General Linear Regression With

Learning Objectives

1. Describe the Linear Regression Model
2. State the Regression Modeling Steps
3. Explain Ordinary Least Squares
4. Compute Regression Coefficients
5. Understand and check model assumptions
6. Predict Response Variable

Models

What is a Model?

1. Representation of Some Phenomenon

   Non-Math/Stats Model

What is a Math/Stats Model?

1. Often Describe Relationship between Variables

2. Types
   - Deterministic Models (no randomness)
   - Probabilistic Models (with randomness)

Deterministic Models

1. Hypothesize Exact Relationships
2. Suitable When Prediction Error is Negligible
3. Example: Body mass index (BMI) is measure of body fat based

- Metric Formula: $BMI = \frac{Weight\ in\ Kilograms}{(Height\ in\ Meters)^2}$

- Non-metric Formula: $BMI = \frac{Weight\ (pounds) \times 703}{(Height\ in\ inches)^2}$

Probabilistic Models

1. Hypothesize 2 Components
   - Deterministic
   - Random Error
2. Example: Systolic blood pressure of newborns Is 6 Times the Age in days + Random Error
   - $SBP = 6 \times age(d) + \varepsilon$
   - Random Error May Be Due to Factors Other Than age in days (e.g. Birthweight)

Let's look at a problem

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.

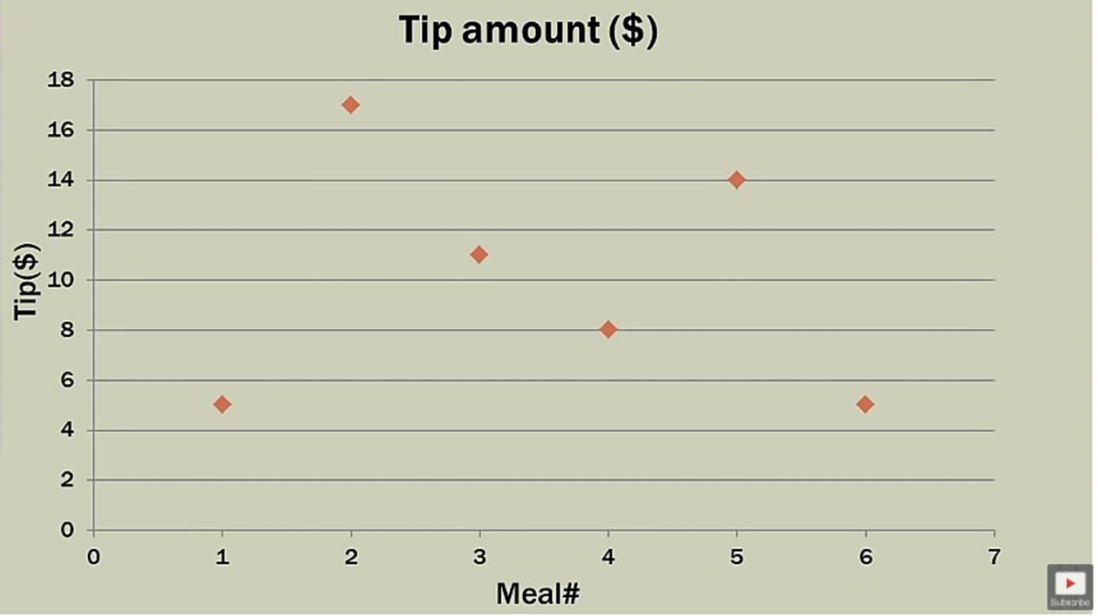
Regression Models

- Relationship between one dependent variable and explanatory variable(s)
- Use equation to set up relationship
  - Numerical Dependent (Response) Variable
  - 1 or More Numerical or Categorical Independent (Explanatory) Variables
- Used Mainly for Prediction & Estimation

Lets look at the tips…

TIPS FOR SERVICE

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



Regression Modeling Steps

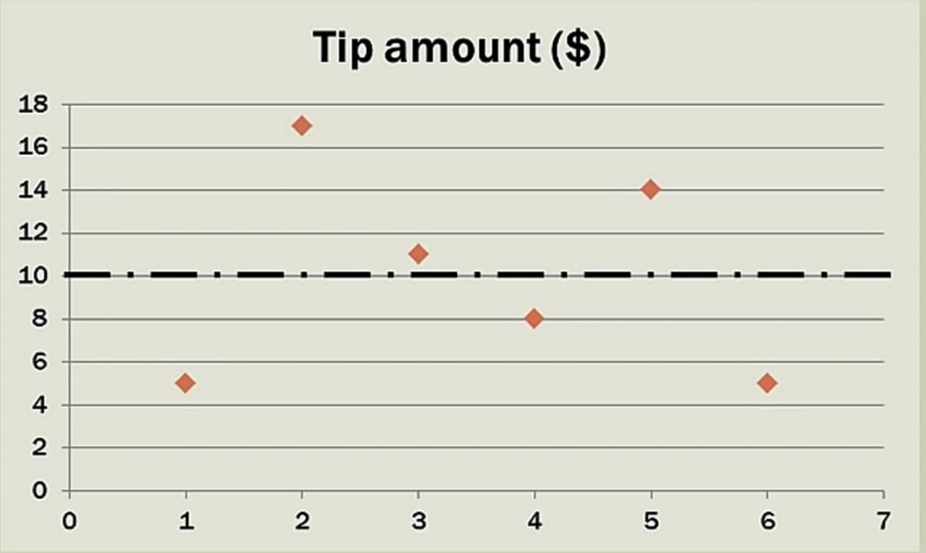
- 1. Hypothesize Deterministic Component
  - Estimate Unknown Parameters
- 2. Specify Probability Distribution of Random Error Term
  - Estimate Standard Deviation of Error
- 3. Evaluate the fitted Model
- 4. Use Model for Prediction & Estimation

What could be a estimate…

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.

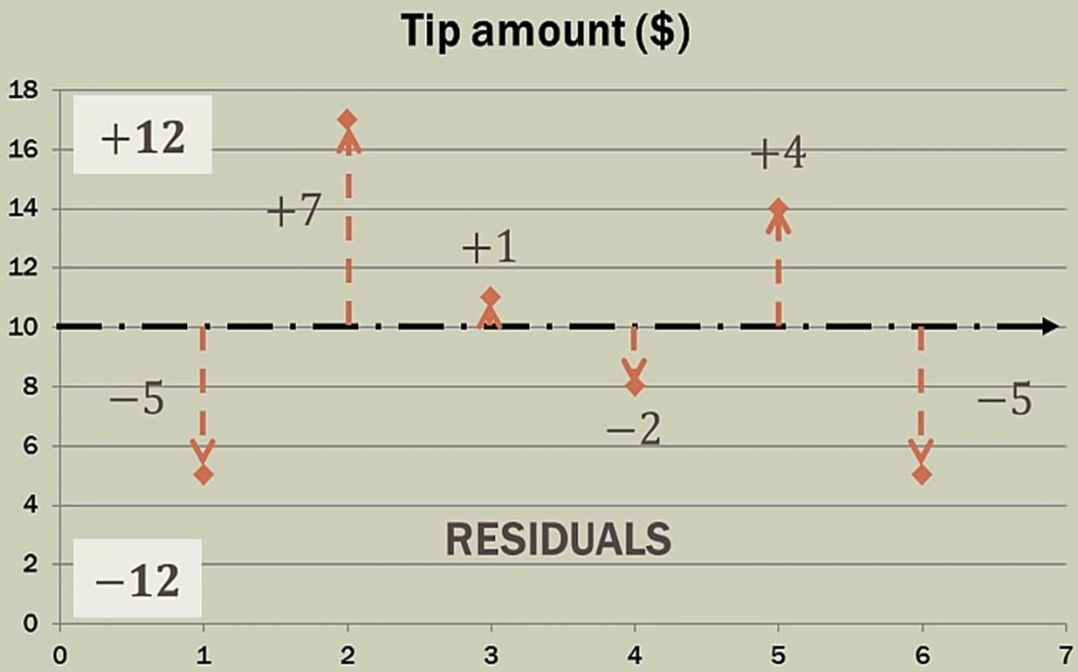
Model Specification

Goodness of estimate…

"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

$\bar{y} = \$10$

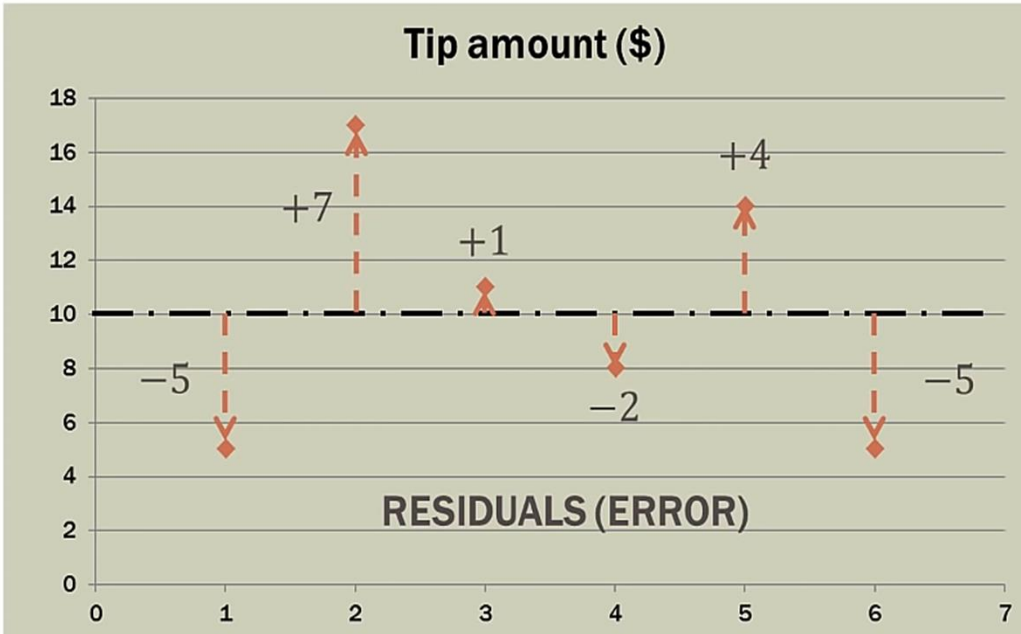


Specifying the deterministic component

- 1. Define the dependent variable and independent variable
- 2. Hypothesize Nature of Relationship
  - Expected Effects (i.e., Coefficients' Signs)
  - Functional Form (Linear or Non-Linear)
  - Interactions

Square of Residuals..

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

Linear Regression Model

Minimise the error

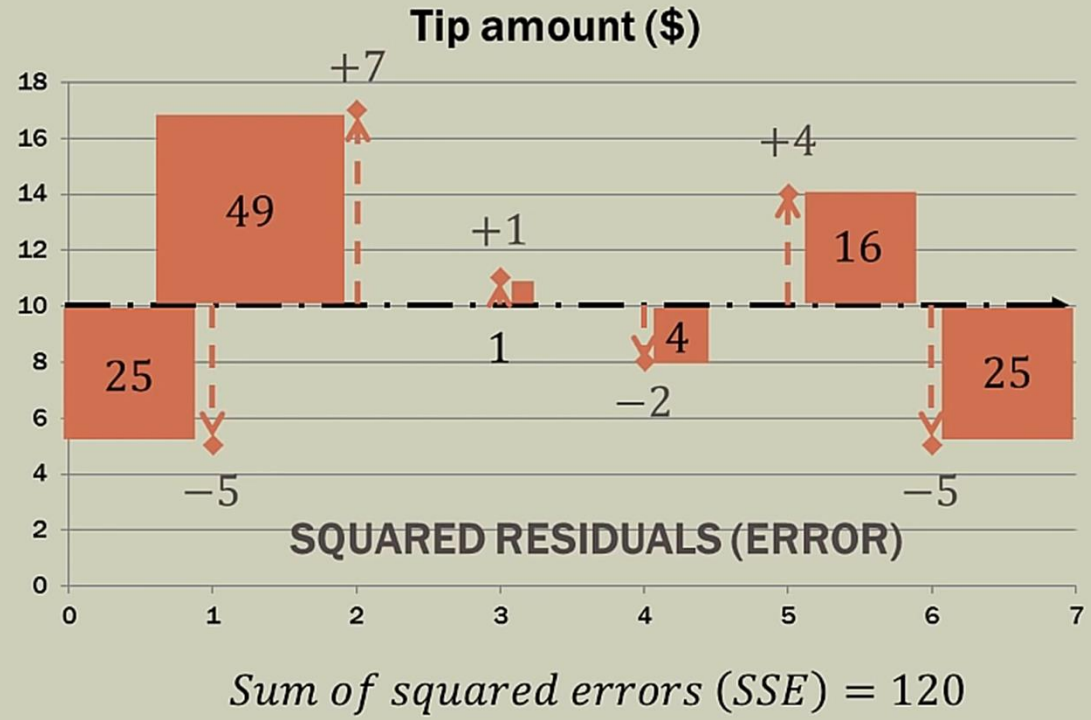
SUM OF SQUARES

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.

Introduction of new Variable

VERY IMPORTANT



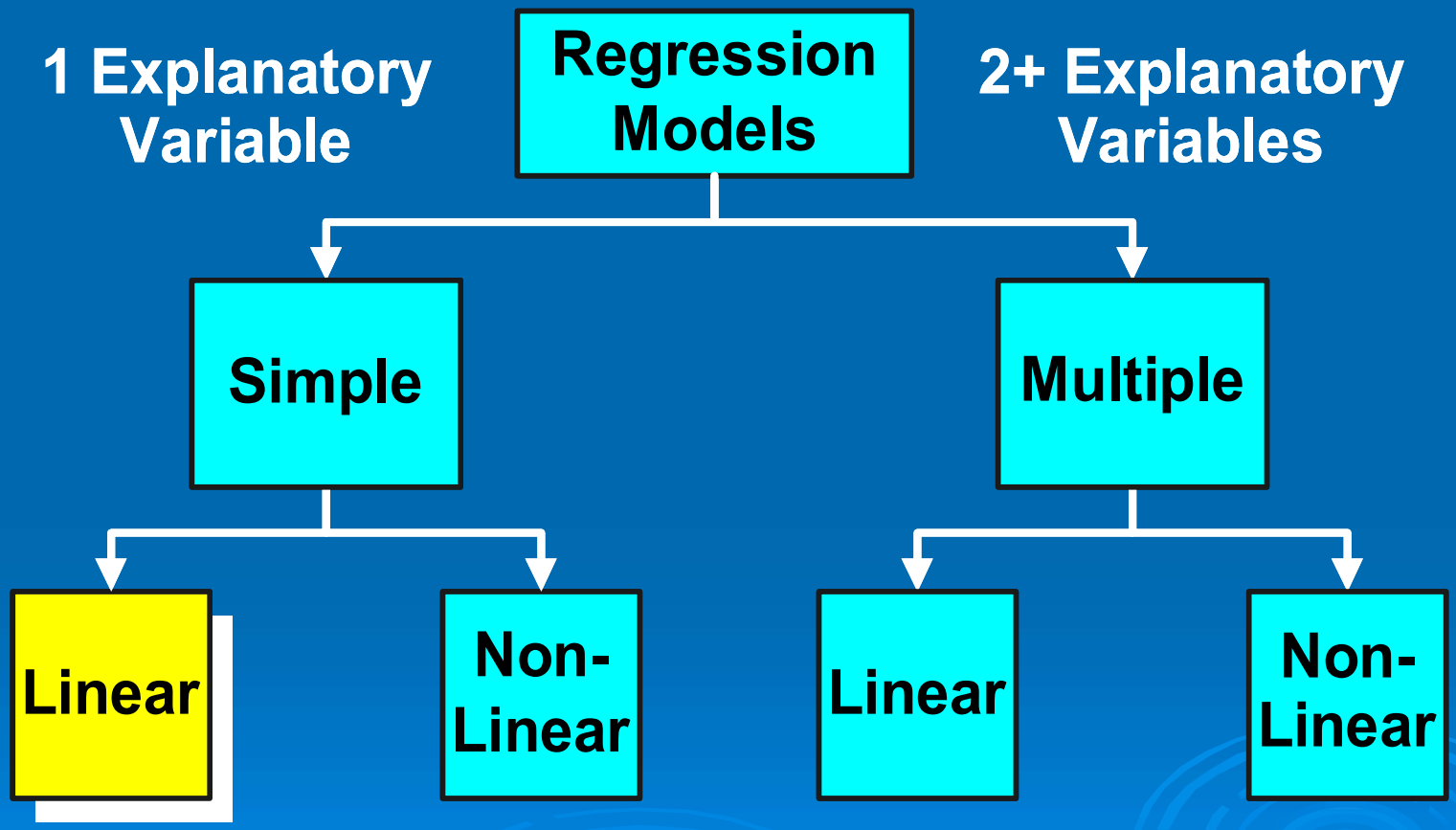
$Sum\ of\ squared\ errors\ (SSE) = 120$

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?

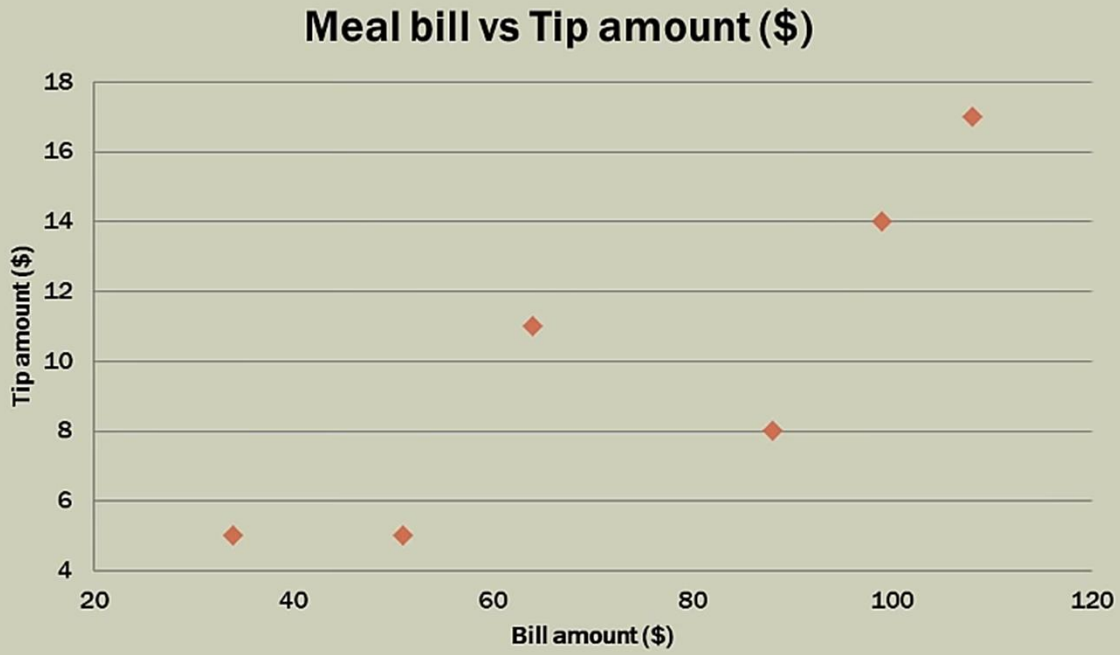
NOTHING.

Types of Regression Models



New Variable

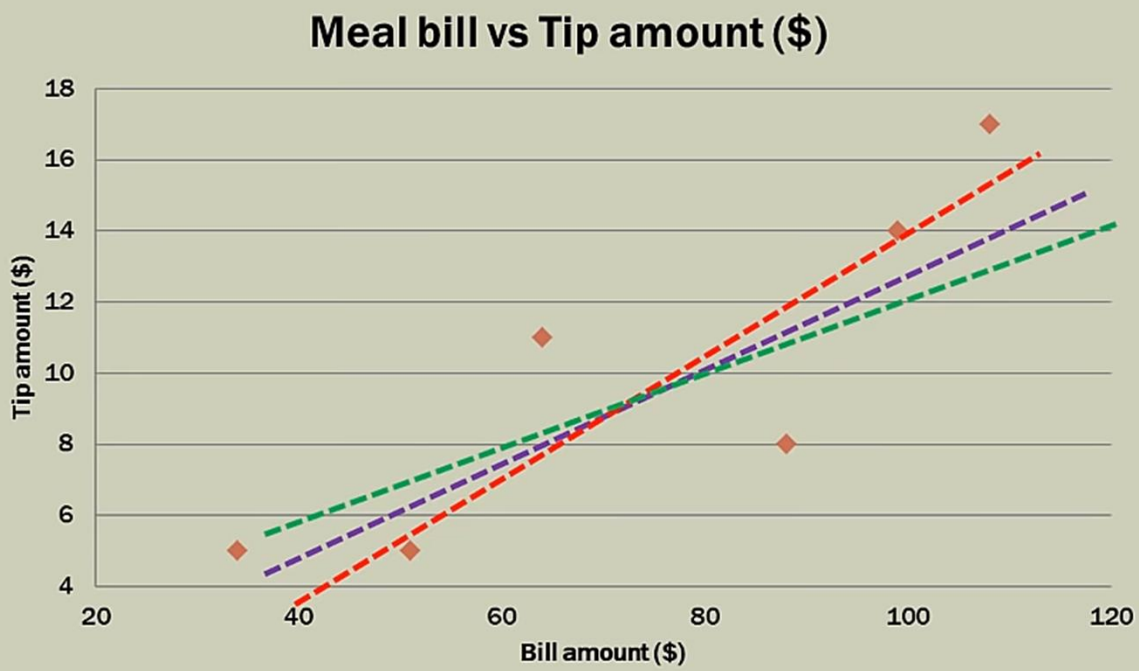
GETTING READY FOR LEAST SQUARES



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

Does it represent a line ?

STEP 2: LOOK FOR A VISUAL LINE

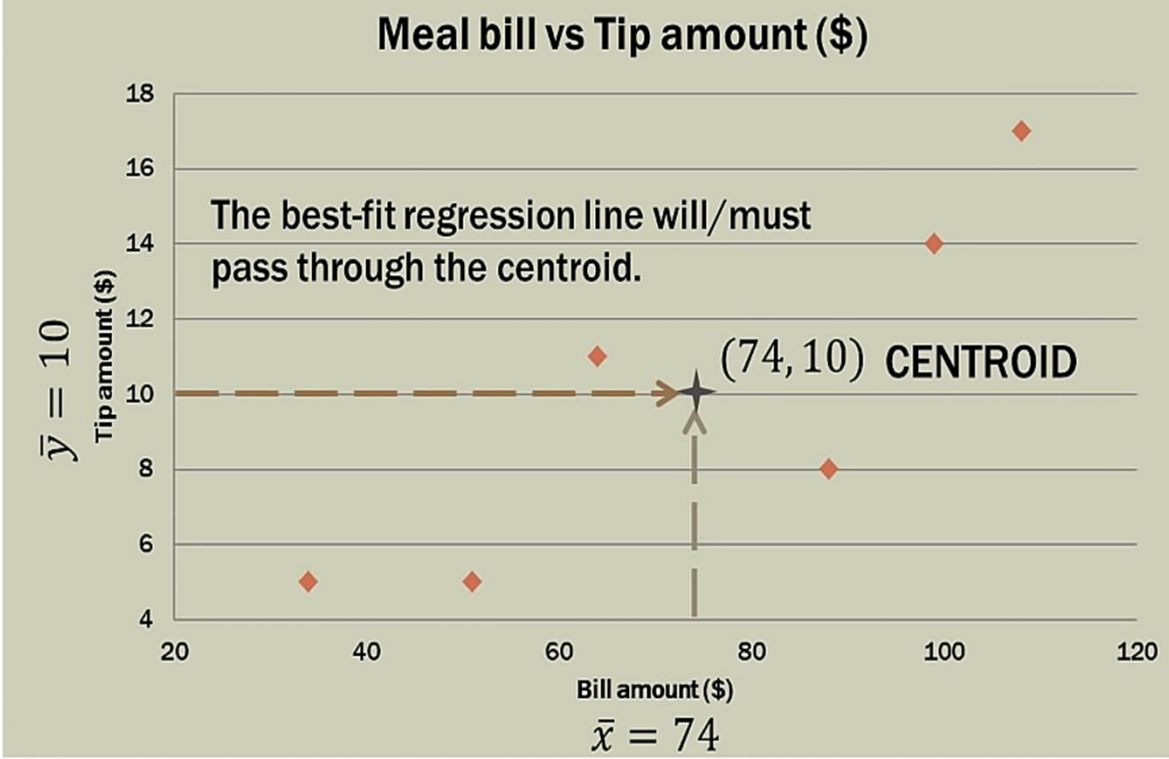


Does the data seem to fall along a line?

In this case, YES! Proceed.

Lets start fitting….centroid.

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
$\bar{x} = 74$	$\bar{y} = 10$

Components of the line…

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

Process to calculate

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.

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

Tabulation process

STEP 5: CALCULATIONS

Meal	Total bill ($)	Tip amount ($)	Bill deviation	Tip Deviations	Deviation Products	Bill Deviations 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
	$\bar{x} = 74$	$\bar{y} = 10$			$\sum = 615$	$\sum = 4206$

Calculation of slope…

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
$\sum = 615$	$\sum = 4206$

Calculation of Intercept

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
$\bar{x} = 74$	$\bar{y} = 10$

Voila…The St. Line Eq

YOUR REGRESSION LINE

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

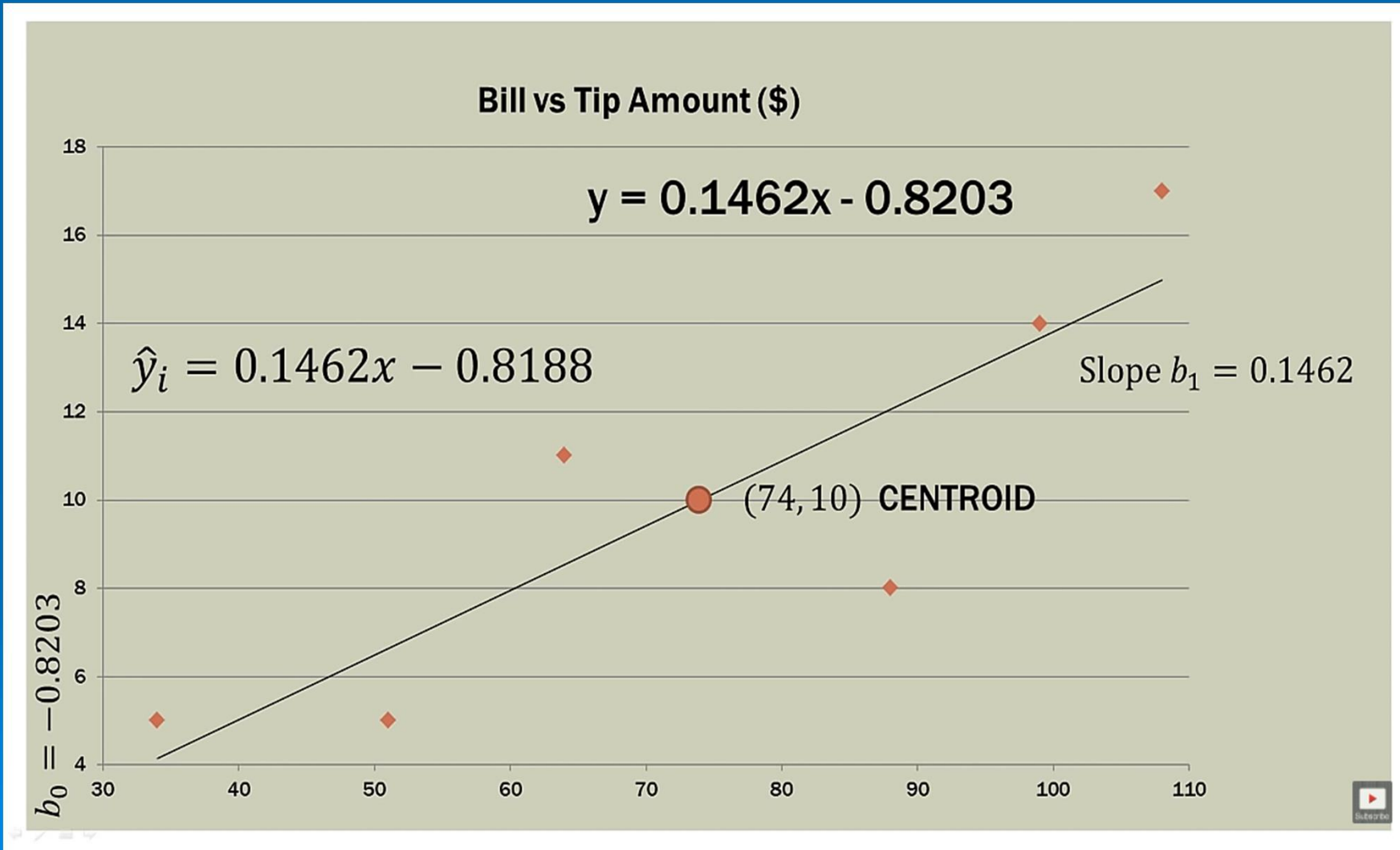
intercept slope

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

OR

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

Graphical representation of our model



Making sense of it…

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."

Is our model better ?

A TALE OF TWO LINES

Tip amount ($)

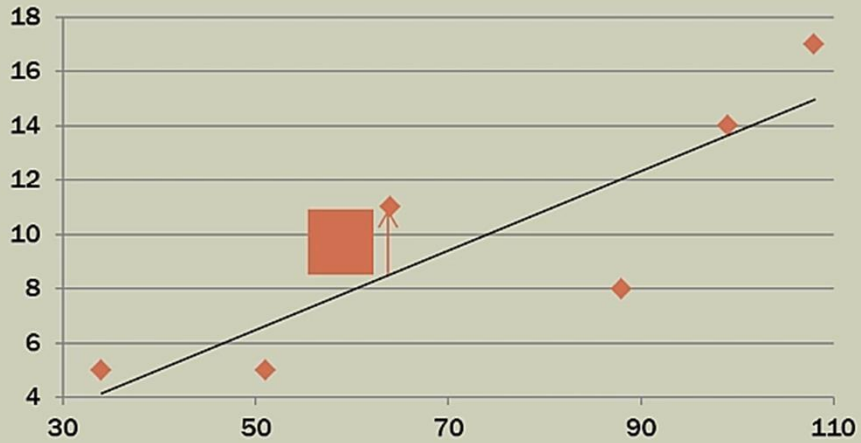
$SSE = 120$

$SSE = SST$

$SST = 120$

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.

Bill vs Tip Amount ($)



$SST = 120$

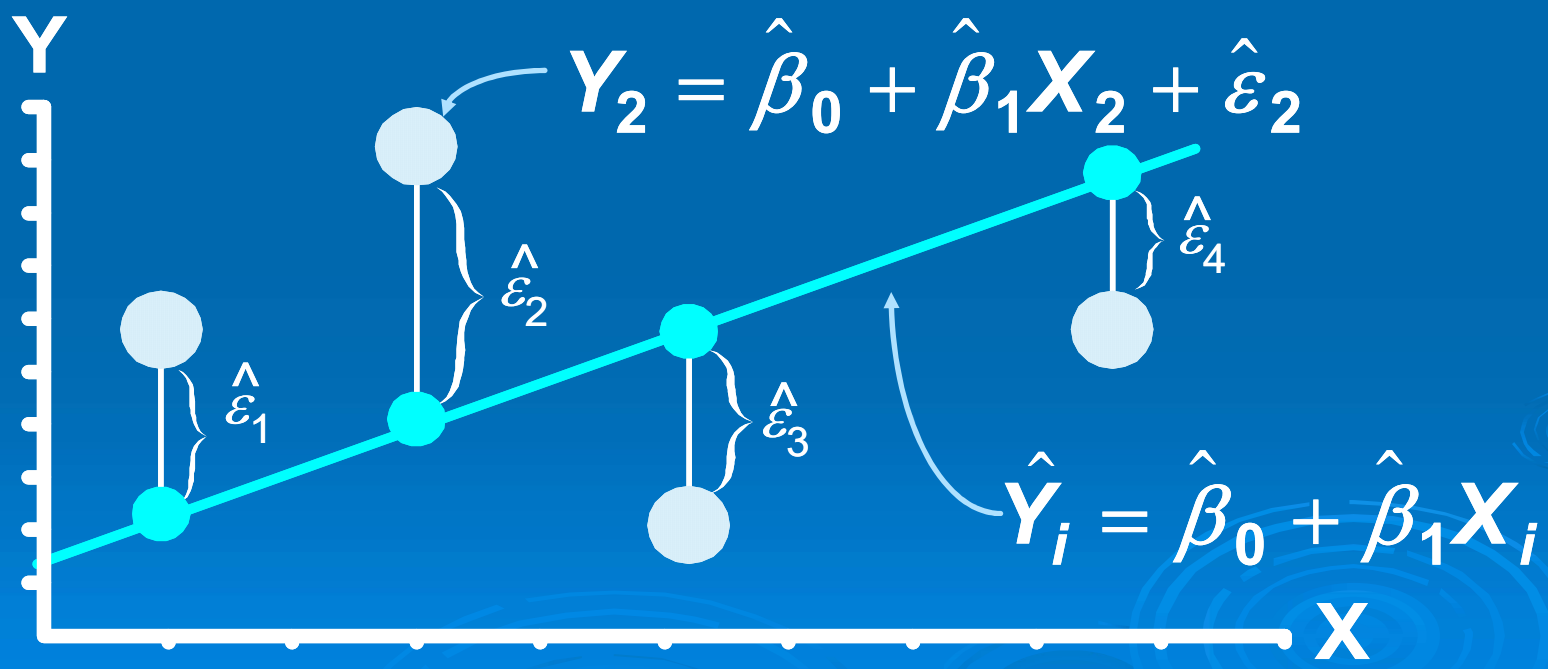
$SSE = ?$

$SST - SSE = SSR$

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.

Least Squares Graphically

LS minimizes $\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + \hat{\varepsilon}_3^2 + \hat{\varepsilon}_4^2$



Error = Observered - Predicted

ESTIMATED REGRESSION VALUES

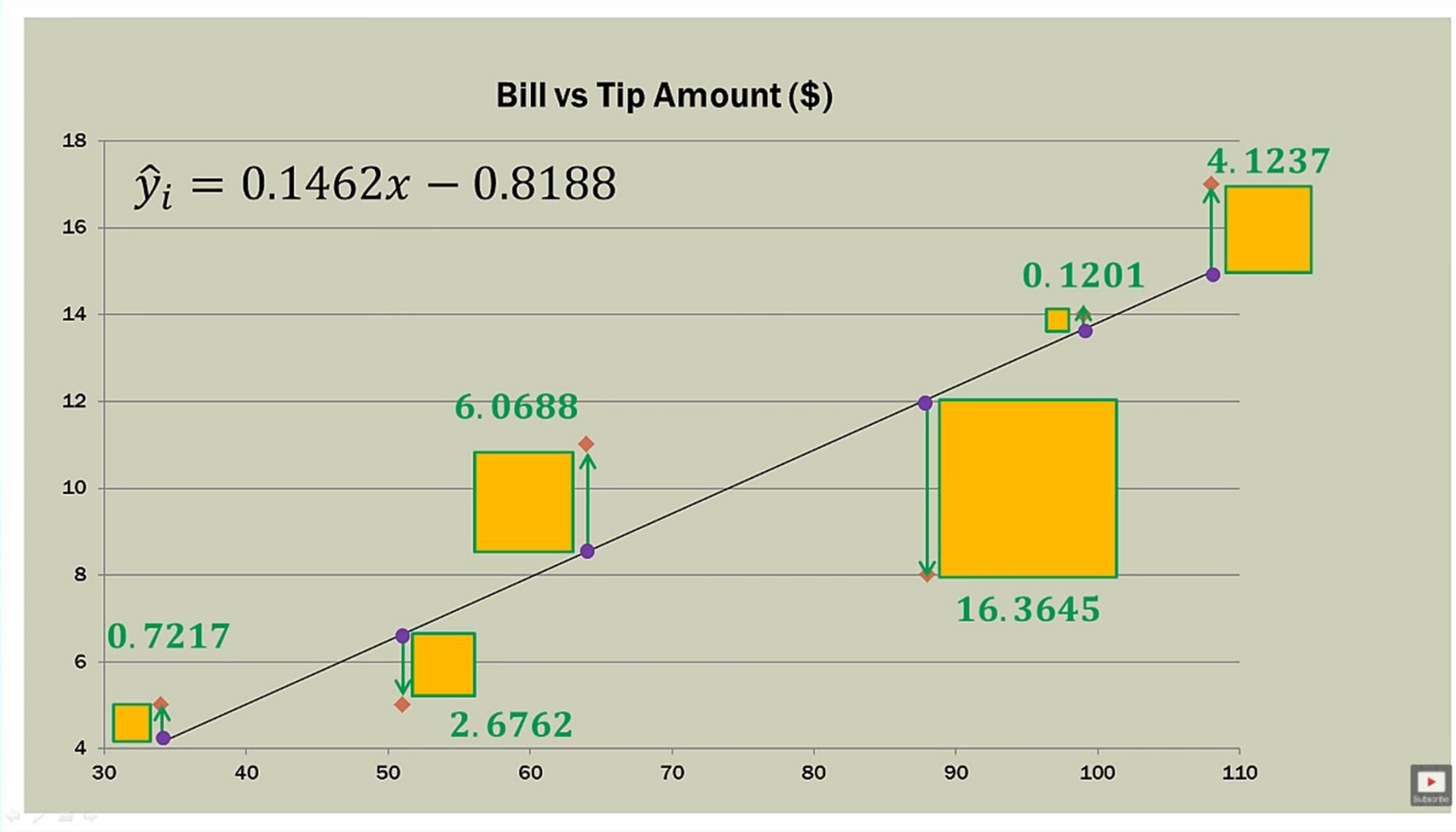
Meal	Total bill ($)	Tip amount ($)	$\hat{y}_i = 0.1462x - 0.8188$	$\hat{y}_i$ (predicted tip amount)
	x	y		
1	34	5	$\hat{y}_i = 0.1462(34) - 0.8188$	4.1505
2	108	17	$\hat{y}_i = 0.1462(108) - 0.8188$	14.9693
3	64	11	$\hat{y}_i = 0.1462(64) - 0.8188$	8.5365
4	88	8	$\hat{y}_i = 0.1462(88) - 0.8188$	12.0453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

The Squared Error

REGRESSION SQUARED ERROR (RESIDUALS)

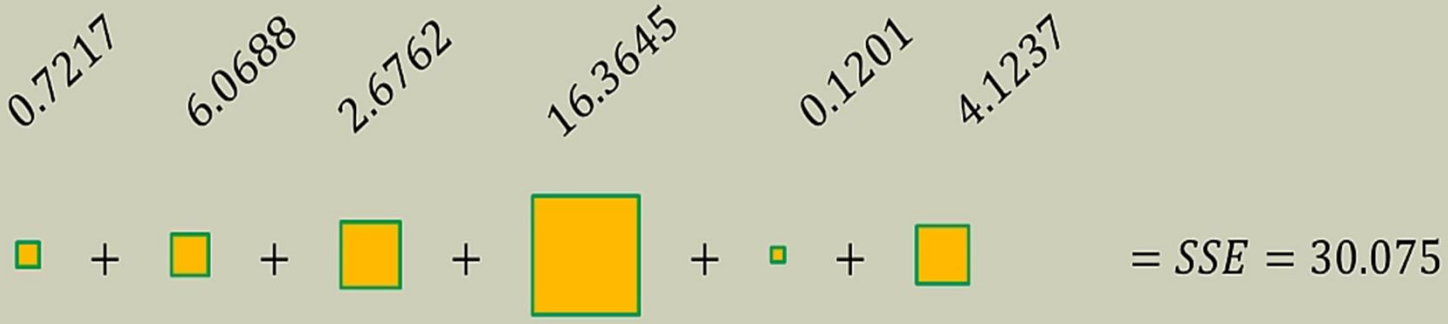
Meal	Total bill ($)	Observed tip amount ($)	$\hat{y}_i$ (predicted tip amount)	Error $(y - \hat{y}_i)$	Squared Error $(y - \hat{y}_i)^2$
	x	y			
1	34	5	4.1505	0.8495	0.7217
2	108	17	14.9693	2.0307	4.1237
3	64	11	8.5365	2.4635	6.0688
4	88	8	12.0453	−4.0453	16.3645
5	99	14	13.6535	0.3465	0.1201
6	51	5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

On the graph…


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
0.7217
2.6762
6.0688
16.3645
0.1201
4.1237

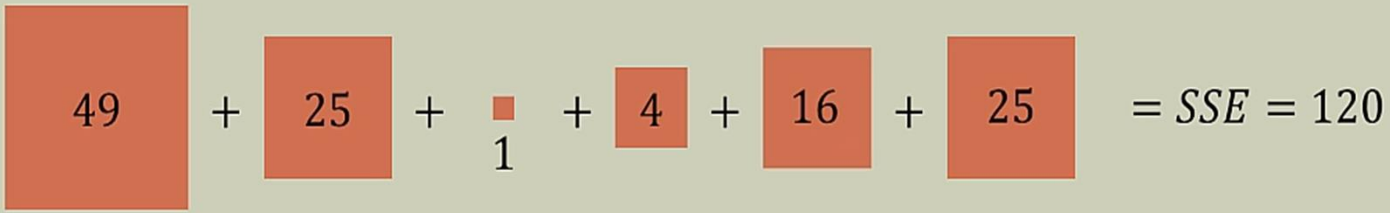
Did error decrease…

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



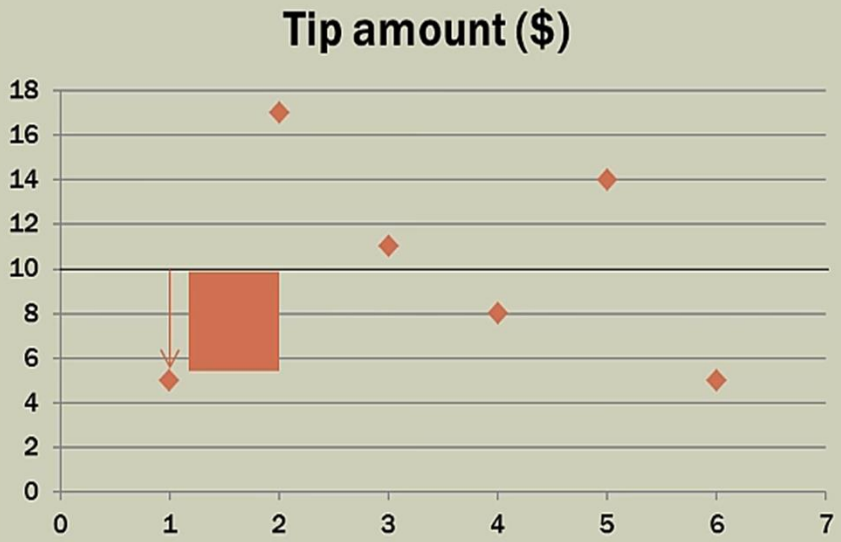
Sum of Squared Errors Comparison

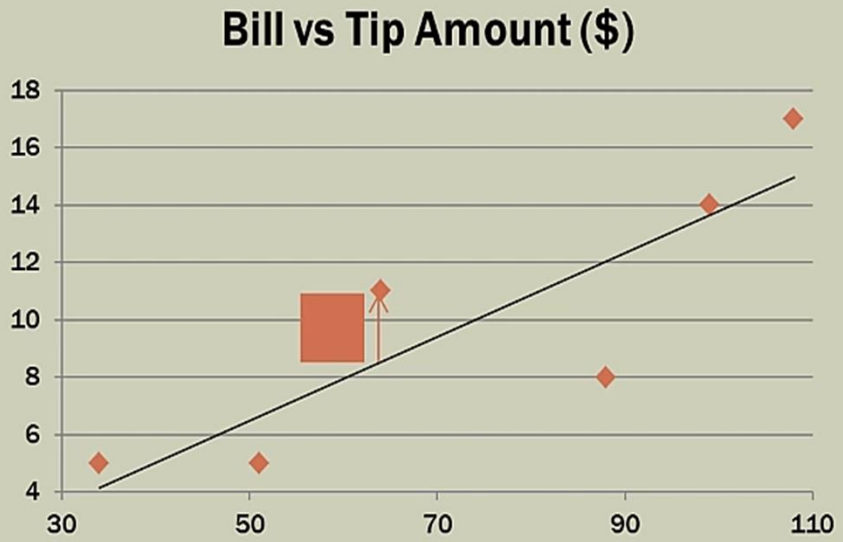
D.V. (Tip amount) ONLY



Visual representation…

A TALE OF TWO LINES





$SSE = 120$

$SSE = SST$

$SST = 120$

$SST = 120$

$120 - 30.075 = SSR$

$120 - 30.075 = 89.925$

$SSE = 30.075$

$SSR = 89.925$

Explanation of errors…

r^2 INTERPRETATION

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

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

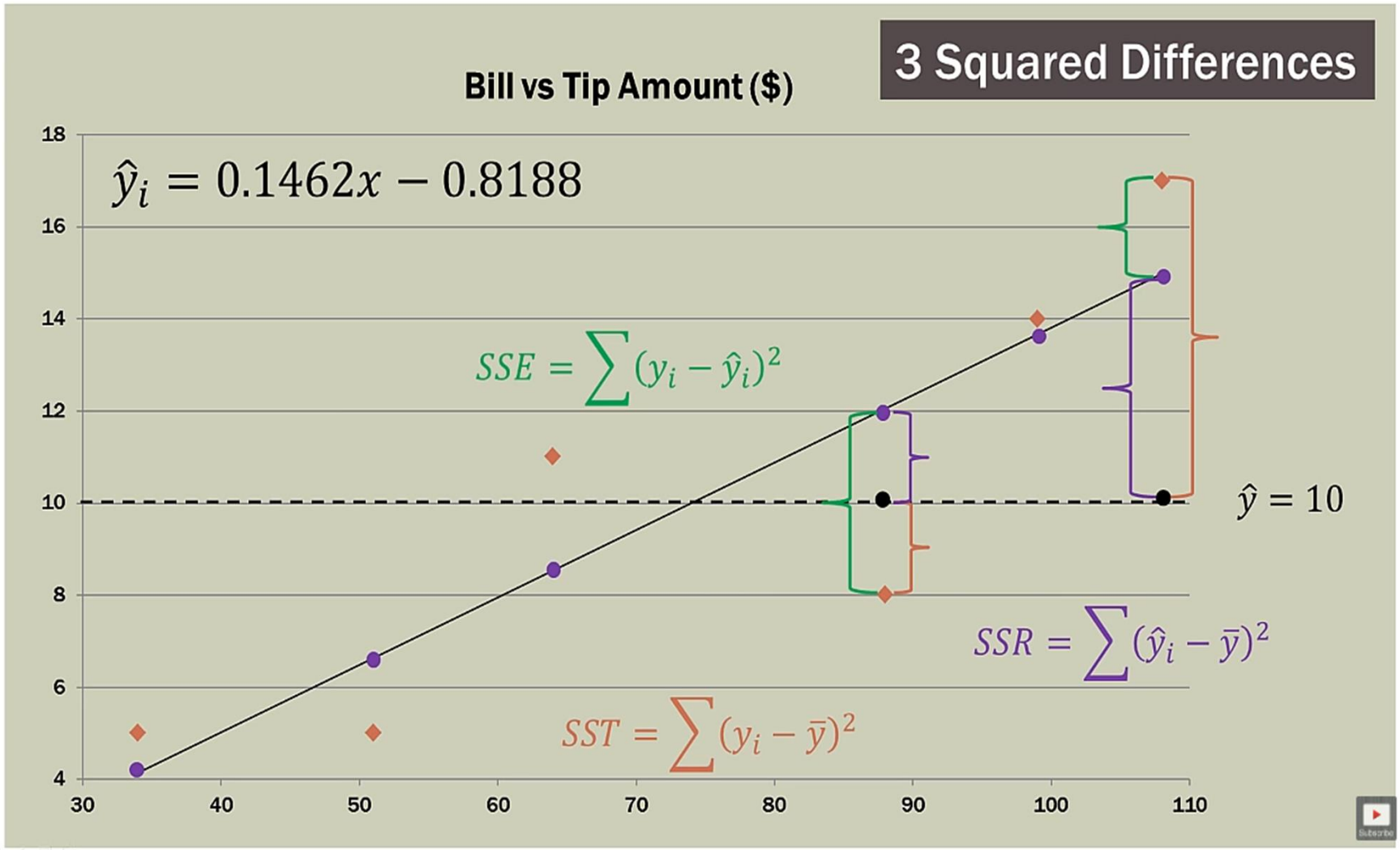
$$Coefficient\ of\ Determination = r^2 = 0.7493\ or\ 74.93\%$$

GOOD FIT!

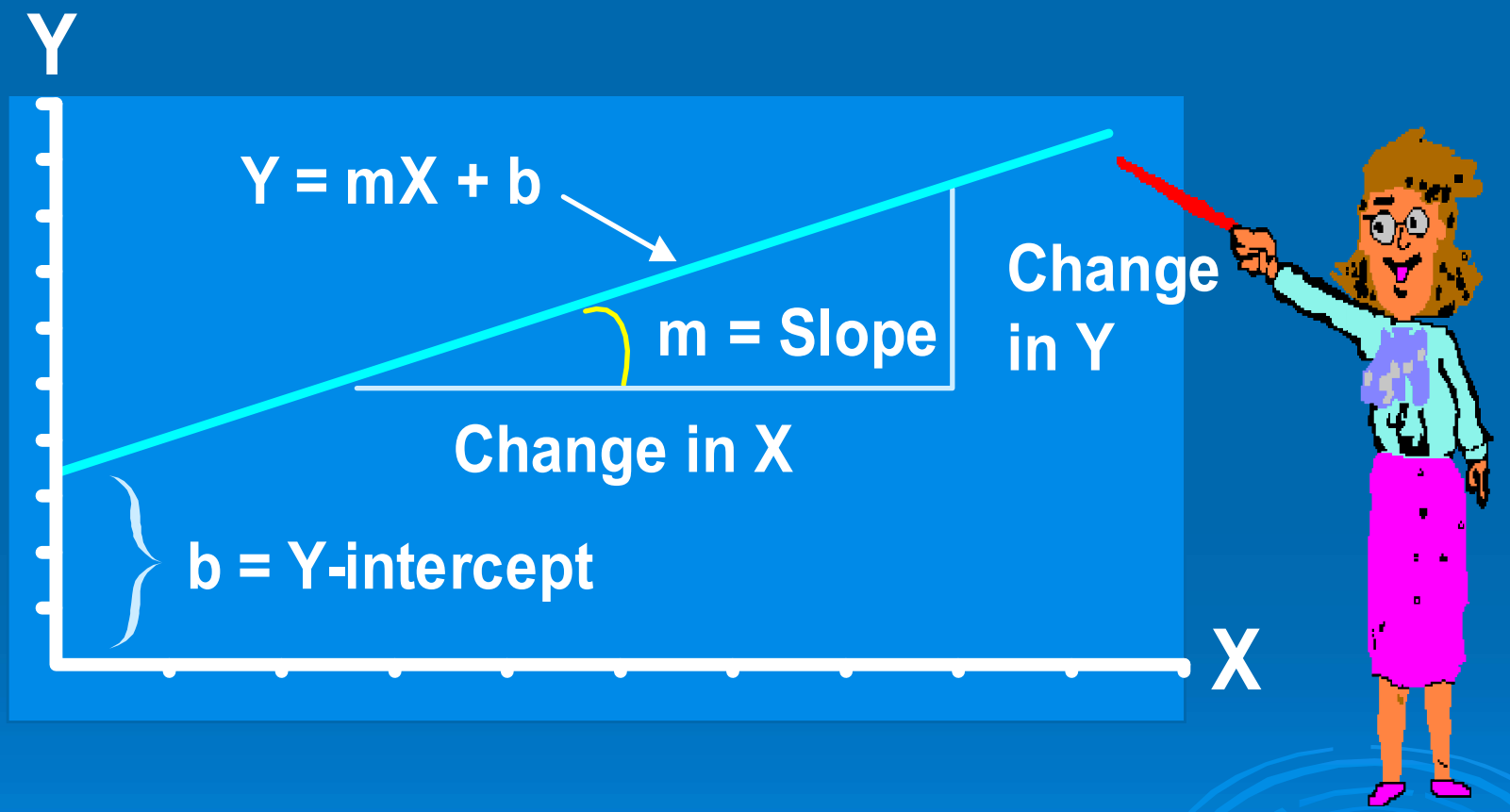
We can conclude that 74.93% of the total sum of squares can be explained by using the estimated regression equation to predict the tip amount. The remainder is error.

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

SST = SSE+SSR


Bill vs Tip Amount ($)
3 Squared Differences
$\hat{y}_i = 0.1462x - 0.8188$
$SSE = \sum (y_i - \hat{y}_i)^2$
$SSR = \sum (\hat{y}_i - \bar{y})^2$
$SST = \sum (y_i - \bar{y})^2$
$\hat{y} = 10$
18
16
14
12
10
8
6
4
30
40
50
60
70
80
90
100
110

Linear Equations



Linear Regression Model

- 1. Relationship Between Variables Is a Linear Function

Population Y-Intercept

Population Slope

Random Error

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Dependent (Response) Variable (e.g., CD+ c.)

Independent (Explanatory) Variable (e.g., Years s. serocon.)

The Explicit Assumptions

These assumptions are explicitly stated by the model:

1. The residuals are independent
2. The residuals are normally distributed
3. The residuals have a mean of 0 at all values of X
4. The residuals have constant variance

The Implicit Assumptions

These assumptions aren't, but the specification of the model implies them. This is the way I've summarized them-they can be written with different terminology, of course.

1. All X are fixed and are measured without error
2. The model is linear in the parameters
3. The predictors and response are specified correctly
4. There is a single source of unmeasured random variance

If there is an assumption you've heard not on this list, chances are it is a logical extension of one of these core assumptions.

Regression with R

GYANVRIKSH

RGui (64-bit)

File Edit Packages Windows Help

C:\Users\mrchakra\Desktop\Desktop_Folder\Learn\LR\SimpleLR.R - R Editor

```
data()
data(airquality)
names(airquality)
attach(airquality)
plot(Ozone~Solar.R)
plot(Ozone~Solar.R,data=airquality)

#Trying the Mean of Ozone
#mean(airquality$Ozone)

#calculate mean ozone concentration (na's removed)
mean.Ozone=mean(airquality$Ozone,na.rm=T)

abline(h=mean.Ozone)

#use lm to fit a regression line through these data:

model1=lm(Ozone~Solar.R,data=airquality)

model1

abline(model1,col="red")
plot(model1)

termplot(model1)
summary(model1)
```

Questions?