STUTERN

Regression (Linear)

SGA07_DATASCI

11th February 2020

Module Overview

- What is a model
- Least square method
- Linear Algebra
- Linear Regression model
- Coefficient of Determination

Book Keeping

- Apologies for last Thursday
- Morning challenge to get you started on group task
- Group task submission set to February 28th
- Concentrate efforts on exploratory analysis

Outcome

After this Module, you will;

- Understand what it means to build a model as a data scientist
- Understand how to use the least squared method to validate assumptions about your model
- Understand the mathematics of linear algebra and how it implies to linear regression models

What is a model

- Experience / Training
- Class / Dependent Variable
- Performance Measure
- Self-learning / Improves with Training

> “A computer program is said to learn from **experience** *E* with respect to some class of **tasks** *T* and **performance** measure *P* if its performance at tasks in T, as measured by P, improves with experience E.”

VAT Service

Let's assume that we work in FIRS and given the new change in VAT, we will like to build a model that allows us to predict what amount of VAT to expect from any given business which provides total amount of sales

FIRS Data

Business_ID	Total_Sales (₦ million)	VAT_Amount (₦ million)
1	34	5
2	108	17
3	64	11
4	88	8
5	99	14
6	51	5
7	45	?
8	78	?
9	123	?

Linear Regression (Def.)

- Statistical Method
- Mathematical relationship between two attributes
- Independent & Dependent attributes

> One of the most common statistical methods is linear regression. At its most basic, it's used when you want to express the mathematical relationship between two variables or attributes.

Least Squared Error

- Dependent variable of FIRS data is VAT_Amount
- We can assume a statistic (such as mean) to replace missing value (predictions)

> “The goal of a simple linear regression is to create a linear model that minimises the sum of squared errors of a predicted value”

Least Square Method

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

$y_i =$ Expected value of dependent attribute

$\hat{y}_i =$ Estimated (predicted) value of dependent attribute

FIRS Data

Business_ID	VAT_Amount	VAT_Mean	Residual	Residual^2
1	5	10	-5	25
2	17	10	7	49
3	11	10	1	1
4	8	10	-2	4
5	14	10	4	16
6	5	10	-5	25
	10			120

Linear Algebra Review

$$y = mx + a$$

$y =$ Dependent variable

$x =$ Independent variable

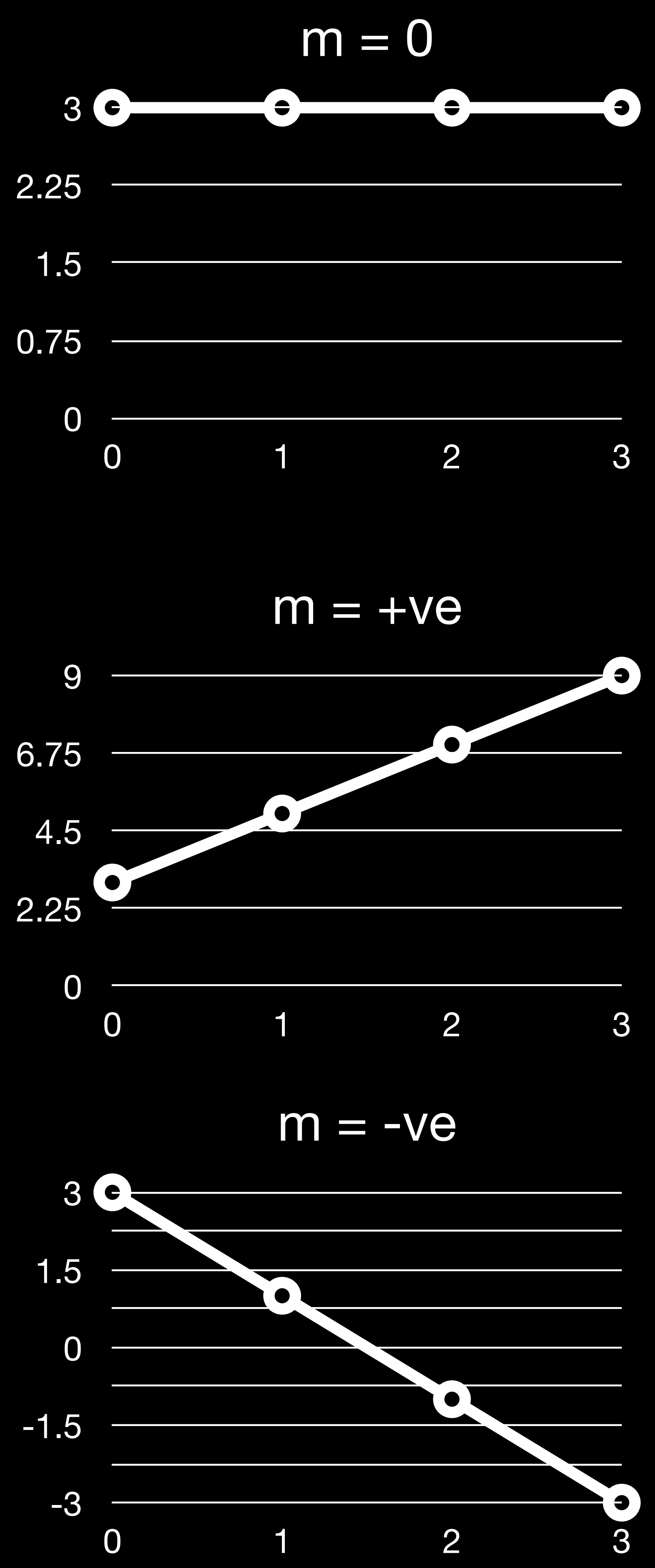
$m =$ Slope of line

$a =$ y-intercept given by $x = 0$

$$y = 2x + 3$$

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

$$a = y = 2(0) + 3 = 3$$



Linear Regression (Formula)

$$\hat{y} = \beta_0 + \beta_1 x$$

$\hat{y} =$ Expected value of dependent variable

$x =$ Independent variable

$\beta_1 =$ Slope of line

$\beta_0 =$ y-intercept given by $x = 0$

$$\beta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$x_i =$ Observed value of independent variable

$\bar{x} =$ Mean value of independent variable

$y_i =$ Observed value of dependent variable

$\bar{y} =$ Mean value of dependent variable

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

FIRS Data

Business_ID	x	y	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})(y_i - \overline{y})$	$(x_i - \overline{x})^2$
1	34	5	-40.00	-5.00	200.00	1600.00
2	108	17	34.00	7.00	238.00	1156.00
3	64	11	-10.00	1.00	-10.00	100.00
4	88	8	14.00	-2.00	-28.00	196.00
5	99	14	25.00	4.00	100.00	625.00
6	51	5	-23.00	-5.00	115.00	529.00
Stat	74	10			615.00	4206.00

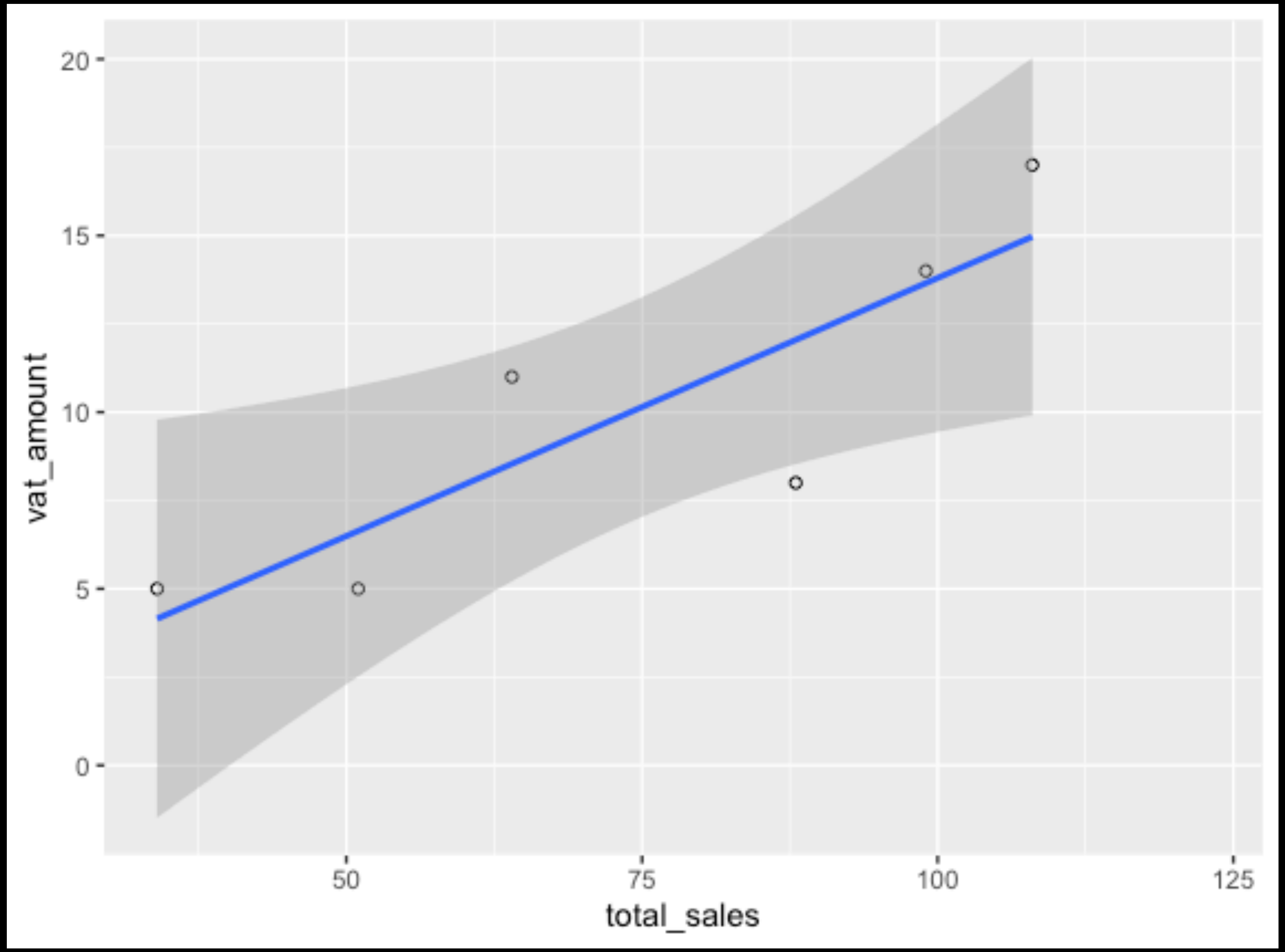
$$\beta_1 = \frac{615}{4206} = 0.1462$$

$$\beta_0 = 10 - (0.1462)74 = - 0.8203$$

$$\hat{y} = 0.1462x_i - 0.8203$$

FIRS Data

Business_ID	x	y	$\hat{y}$	Residual	Residual^2
1	34	5	4.15	0.85	0.72
2	108	17	14.97	2.03	4.12
3	64	11	8.54	2.46	6.07
4	88	8	12.05	-4.05	16.36
5	99	14	13.65	0.35	0.12
6	51	5	6.64	-1.64	2.68
					30.07

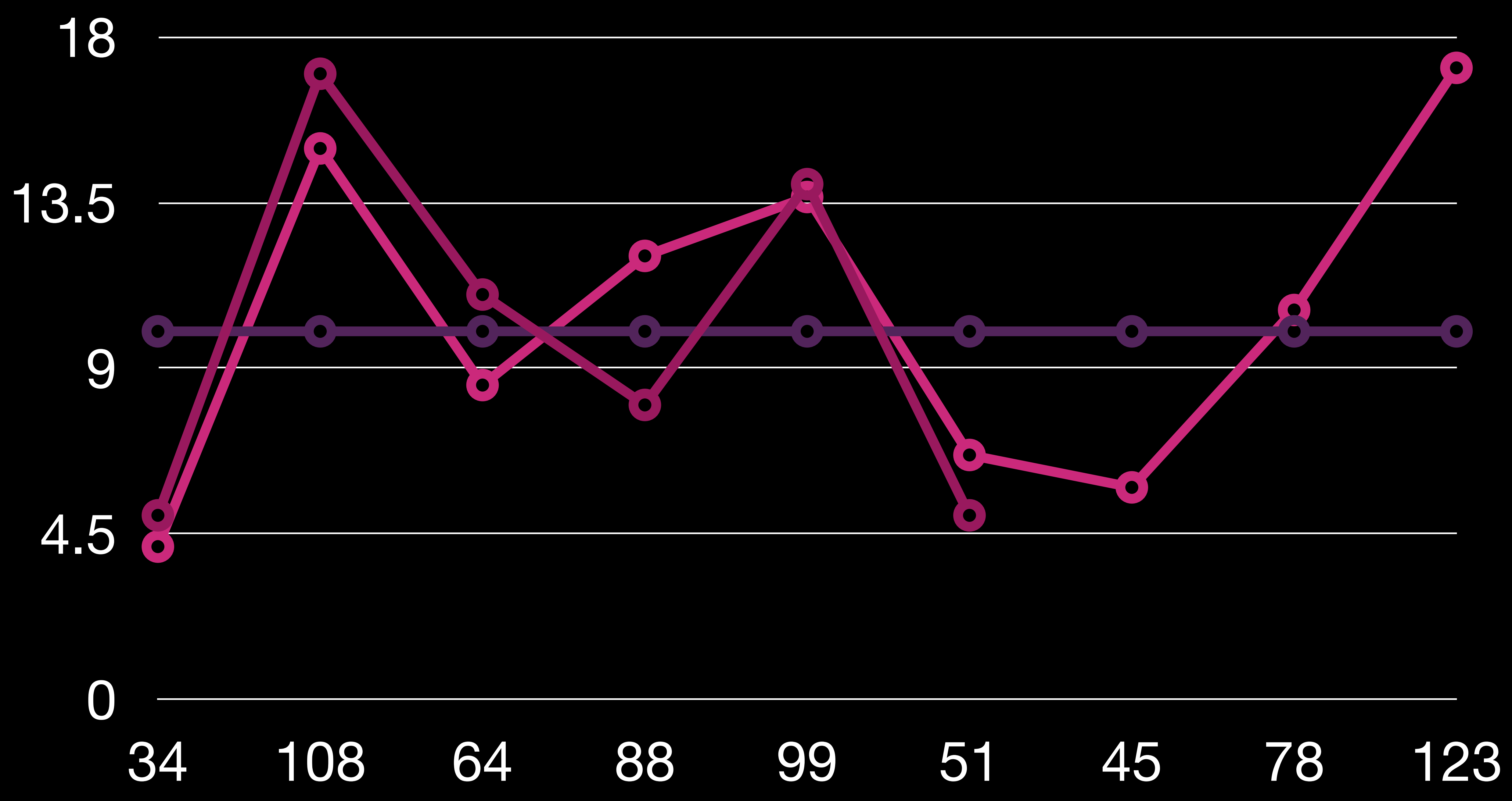


Coefficient of Determination

$$r^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = \frac{120 - 30.07}{120} = \frac{89.93}{120} = 0.749$$

We can conclude that 74.9% of the total sum of squares can be explained by using the estimated regression model to predict the VAT amount given a total sales amount. The reminder is error.

Model Fitting



FIRS Data

Business_ID	Total_Sales (₦ million)	VAT_Amount (₦ million)	y_est_mean	y_est_reg
1	34	5	10	4.15
2	108	17	10	14.97
3	64	11	10	8.54
4	88	8	10	12.05
5	99	14	10	13.65
6	51	5	10	6.64
7	45	?	10	5.76
8	78	?	10	10.58
9	123	?	10	17.16

Practice Lab

Build a predictive linear regression model using R

Use the following Instructions:

- Get your data in R
- Explore the data (Univariate & Bivariate)
- Build a linear model
- Apply linear model to test data to validate model

Recap/Summary

At the end of this Module, you should understand;

- Understand what it means to build a model as a data scientist
- Understand how to use the least squared method to validate assumptions about your model
- Understand the mathematics of linear algebra and how it implies to linear regression models

Suggested Material

- O'Reilly Doing Data Science by Carthy O'Neil and Rachel Schutt Pages 55 - 71
- https://www.youtube.com/playlist?list=PLIeGtxpvyG-LoKUpV0fSY8BGKIMIdmfCi