUALS)

Meal	Total bill ($)	Observed tip amount ($)	$\hat{y}_i$ (predicted tip amount)	Error $(y - \hat{y}_i)$	Squared Error $(y - \hat{y}_i)^2$
	x	y			
1	34	5	4.1505	0.8495	0.7217
2	108	17	14.9693	2.0307	4.1237
3	64	11	8.5365	2.4635	6.0688
4	88	8	12.0453	−4.0453	16.3645
5	99	14	13.6535	0.3465	0.1201
6	51	5	6.6359	−1.6359	2.6762
	$\bar{x} = 74$	$\bar{y} = 10$		SSE=	$\sum = 30.075$

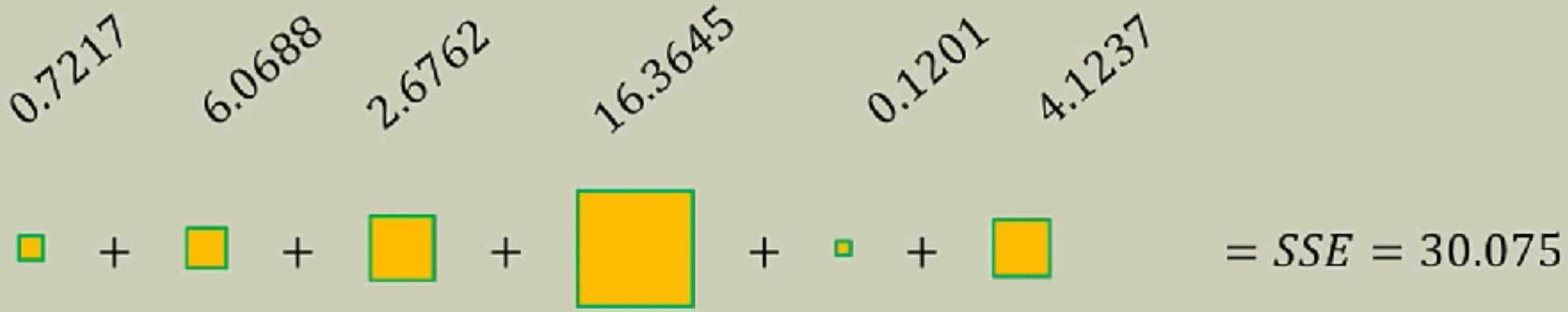

Bill vs Tip Amount ($)
$\hat{y}_i = 0.1462x - 0.8188$
18
16
14
12
10
8
6
4
30
40
50
60
70
80
90
100
110
4.1237
0.1201
6.0688
16.3645
0.7217
2.6762

Sum of Squared Errors Comparison

D.V. (Tip amount) ONLY

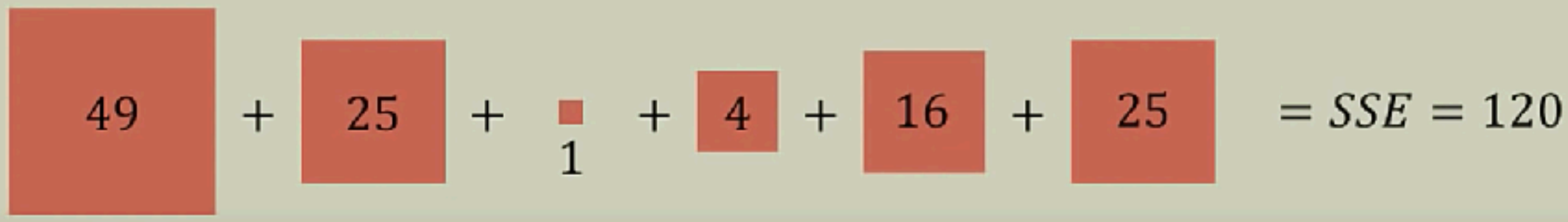
$$49 + 25 + 1 + 4 + 16 + 25 = SSE = 120$$

D.V. and I.V. (Tip amount as a function of meal amount)



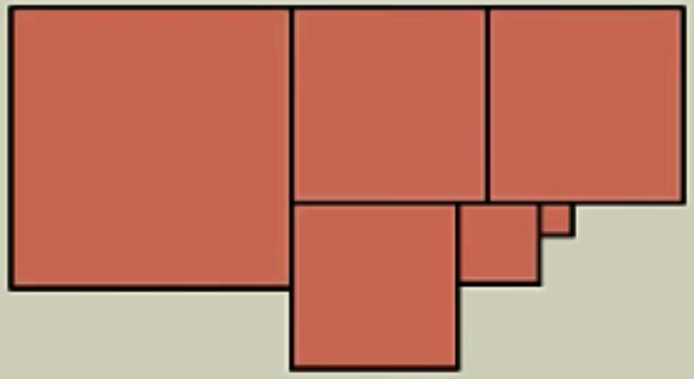
Sum of Squared Errors Comparison

D.V. (Tip amount) ONLY



$= 30.075$

Sum of Squared Errors Comparison



$= 120$

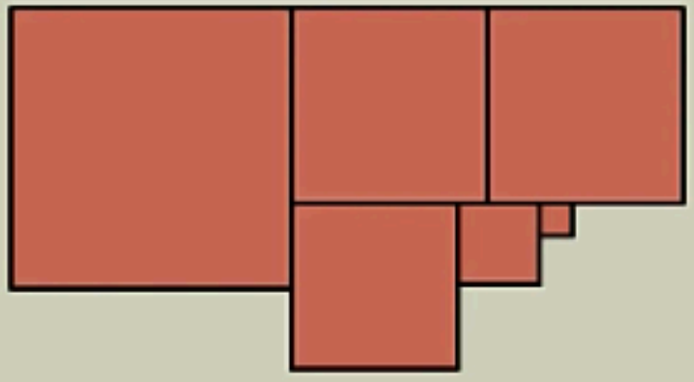
So when we conducted the regression, the SSE decreased from 120 to 30.075. That is, 30.075 of the sum of squares was explained or allocated to ERROR.

Where did the other 89.925 go?

The 89.925 is the sum of squares due to REGESSION.

$= 30.075$

Sum of Squared Errors Comparison



$= 120$

So when we conducted the regression, the SSE decreased from 120 to 30.075. That is, 30.075 of the sum of squares was explained or allocated to ERROR.

Where did the other 89.925 go?

The 89.925 is the sum of squares due to REGESSION.

$$SST = SSR + SSE$$

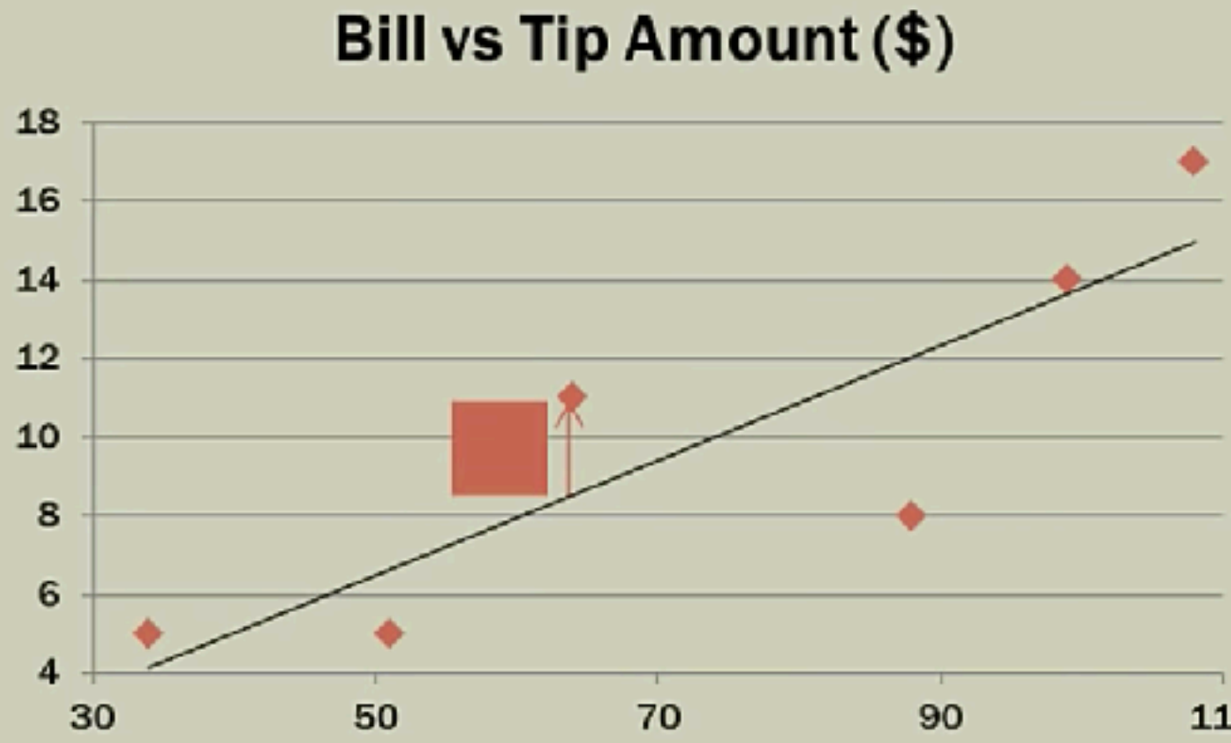
$$120 = 89.925 + 30.075$$

A TALE OF TWO LINES

$SSE = 120$

$SSE = SST$

$SST = 120$



$SST = 120$

$120 - 30.075 = SSR$

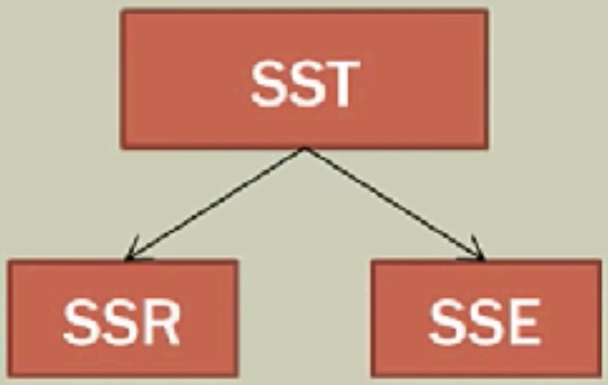
$120 - 30.075 = 89.925$

$SSE = 30.075$

COEFFICIENT OF DETERMINATION

How well does the estimated regression equation fit our data?

This is where regression begins to look a lot like ANOVA; the total sum of squares is partitioned or allocated to SSE and SSR.



If SSR is large, it uses up more of SST and therefor SSE is smaller relative to SST. The coefficient of determination quantifies this ratio as a percentage.

$$Coefficient\ of\ Determination = r^2 = \frac{SSR}{SST}$$

r^2 INTERPRETATION

$$Coefficient\ of\ Determination = r^2 = \frac{SSR}{SST}$$

$$Coefficient\ of\ Determination = r^2 = \frac{89.925}{120}$$

$$Coefficient\ of\ Determination = r^2 = 0.7493\ or\ 74.93\%$$

r^2 INTERPRETATION

$$Coefficient\ of\ Determination = r^2 = \frac{SSR}{SST}$$

GOOD FIT!

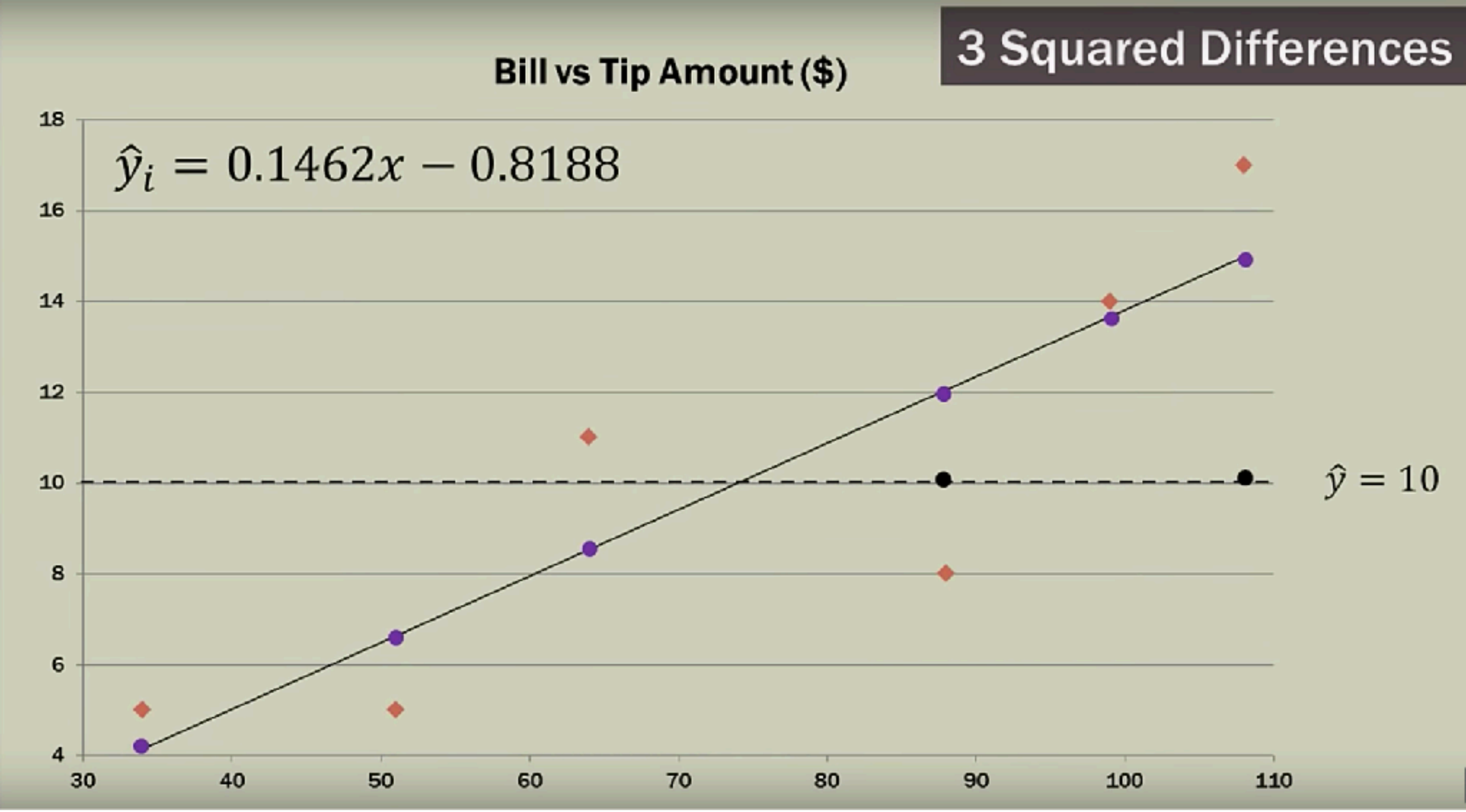
$$Coefficient\ of\ Determination = r^2 = \frac{89.925}{120}$$

$$Coefficient\ of\ Determination = r^2 = 0.7493\ or\ 74.93\%$$

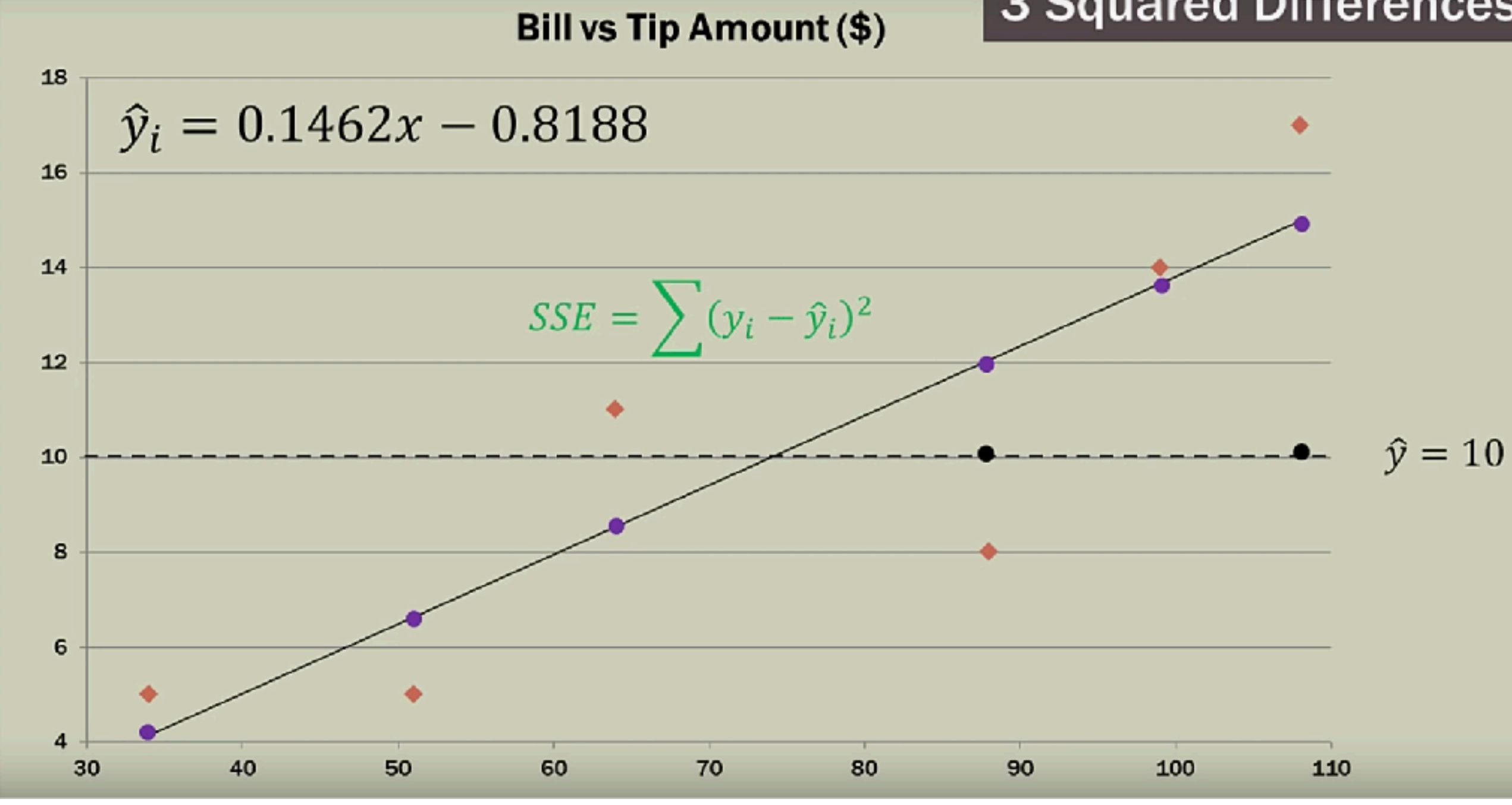
We can conclude that 74.93% of the total sum of squares can be explained by using the estimated regression equation to predict the tip amount. The remainder is error.

$\hat{y}_i = 0.1462x - 0.8188$ **Where x is the dollar amount of the bill.**

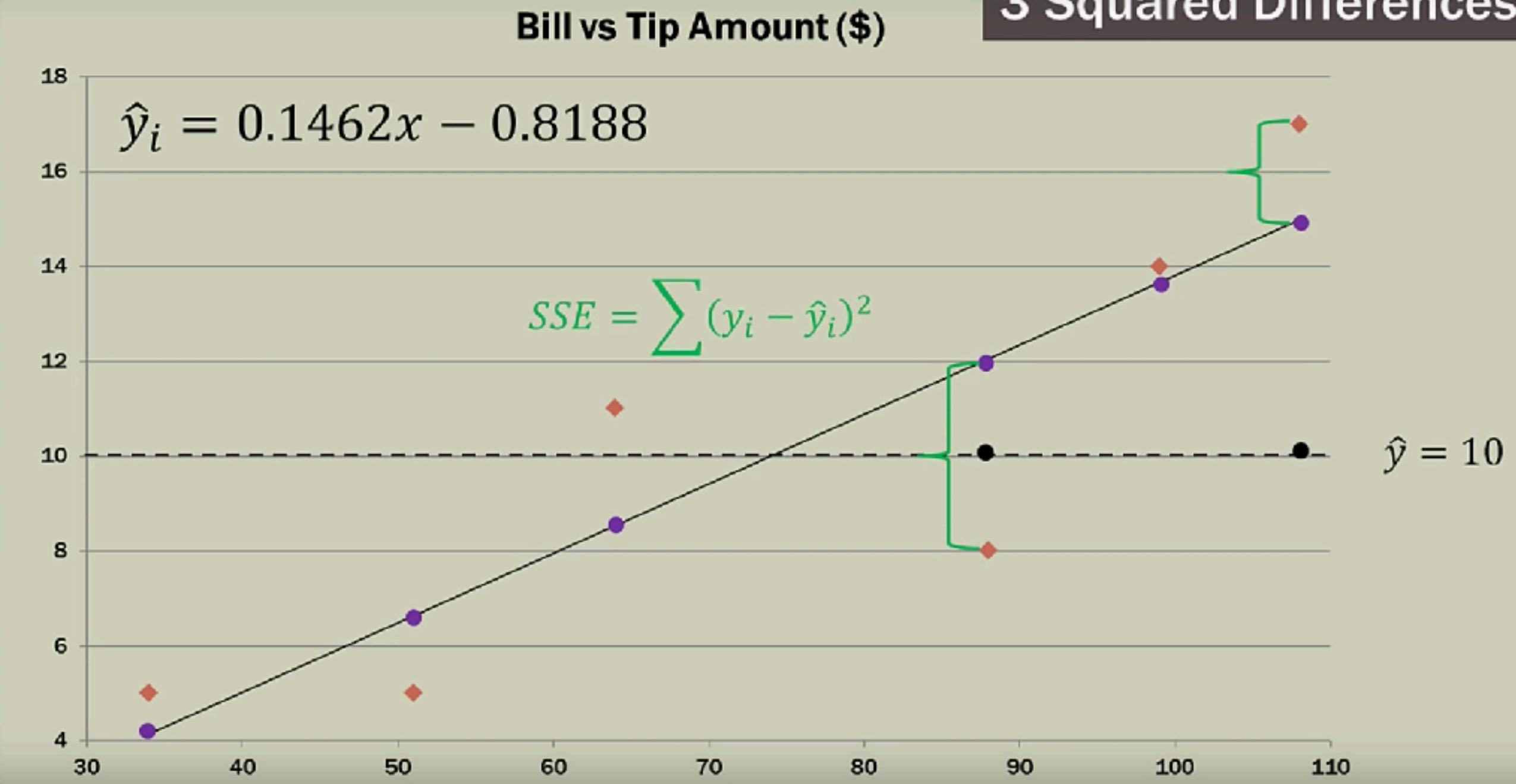
3 Squared Differences



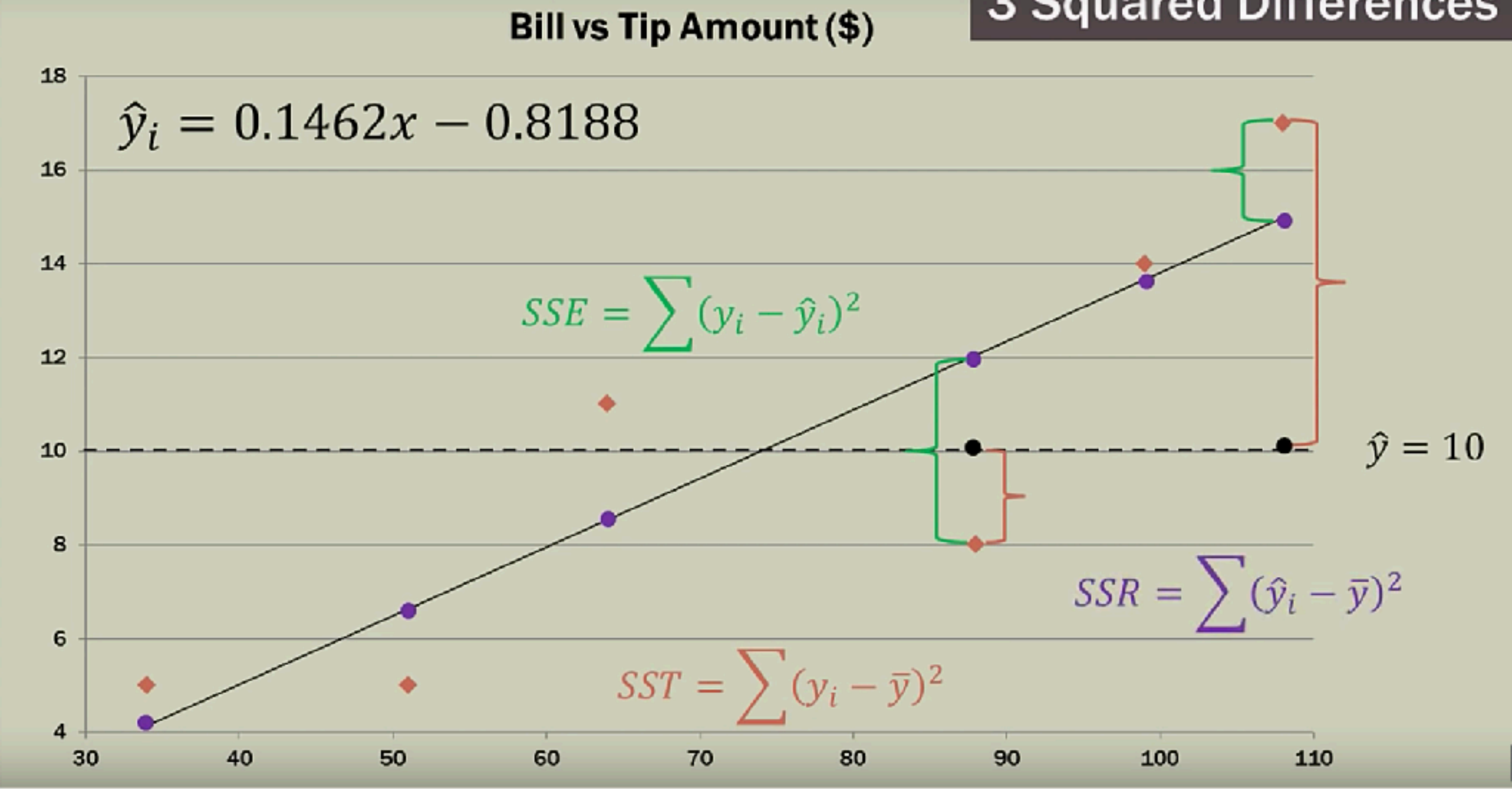
3 Squared Differences



3 Squared Differences



3 Squared Differences



3 Squared Differences