Simple linear regression

Simple linear regression

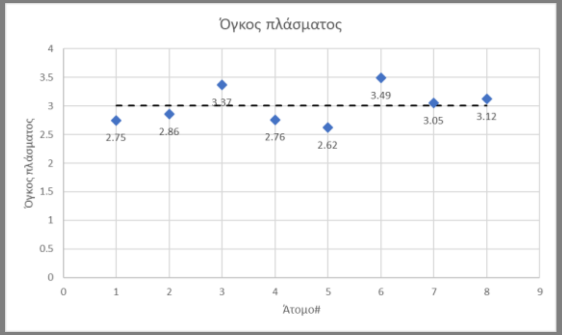
Elias Zintzaras, M.Sc., Ph.D.

Professor in Biomathematics-Biometry
Department of Biomathematics
School of Medicine
University of Thessaly

Institute for Clinical Research and Health Policy Studies
Tufts University School of Medicine
Boston, MA, USA

Theodoros Mprotsis, MSc, PhD
Teacher & Research Fellow
(http://biomath.med.uth.gr)
University of Thessaly
Email: tmprotsis@uth.gr

The basics


Όγκος πλάσματος
4
3.5
3
2.5
2
1.5
1
0.5
0
Όγκος πλάσματος
2.75
2.86
3.37
2.76
2.62
3.49
3.05
3.12
0
1
2
3
4
5
6
7
8
9
Άτομο#

Plasma volume

Suppose we want to create a **model** to **predict** plasma volume based on the body weight of a healthy man

Plasma volume

- The data collected is shown in the table on the right
- In retrospect, it was found that only the plasma volume data was collected.

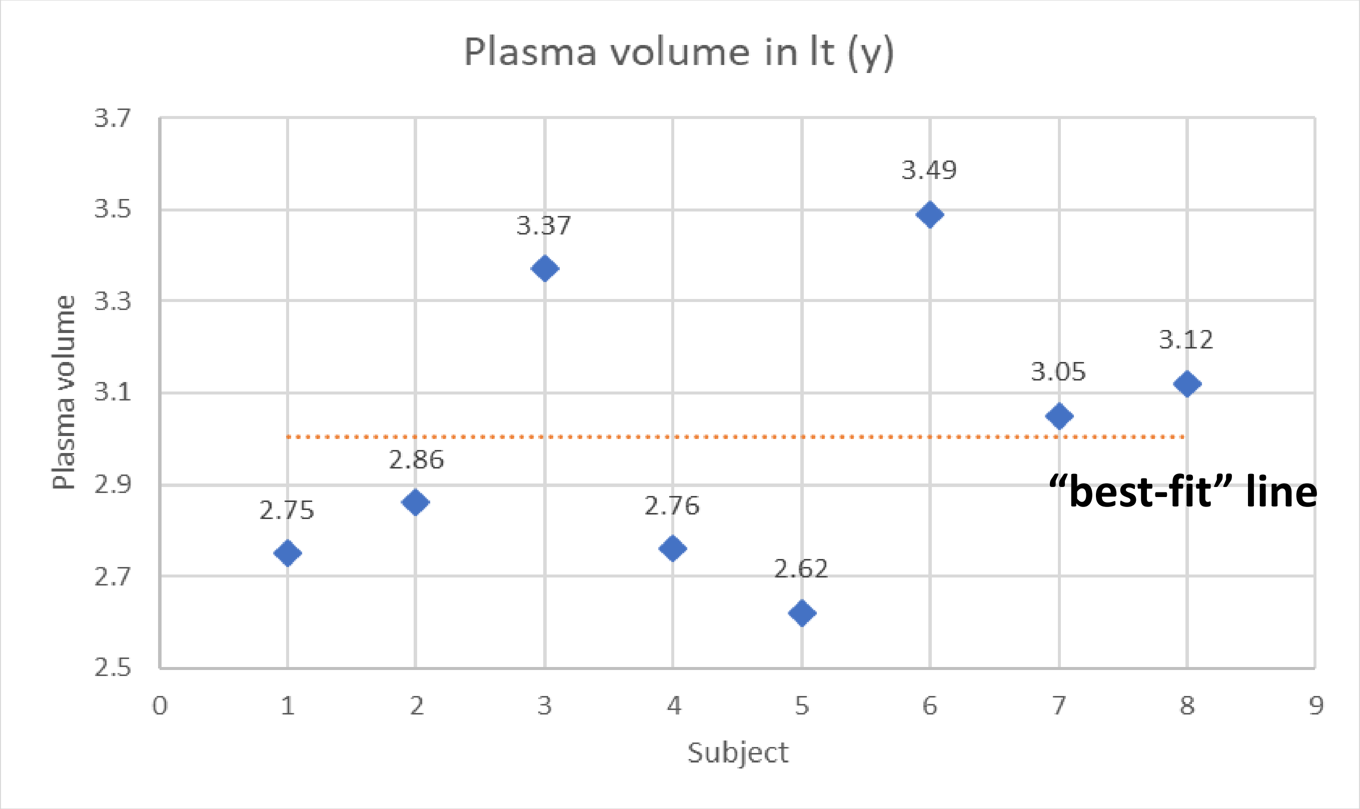
How can we predict the plasma volume for the next healthy subject based only on the data collected?

Subject#	Plasma volume in lt (y)
1	2.75
2	2.86
3	3.37
4	2.76
5	2.62
6	3.49
7	3.05
8	3.12

Plasma volume

Subject#	Plasma volume in lt (y)
1	2.75
2	2.86
3	3.37
4	2.76
5	2.62
6	3.49
7	3.05
8	3.12

$$\bar{y} = 3.0025$$

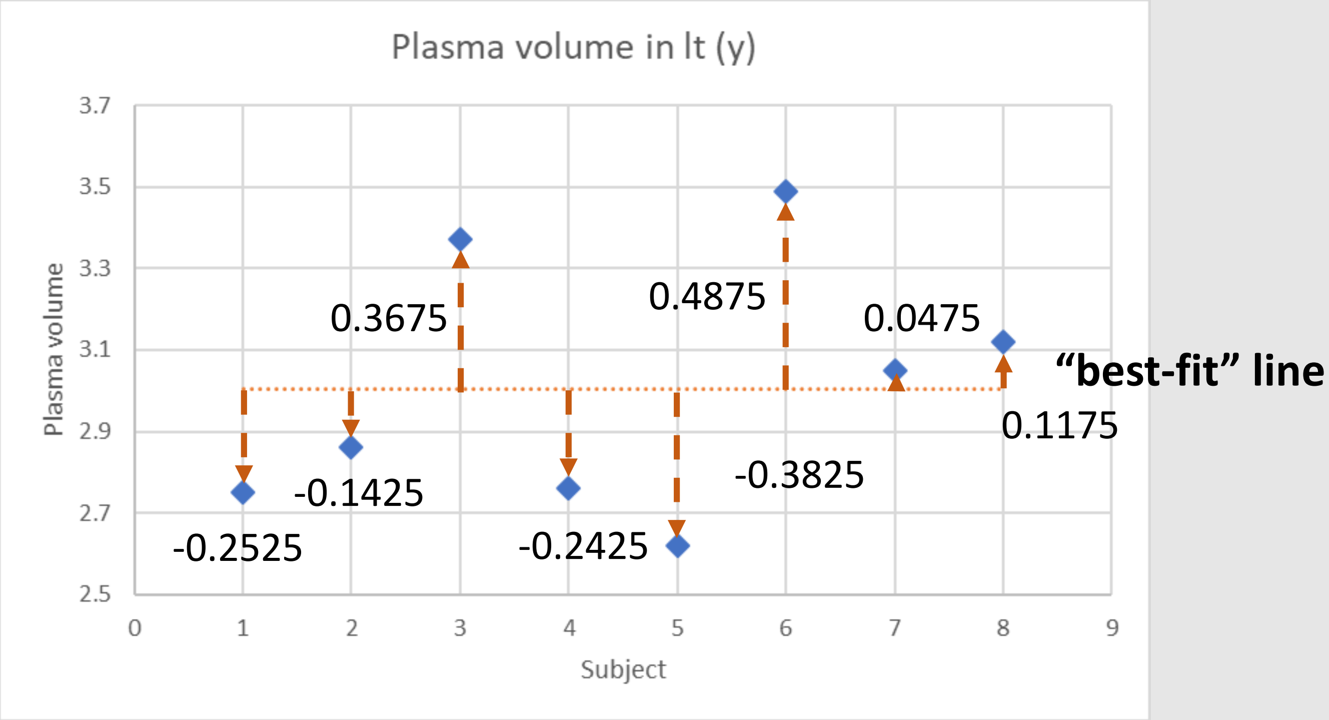


With only one variable and no additional information, the best prediction for the next measurement is the mean of the sample. The variability in plasma volumes can only be explained by the plasma volumes themselves.

Goodness of fit for the plasma volumes

Subject#	Plasma volume in lt (y)
1	2.75
2	2.86
3	3.37
4	2.76
5	2.62
6	3.49
7	3.05
8	3.12

$$\bar{y} = 3.0025$$

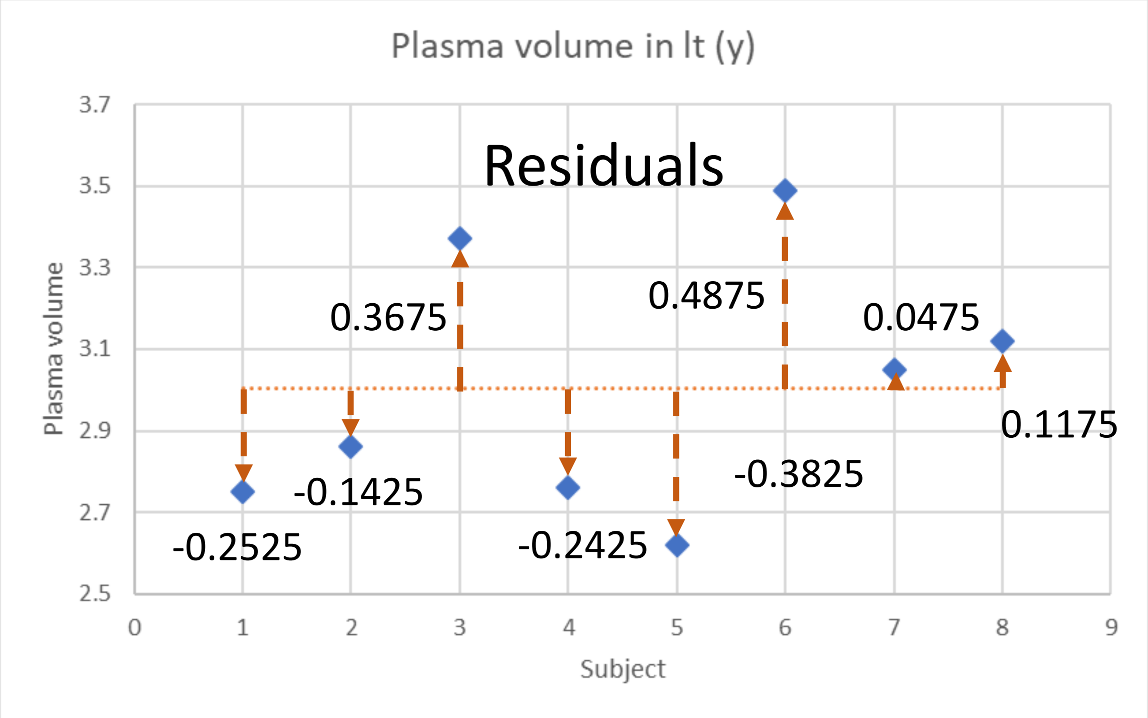


1. The data points, the observed values, do not fall on that line
2. Some are above some are below it
3. This tells us how good this line fits these observed data points
4. One way we can do that is to measure the distance they are from that best-fit line (standard deviation)

Goodness of fit for the plasma volumes

Subject#	Plasma volume in lt (y)
1	2.75
2	2.86
3	3.37
4	2.76
5	2.62
6	3.49
7	3.05
8	3.12

$$\bar{y} = 3.0025$$

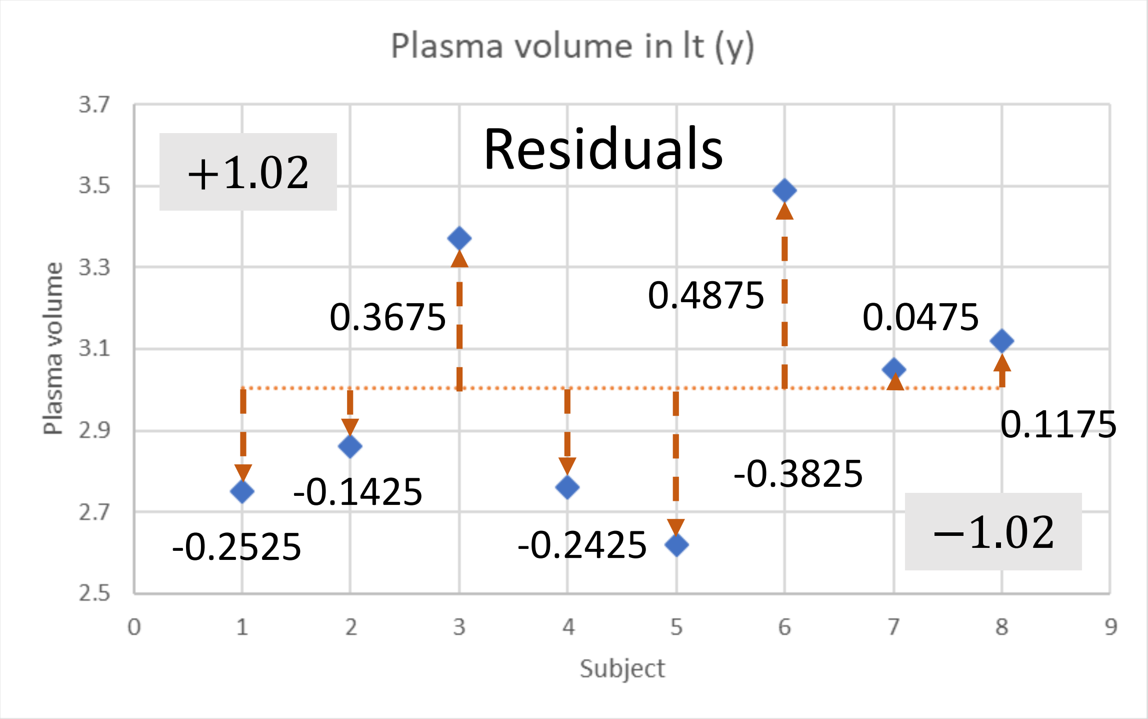


- The distances between the best-fit line and the observed values are called **residuals**
- They are also referred to as **errors** because they represent how far the observed values are from the best-fit line.

Goodness of fit for the plasma volumes

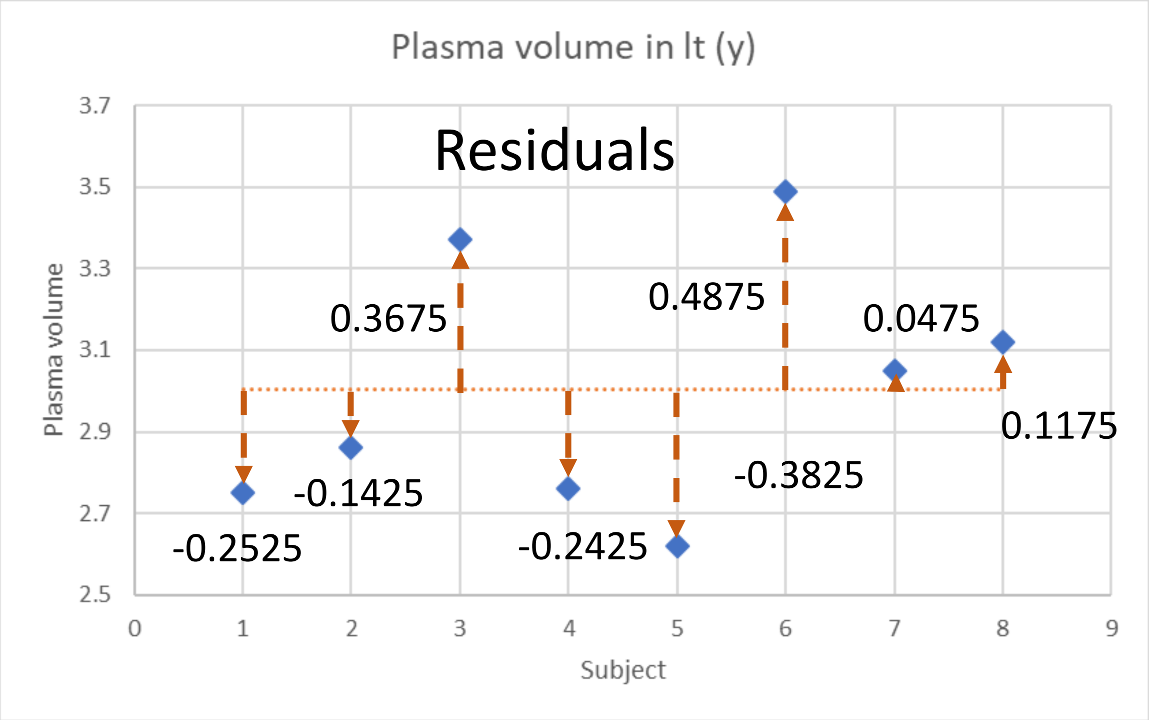
Subject#	Plasma volume in lt (y)
1	2.75
2	2.86
3	3.37
4	2.76
5	2.62
6	3.49
7	3.05
8	3.12

$$\bar{y} = 3.0025$$



- If we add up the residuals above the best-fit line, we get a total of +1.02
- If we add up the residuals below the best-fit line, we get a total of -1.02
- Therefore, the residuals always add up to zero

Squaring the residuals (error)



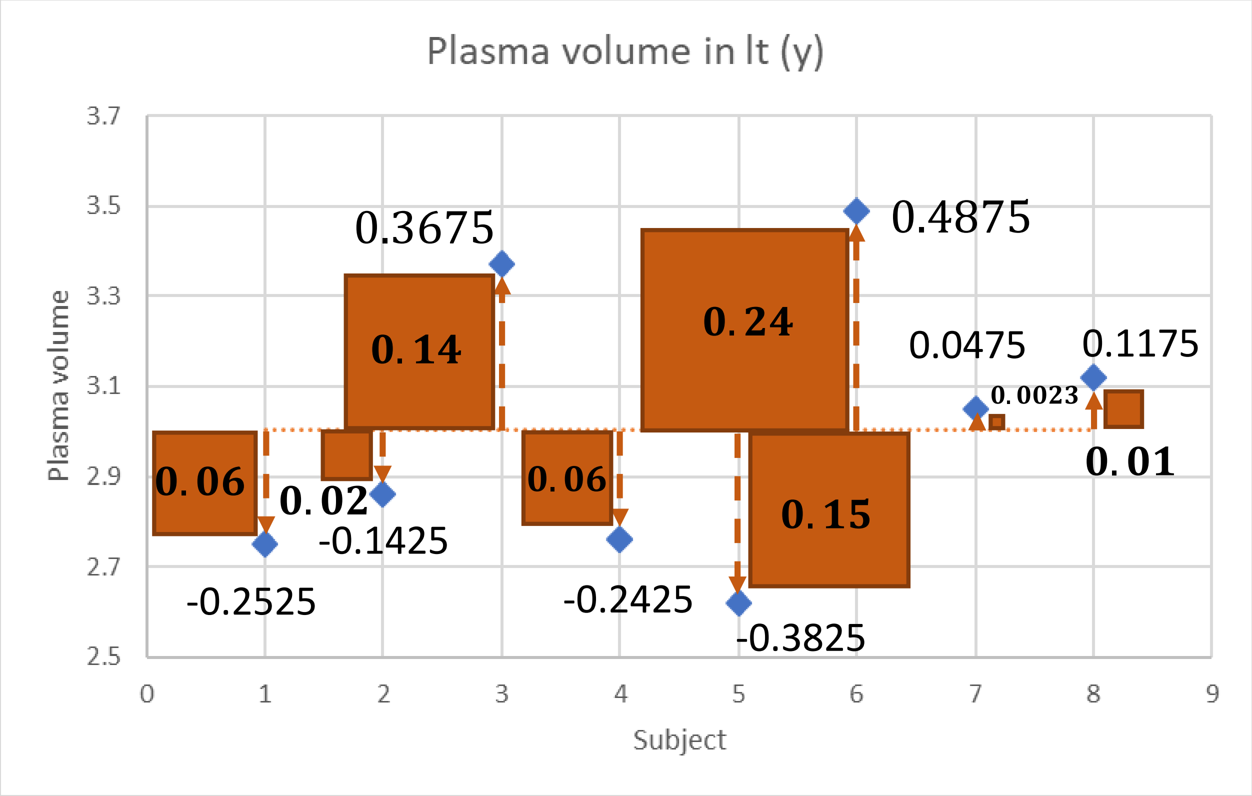
Subject#	Residual	Residual^2
1	-0.2525	0.06
2	-0.1425	0.02
3	0.3675	0.14
4	-0.2425	0.06
5	-0.3825	0.15
6	0.4875	0.24
7	0.0475	0.0023
8	0.1175	0.01

Why square the residuals?

1. Make them positive
2. Emphasizes larger deviations

$$Sum\ of\ squared\ errors\ (SSE) = 0.68$$

Squaring the residuals (error)

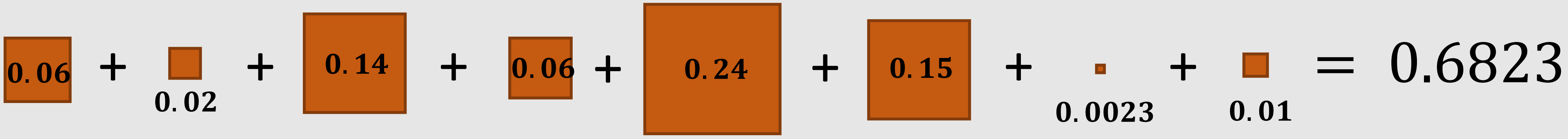


Subject#	Residual	Residual^2
1	-0.2525	0.06
2	-0.1425	0.02
3	0.3675	0.14
4	-0.2425	0.06
5	-0.3825	0.15
6	0.4875	0.24
7	0.0475	0.0023
8	0.1175	0.01

$$Sum\ of\ squared\ errors\ (SSE) = 0.68$$

When we say squaring the residuals we literally mean squaring them

Sum of squares

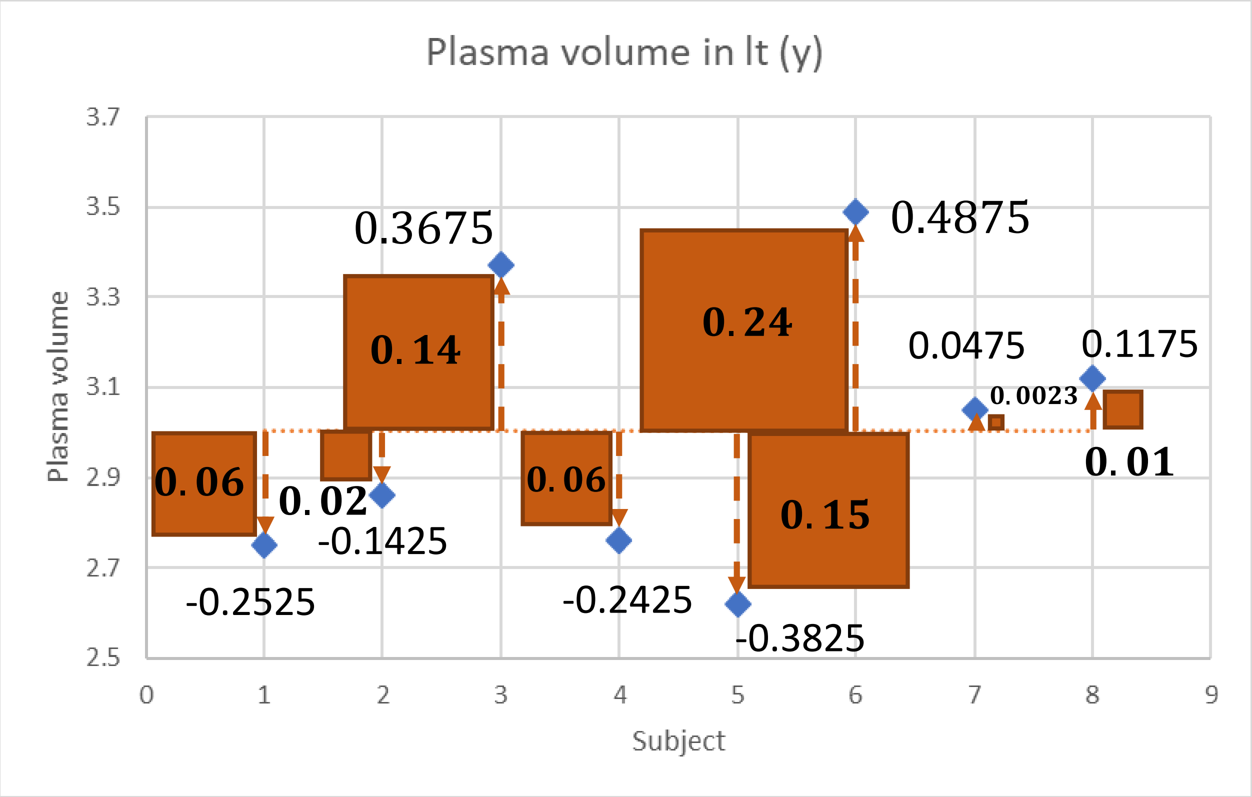


$$0.06 + 0.02 + 0.14 + 0.06 + 0.24 + 0.15 + 0.0023 + 0.01 = 0.6823$$

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 reduce the Sum of Squared Errors (SSE) that we had when we assumed that the independent variable did not exist. The regression line will/should literally “fit” the data better. It will minimize the residuals.

Very important



$$Sum\ of\ squared\ errors\ (residuals) = 0.6823$$

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

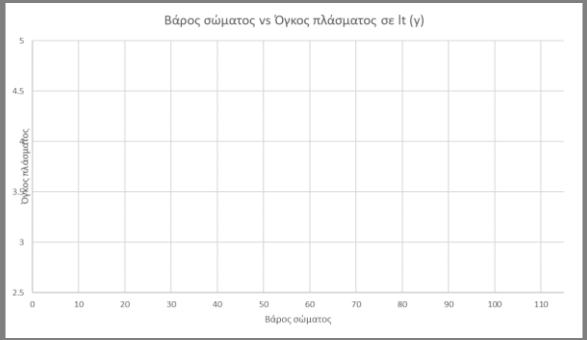
If a two-variable regression model looks like this example, what does the other variable do to help explain the depended variable?

NOTHING

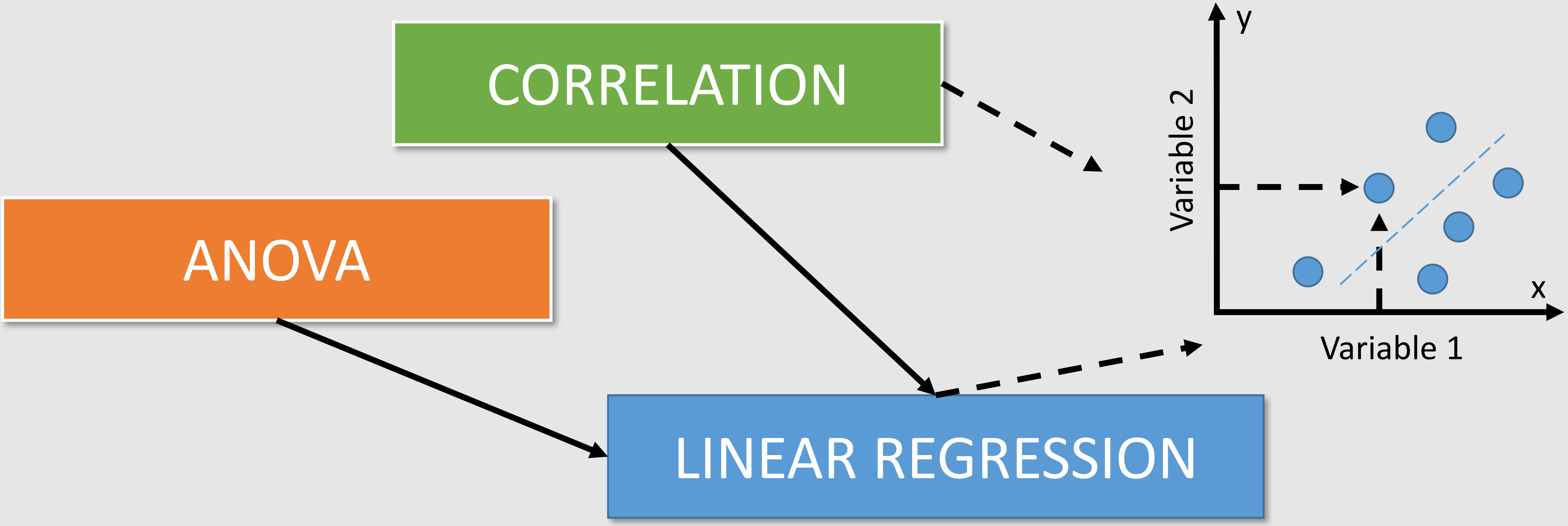
Quick review

- Simple linear regression is really a **comparison of two models**
  - One is where the independent variable does not even exists
  - 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 (error)**
- The residuals are squared and then added together to generate **sum of squares residuals / error, SSE**
- Simple linear regression is designed to find the best fitting line through the data that **minimizes the SSE**

Algebra, correlation, and Graphs



Bivariate statistics



CORRELATION

ANOVA

LINEAR REGRESSION

The value of **one variable (variable 1)**, is a function of **the other variable (variable 2)**

The value of $\boldsymbol{y}$ is a function of $\boldsymbol{x}$; $y = f(x)$

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

Algebra review: Lines

The slope-intercept form of a line

$y = mx + b$

$x =$ some random variable

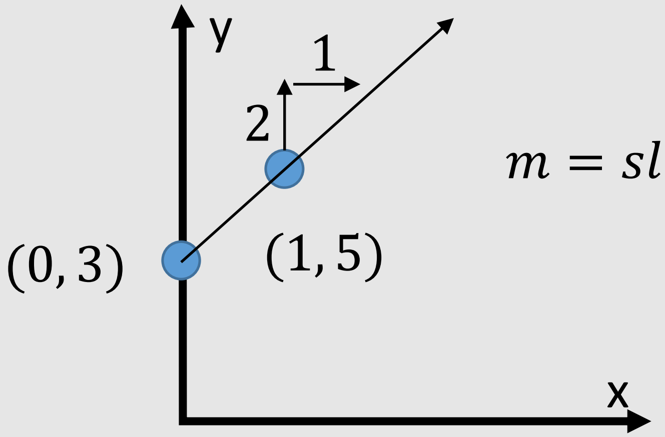
$m =$ slope of the line

$b = y -$ intercept (crosses y $-$ axis)

$y -$ intercept is where $x = 0$

Coordinate of $(0, y)$

$$y = 2x + 3$$



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

$y = 2(0) + 3$

$y = 3$

$(0, 3)$

$y = 2(1) + 3$

$y = 5$

$(1, 5)$

Simple linear regression model

$$\begin{tikzcd}
 & y = \beta_0 + \beta_1 x + \varepsilon \\
y = mx + b \arrow[ur, dashed] &
\end{tikzcd}$$

$\beta_0 = y - intercept$ of the population parameter

$\beta_1 = slope$ of the population parameter

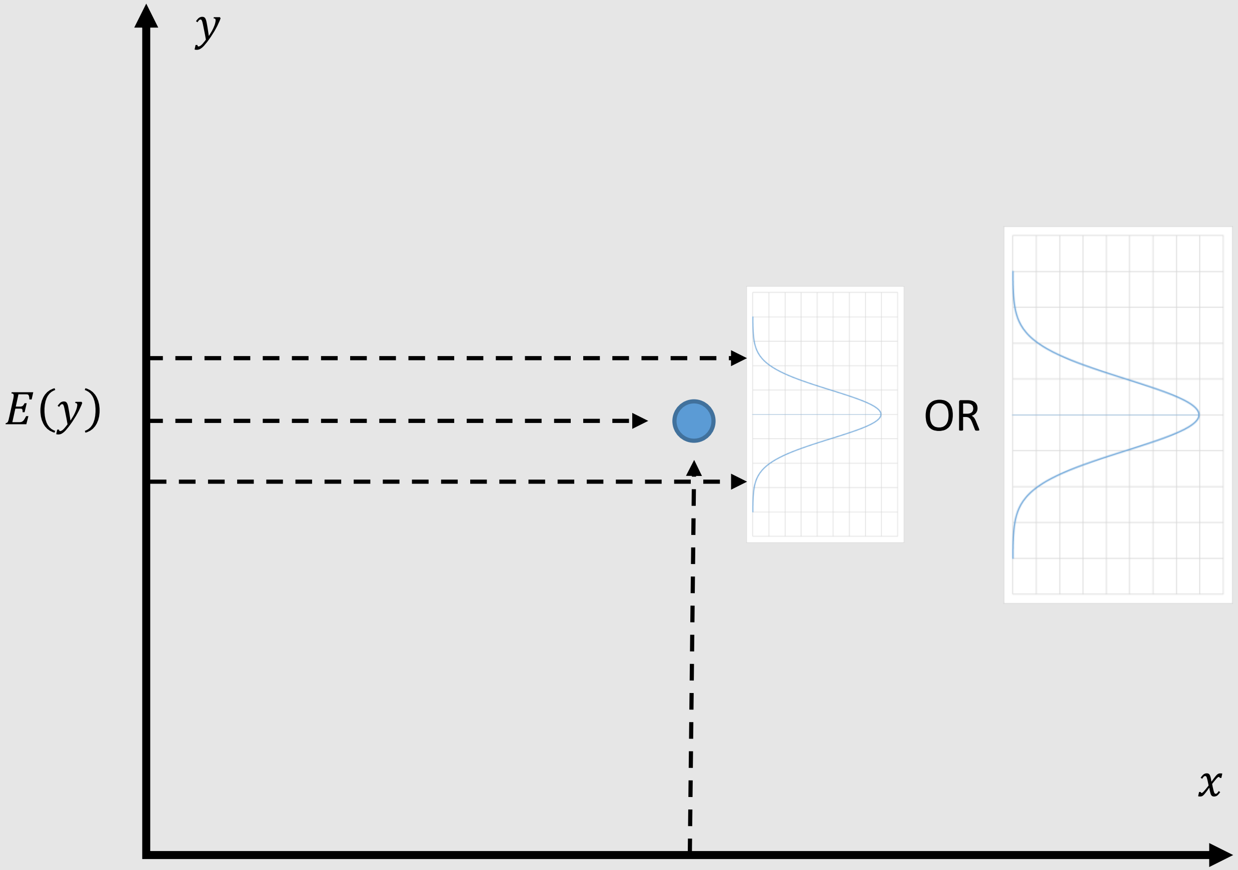
$\varepsilon =$ error term, unexplained variance in y

Simple Linear Regression Equation

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

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

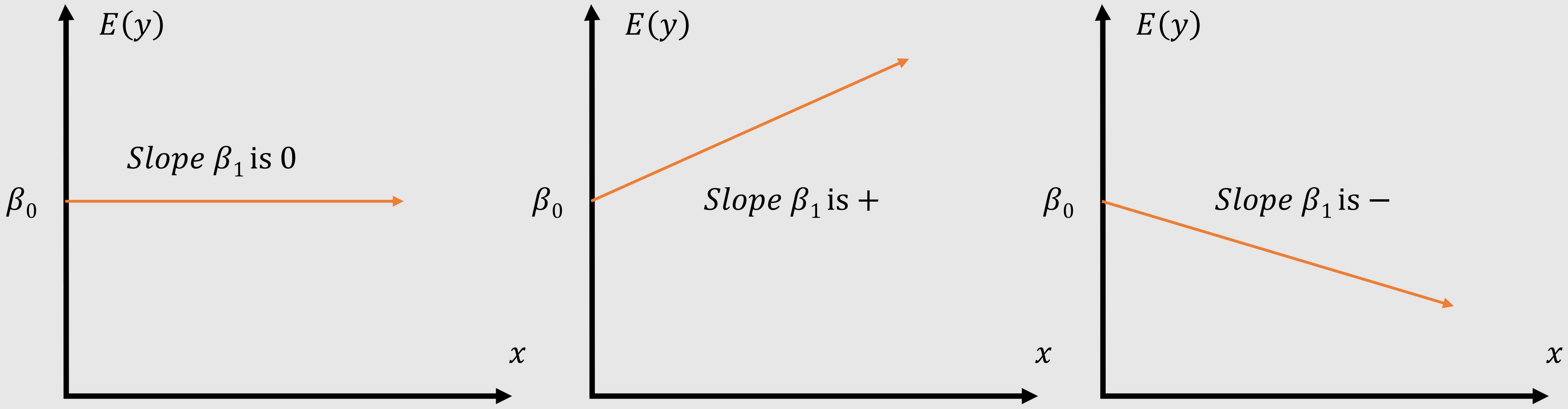
Distribution of y – values



- The expected value is the **mean**. What does it mean?
- If we choose a value of x that corresponds to a value of y. But that is the expected value of y. And that is not so simple. There is actually a distribution of $y's$ for that given x.
- Remember our regression model is not going to be perfect. Any expected value of y is at best an approximation.
- **Therefore, when we say the expected value, what we mean is that it is the mean of a small distribution for that y.**

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 paratemrs, β_0 and β_1, we could use the **Simple Linear Regression**

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

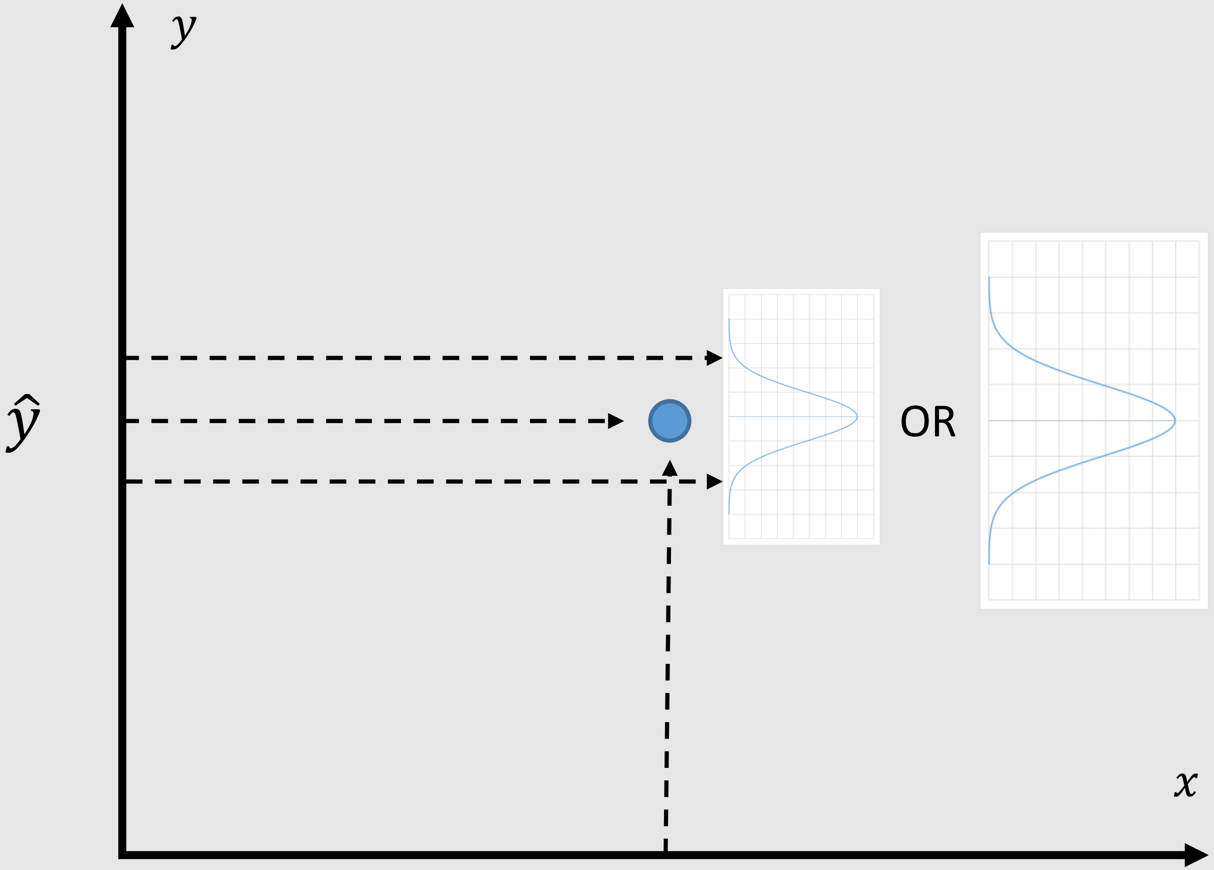
In reality we almost never have the population parameters. Therefore we will estimate them using sample data. When using sample data, we have to change our equation a little bit.

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

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

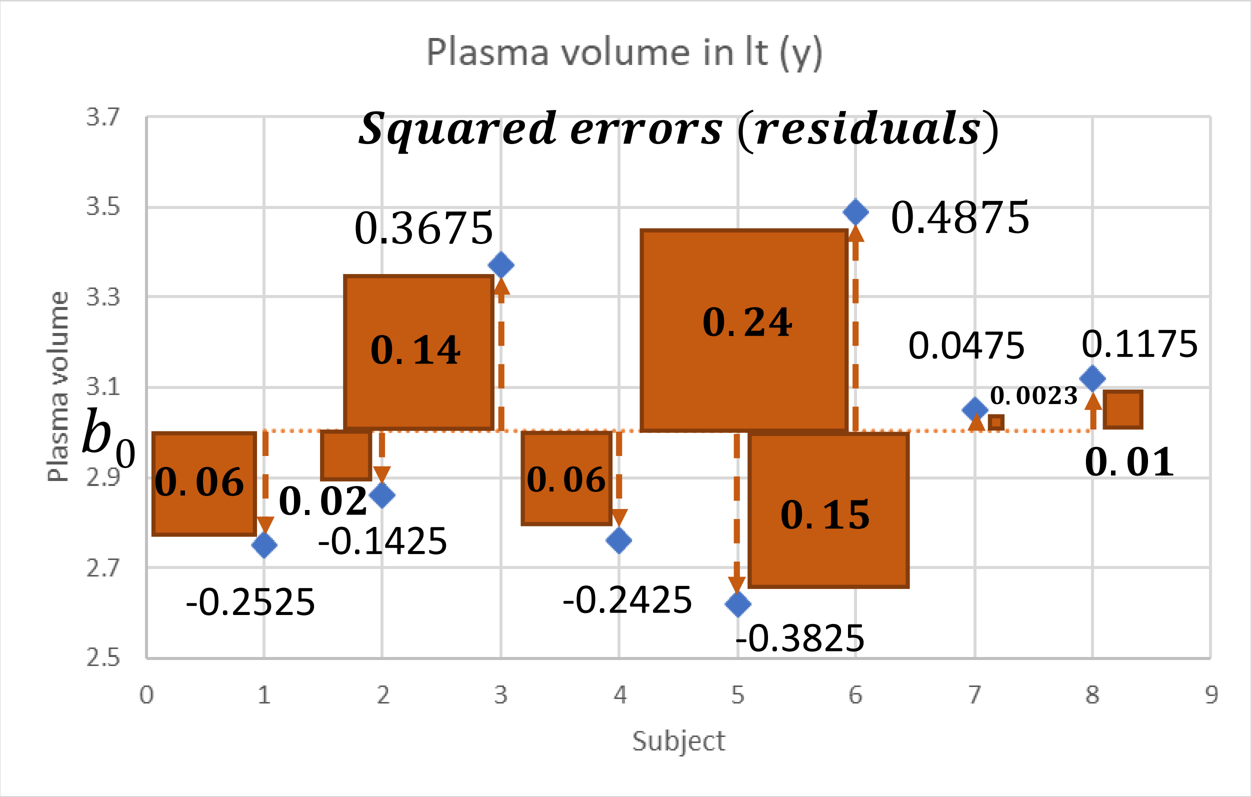
$\hat{y}$ is the mean value of y for a given value of x

Distribution of sample y - values



- We use $\hat{y}$ since we are using samples
- The basic idea is the same
- $\hat{y}$ is the mean of the excepted values of y for any given value of x

When the slope, $\boldsymbol{\beta_1 = 0}$



$$Sum\ of\ squared\ errors\ (residuals) = 0.6823$$

When conducting simple linear regression with two variables, we will determine how good the regression like “fits” the data by comparing it to this type; where we pretend the second variable does not even exist;

the slope, $\boldsymbol{\beta_1 = 0}$

In this case, the value of $\hat{y}$ is 3.0025 **for every value of** x.

$$\hat{y} = b_0 + b_1x$$

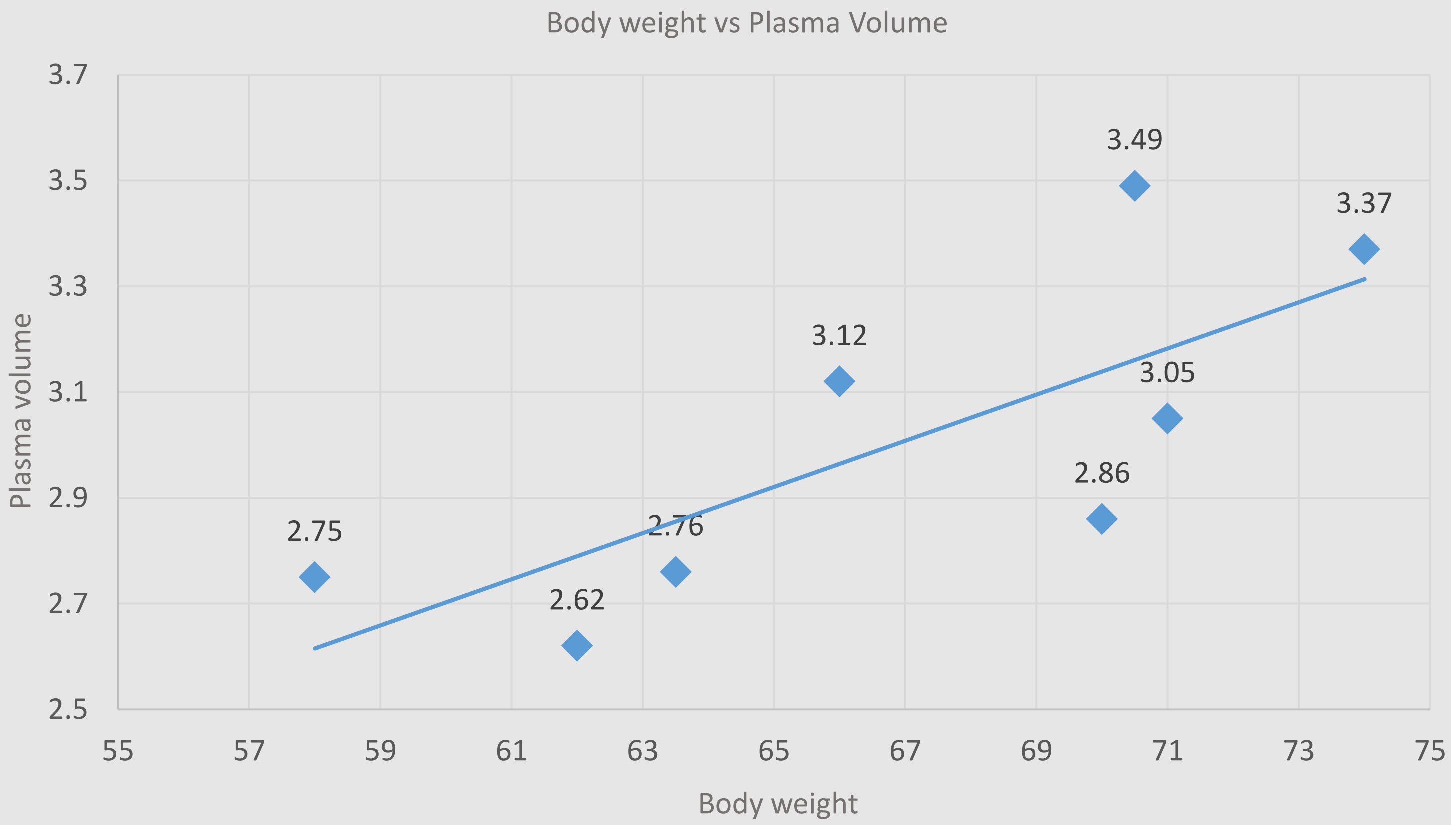
$$\hat{y} = b_0 + (0)x$$

$$\hat{y} = b_0$$

$$b_0 = 3.0025$$

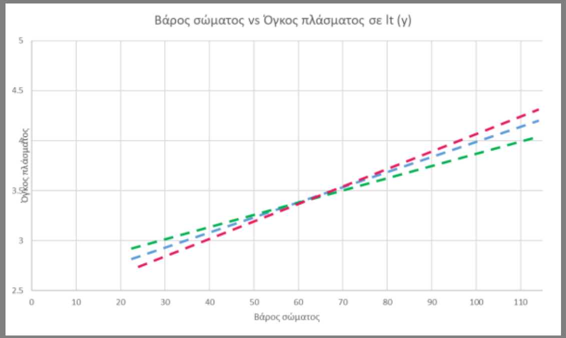
$$\hat{y} = 3.0025$$

Getting ready for least squares



Weight in Kg (x)	Plasma volume in lt (y)
58.0	2.75
70.0	2.86
74.0	3.37
63.5	2.76
62.0	2.62
70.5	3.49
71.0	3.05
66.0	3.12

The least square method



Example

Until now, we only had **plasma volume**, but we also managed to find the **body weight**. Now, we're working with two paired variables

We aim to assess how well plasma volume can be predicted based on body weight

In this case, the **dependent variable** is plasma volume, and the **independent variable** is body weight

Body weight in Kg (x)	Plasma volume in lt (y)
58.0	2.75
70.0	2.86
74.0	3.37
63.5	2.76
62.0	2.62
70.5	3.49
71.0	3.05
66.0	3.12

Least squares criterion

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

y_i = observed value of dependent variable (plasma $volume$)

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

- What we notice is that, for each x value in the graph, we have two corresponding values
- The observed value y_i and the estimated value $\hat{y}_i$ provided by the model
- These values are not identical, and there will be differences between them
- We square these differences and then sum them up
- Our goal is to make this sum as small as possible

Least squares criterion

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

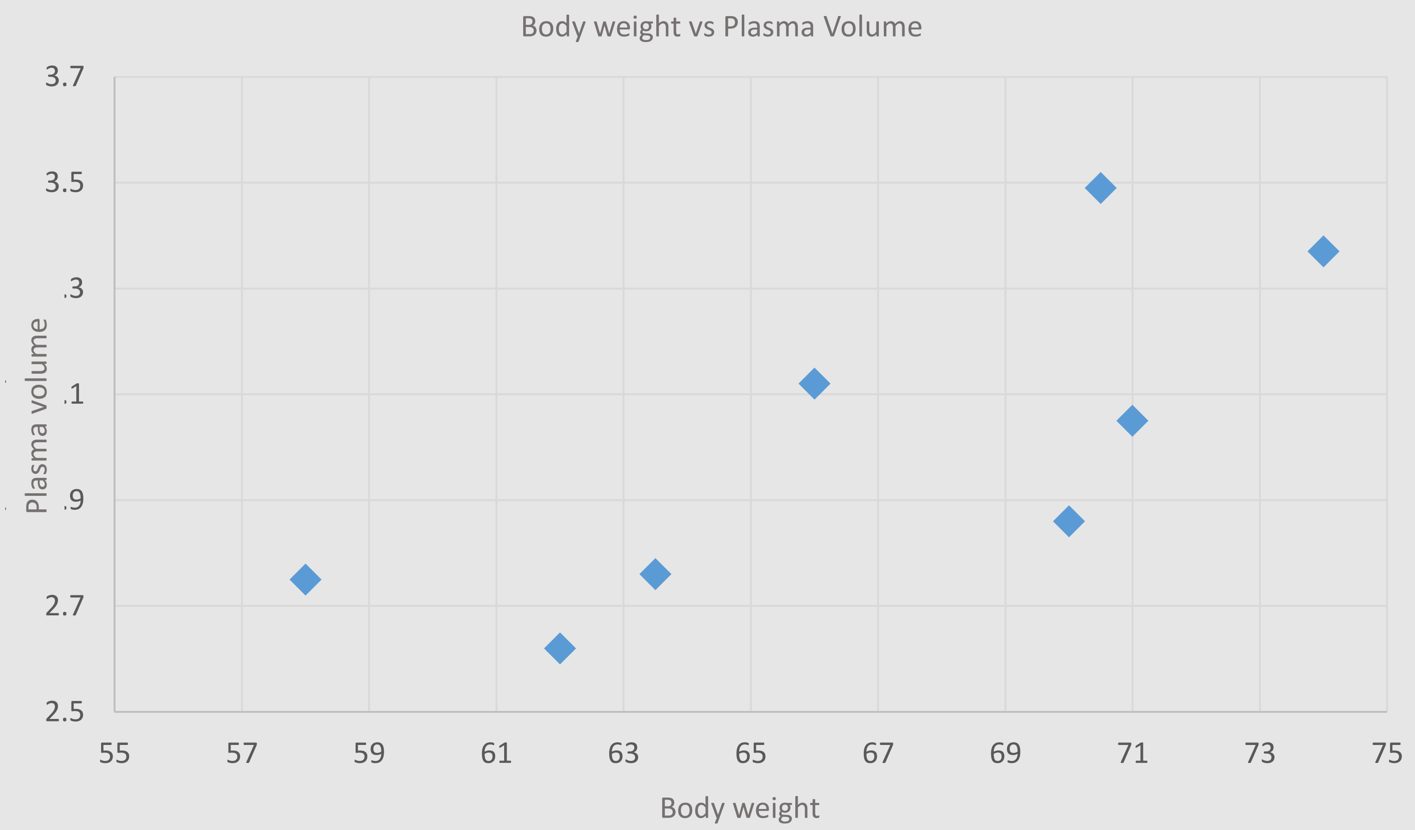
y_i = observed value of dependent variable (plasma $volume$)

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

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 squared residuals.

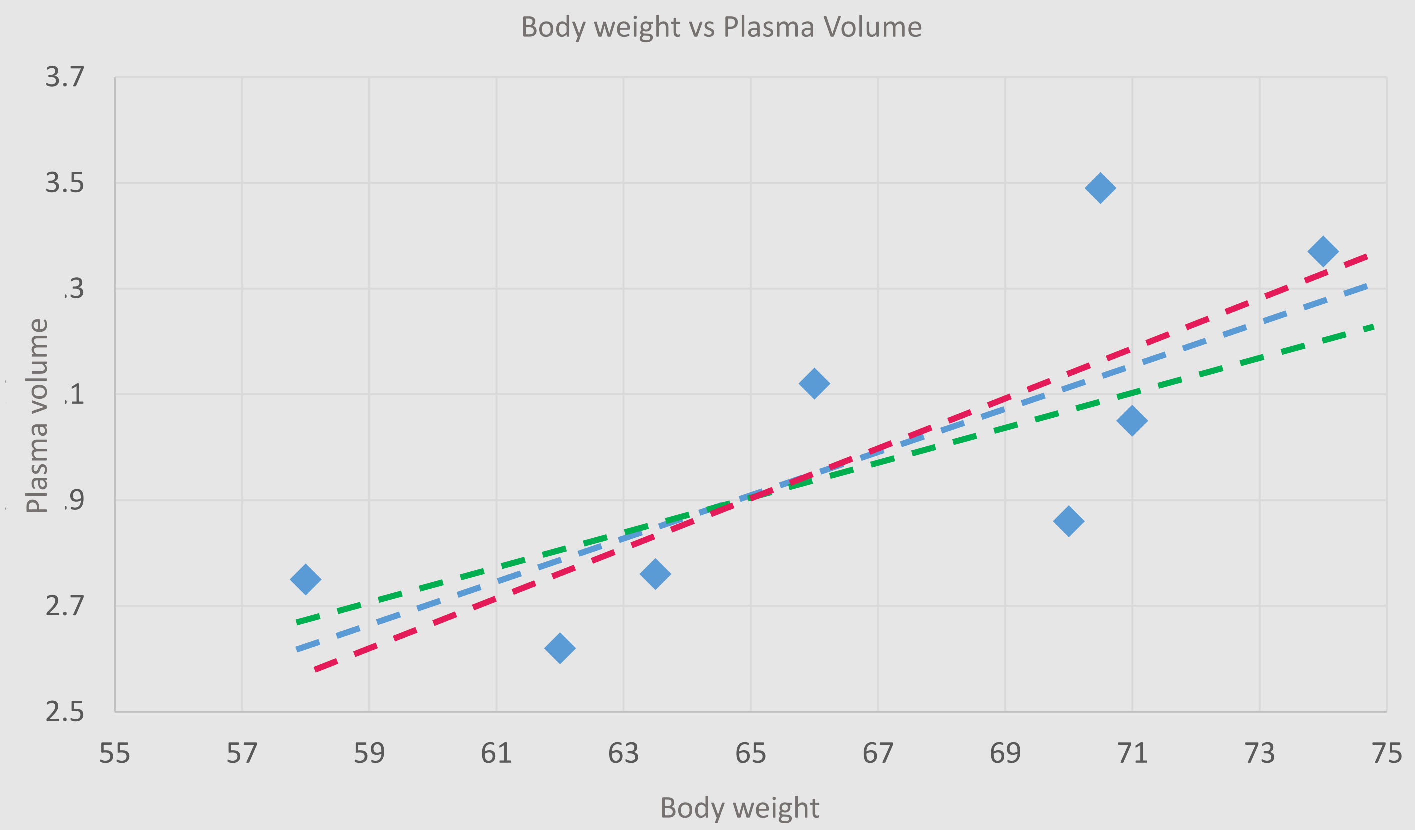
Not only that, but the sum of squared residuals should be much smaller than we used just the dependent variable alone; $\beta_1 = 0, \hat{y} = 3.0025$ for all values of x. The sum of squared residuals was 0.6823.

Step 1: Scatter plot



Body weight in Kg (x)	Plasma volume in lt (y)
58.0	2.75
70.0	2.86
74.0	3.37
63.5	2.76
62.0	2.62
70.5	3.49
71.0	3.05
66.0	3.12

Step 2: Look for a visual line



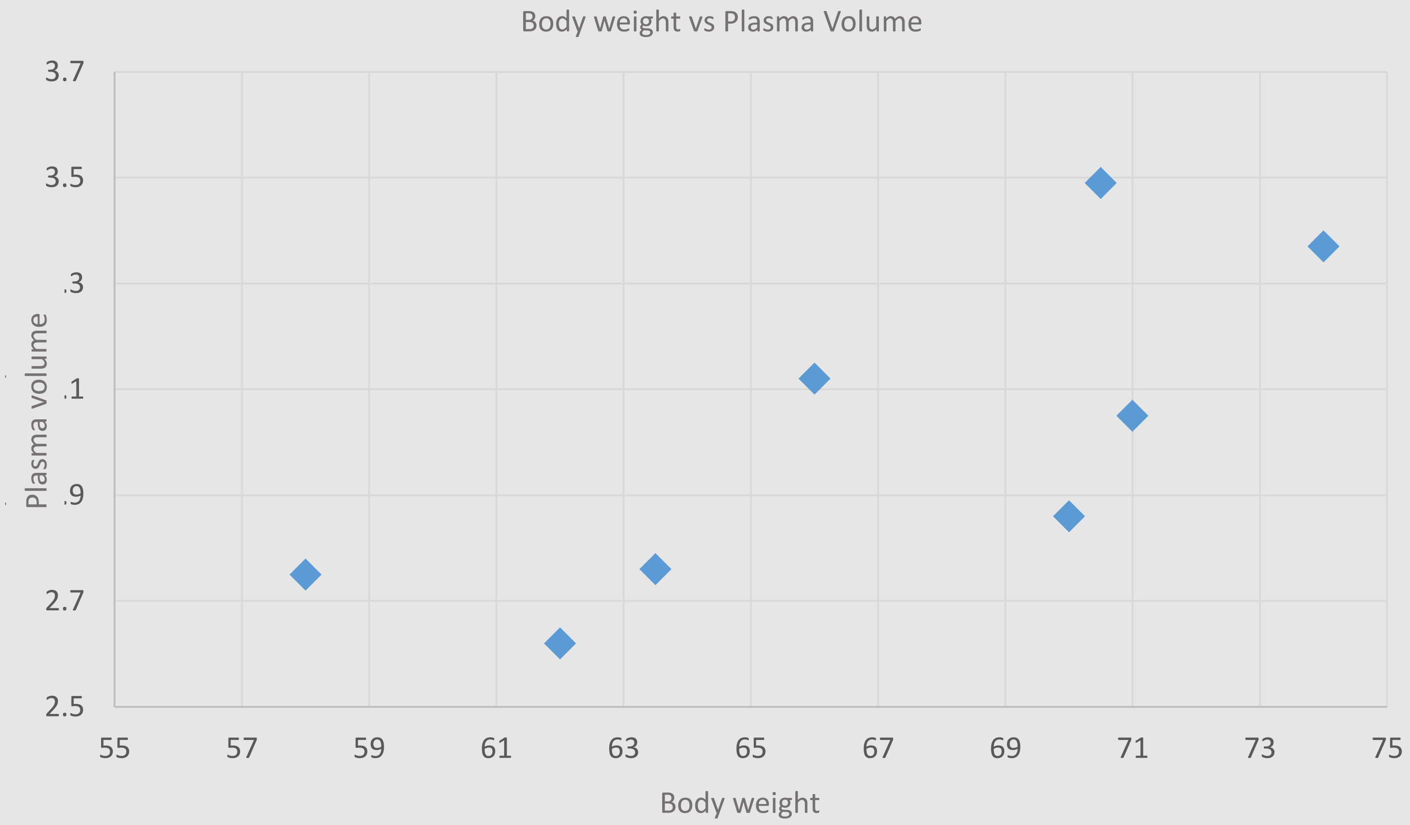
Does this data seem to fall along a line?

In this case, YES! Proceed.

But we do not know which of these lines is the regression line.

If our data appears somewhat disorganized (a little bit messy), then we stop.

Step 3: Correlation (optional)



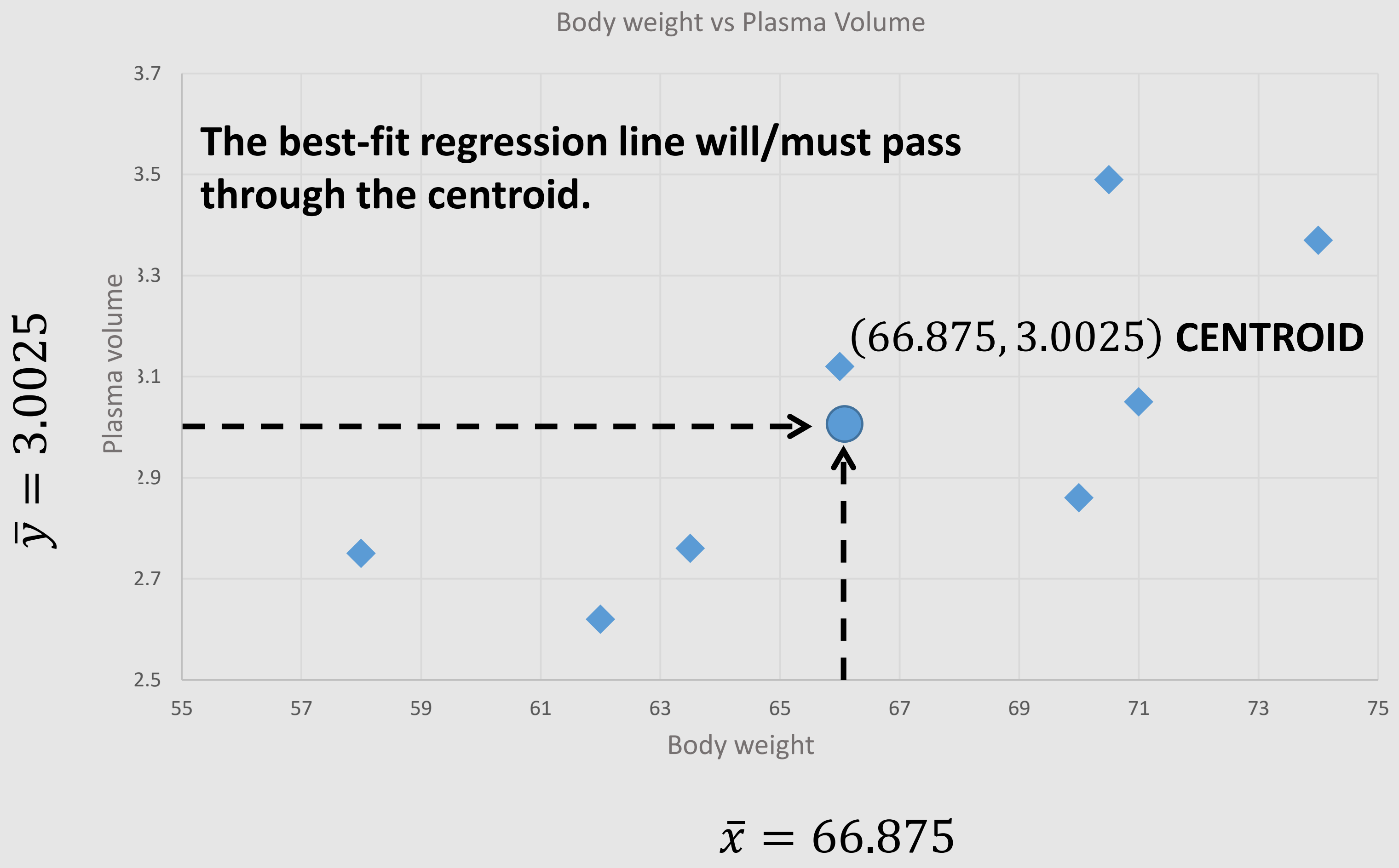
What is the correlation coefficient r?

In this case, $r = 0.743$

Is the relationship strong?

In this case, **YES**!

Step 4: Descriptive Statistics / Centroid



The best-fit regression line will/must pass through the centroid.

$(66.875, 3.0025)$ **CENTROID**

Body weight in Kg (x)	Plasma volume in lt (y)
58.0	2.75
70.0	2.86
74.0	3.37
63.5	2.76
62.0	2.62
70.5	3.49
71.0	3.05
66.0	3.12
$\bar{x} = 66.875$	$\bar{y} = 3.0025$

Step 5: Calculations

Intercept

$$\widehat{y_i} = b_0 + b_1 x_i$$

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 depedent 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. Multiple 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 square values from step 2

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

Step 5: Calculations

Subject#	Body weight in Kg	Plasma volume in lt (y)	Body deviation	Plasma volume deviation	Deviation products	Body weight 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	58	2.75	−8.875	−0.2525	2.2409375	78.765625
2	70	2.86	3.125	−0.1425	−0.4453125	9.765625
3	74	3.37	7.125	0.3675	2.6184375	50.765625
4	63.5	2.76	−3.375	−0.2425	0.8184375	11.390625
5	62	2.62	−4.875	−0.3825	1.8646875	23.765625
6	70.5	3.49	3.625	0.4875	1.7671875	13.140625
7	71	3.05	4.125	0.0475	0.1959375	17.015625
8	66	3.12	−0.875	0.1175	−0.1028125	0.765625
	$\bar{x} = \mathbf{66.875}$	$\bar{y} = \mathbf{3.0025}$			$\sum = \mathbf{8.9575}$	$\sum = \mathbf{205.375}$

β_1 calculations (slope)

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

$$b_1 = \frac{8.9575}{205.375}$$

$$b_1 = 0.04362$$

Therefore, the slope of our regression line is 0.04362

Deviation products	Body weight deviations squared
$(\boldsymbol{x_i - \bar{x}})(\boldsymbol{y_i - \bar{y}})$	$(\boldsymbol{x_i - \bar{x}})^2$
$\mathbf{2.2409375}$	$\mathbf{78.765625}$
$\mathbf{-0.4453125}$	$\mathbf{9.765625}$
$\mathbf{2.6184375}$	$\mathbf{50.765625}$
$\mathbf{0.8184375}$	$\mathbf{11.390625}$
$\mathbf{1.8646875}$	$\mathbf{23.765625}$
$\mathbf{1.7671875}$	$\mathbf{13.140625}$
$\mathbf{0.1959375}$	$\mathbf{17.015625}$
$\mathbf{-0.1028125}$	$\mathbf{0.765625}$
$\sum$ = 8.9575	**$\sum$ = 205.375**

β_0 calculations (y – intercept)

$$b_0 = \bar{y} - b_1\bar{x} \qquad b_1 = 0.04362$$

$$b_0 = 3.0025 - 0.04362 \cdot 66.875$$

$$b_0 = 0.0857$$

Body Weight in kg (x)	Plasma volume in lt (y)
x	y
58	2.75
70	2.86
74	3.37
63.5	2.76
62	2.62
70.5	3.49
71	3.05
66	3.12
$\bar{x} = 66.875$	$\bar{y} = 3.0025$

Regression line

$\widehat{y}_i = b_0 + b_1 x_i$

$b_0 = 0.0857$

intercept

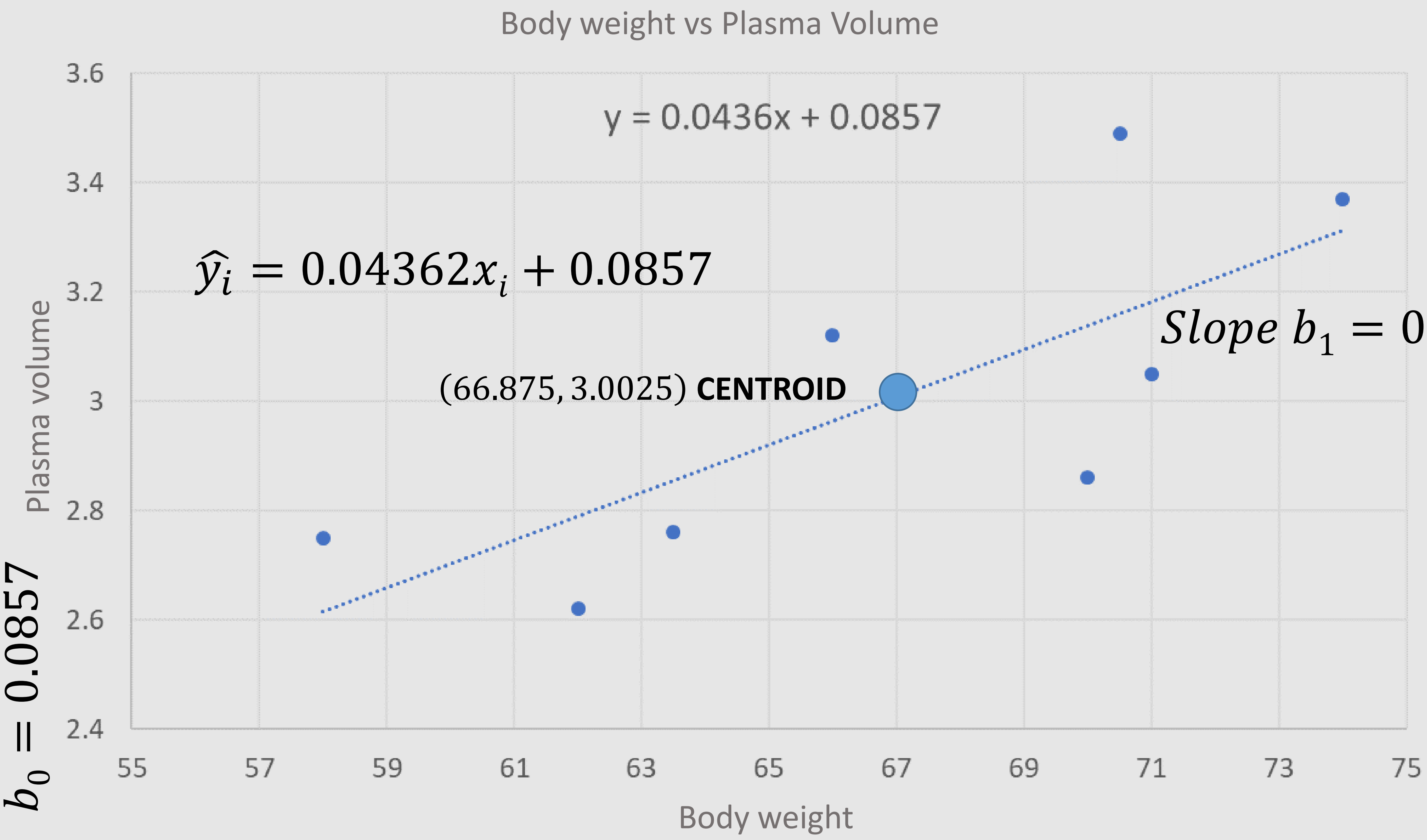
$b_1 = 0.04362$

slope

$$\widehat{y}_i = 0.0857 + 0.04362x_i$$

OR

$$\widehat{y}_i = 0.04362x_i + 0.0857$$

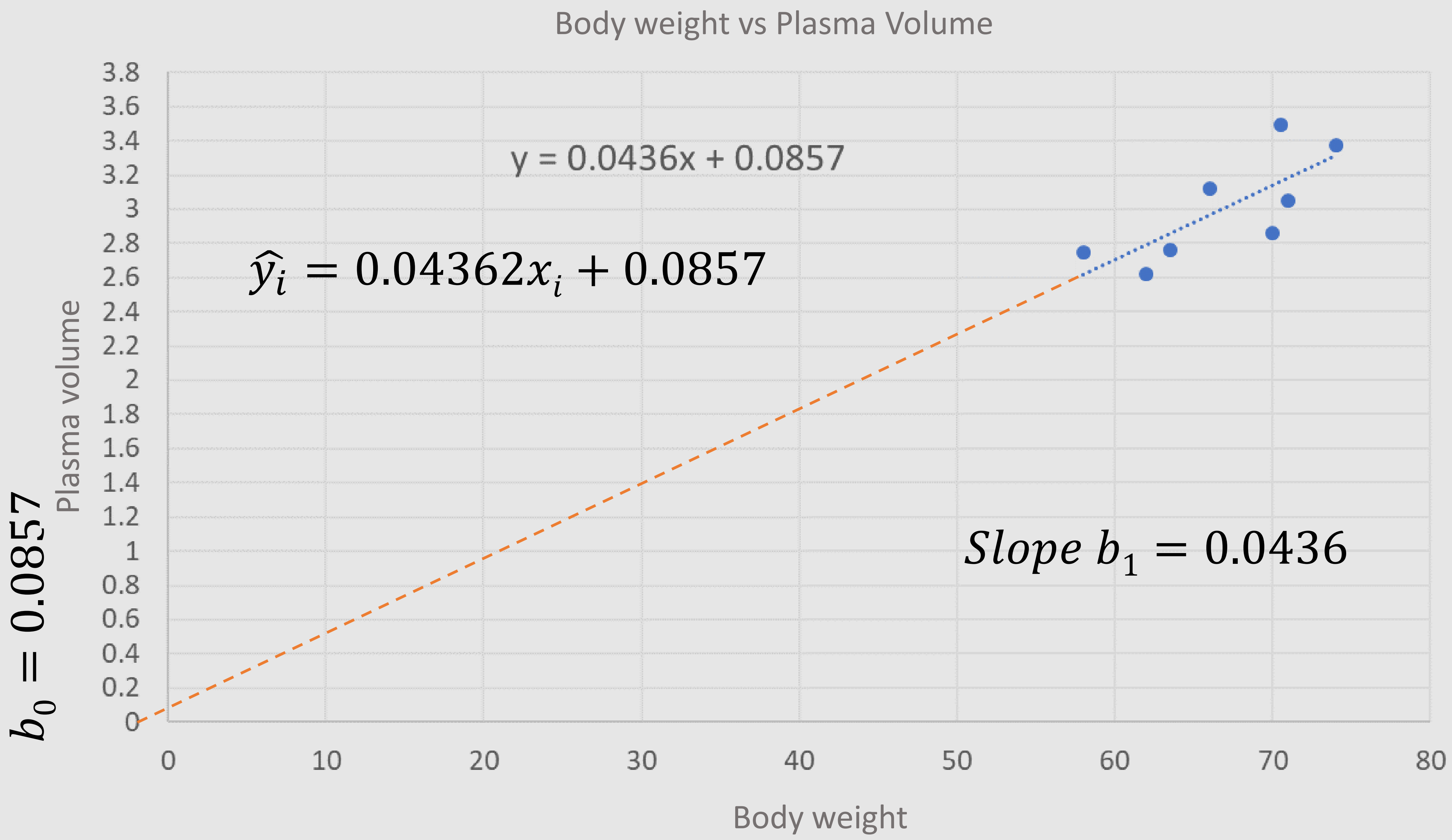


$$\hat{y}_i = 0.04362x_i + 0.0857$$

$(66.875, 3.0025)$ **CENTROID**

$Slope\ b_1 = 0.0436$

$b_0 = 0.0857$



$$\widehat{y_i} = 0.04362x_i + 0.0857$$

$$Slope\ b_1 = 0.0436$$

$$b_0 = 0.0857$$

Quick interpretation

$$\hat{y_i} = 0.04362x_i + 0.0857$$

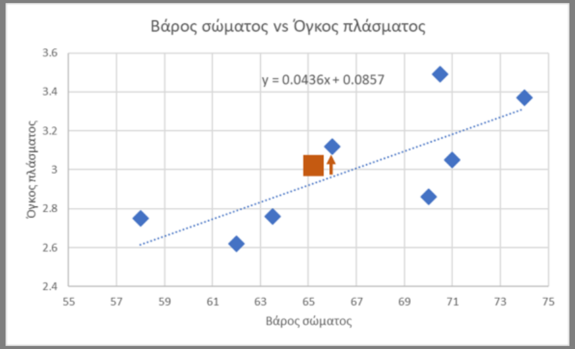
The slope of the regression line is **0.04362**. This means that for each additional kilogram (1 Kg.) of body weight, the plasma volume is predicted to increase by **0.04362 liters**.

The intercept of the regression line is **0.0857**. This represents the predicted plasma volume when the body weight is zero. While a body weight of zero kilograms is not realistic in practice, the intercept can still be useful for understanding the starting point of the regression model.

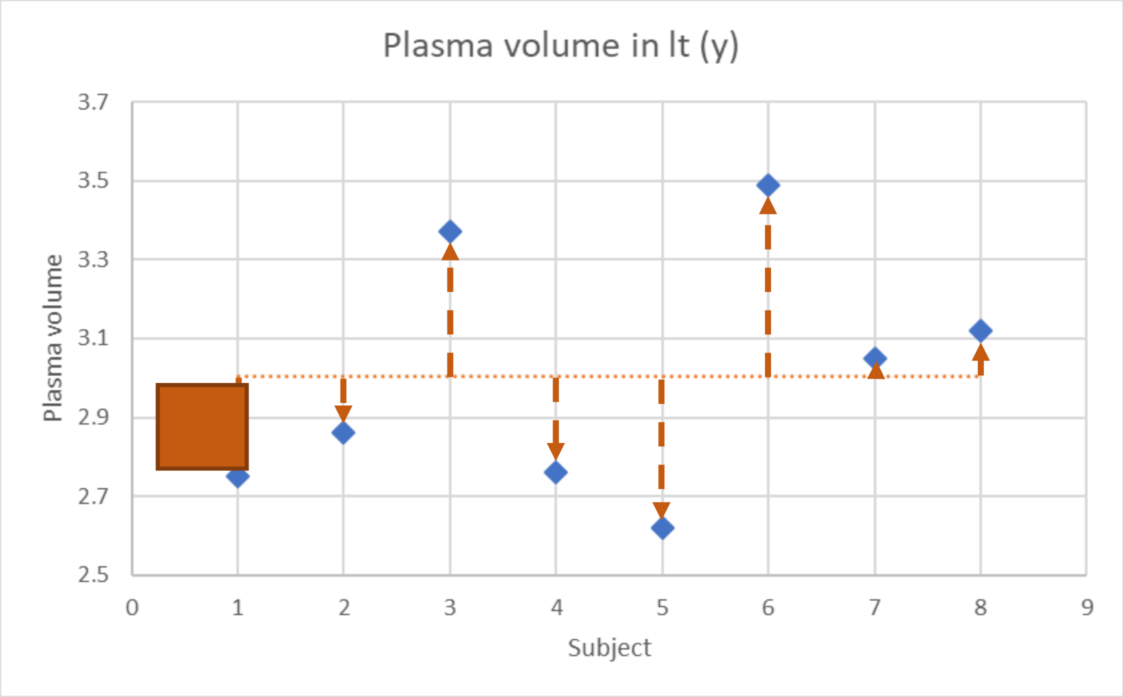
But …

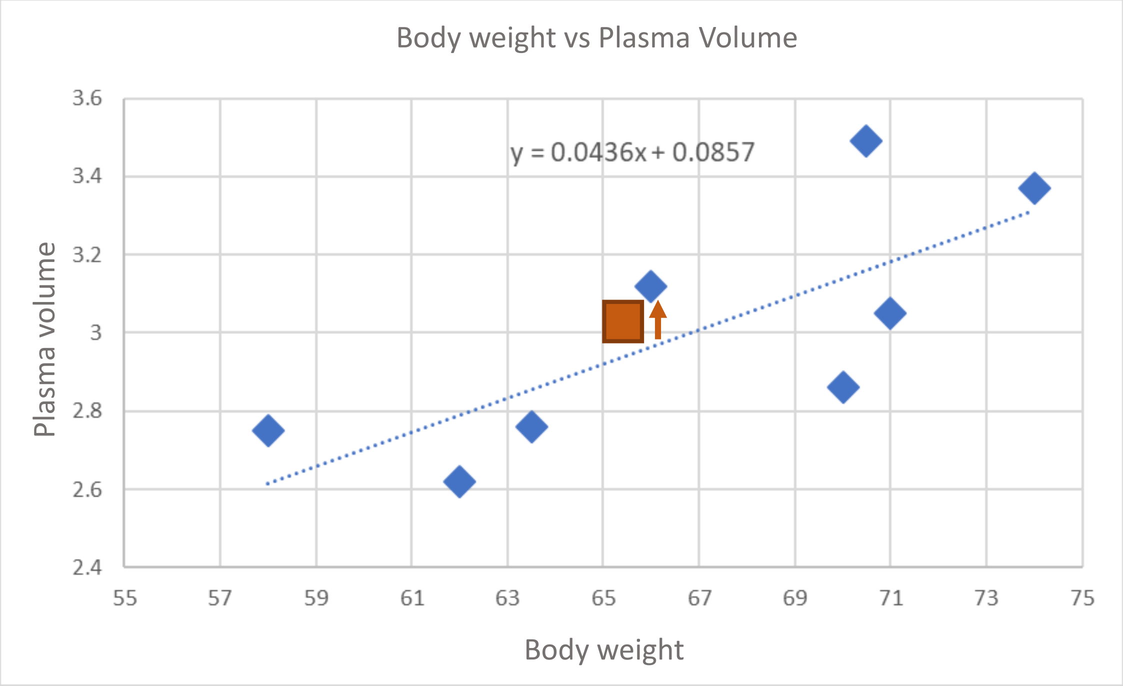
Is this regression line model any good?!?!?!

Coefficient of determination, R^2, R squared



Regression lines





$SSE = 0.6823$

$SSE = SST$

$SST = 0.6823$

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.

$SST = 0.6823$

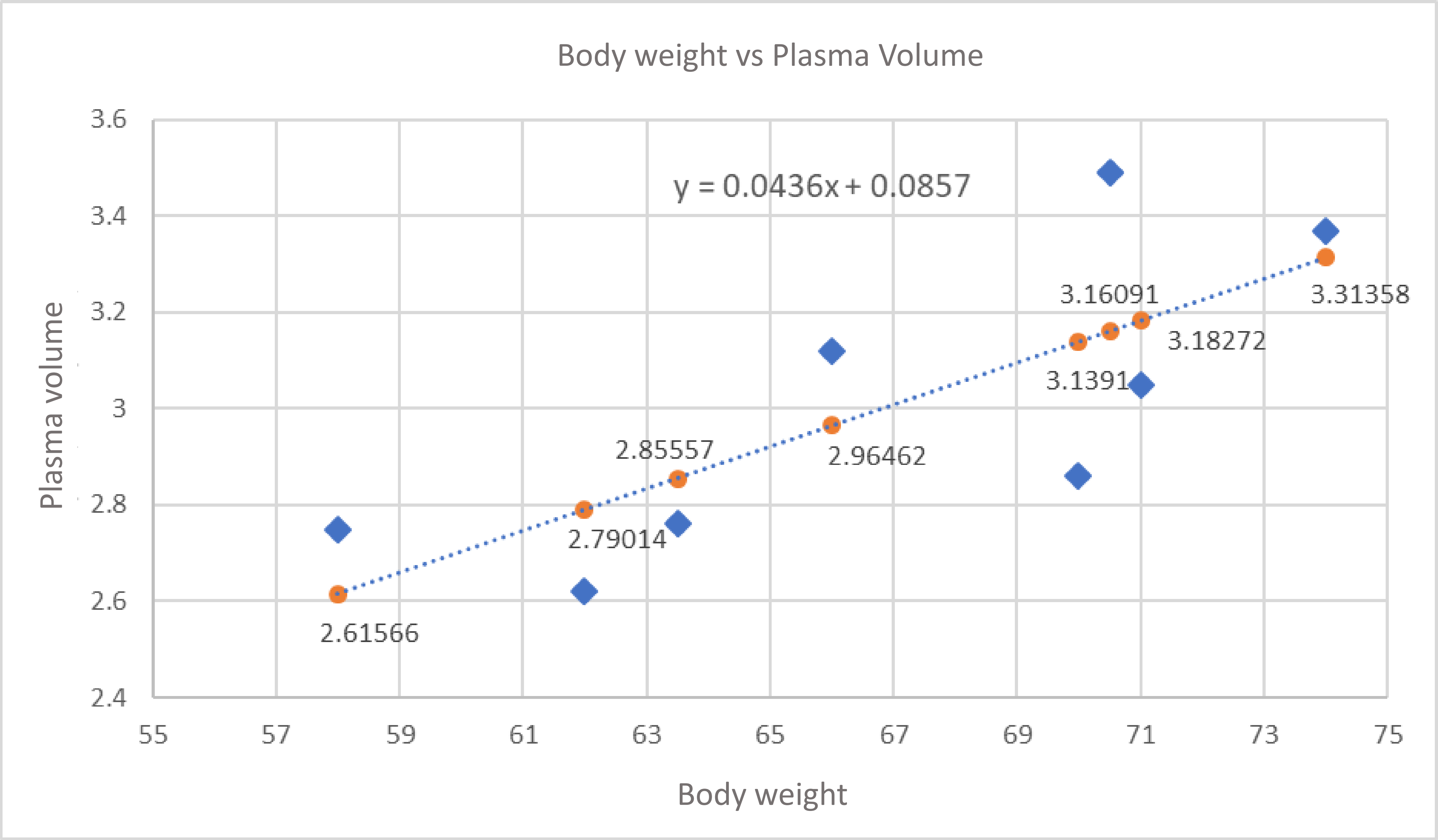
$\mathrm{SSE} = ?$

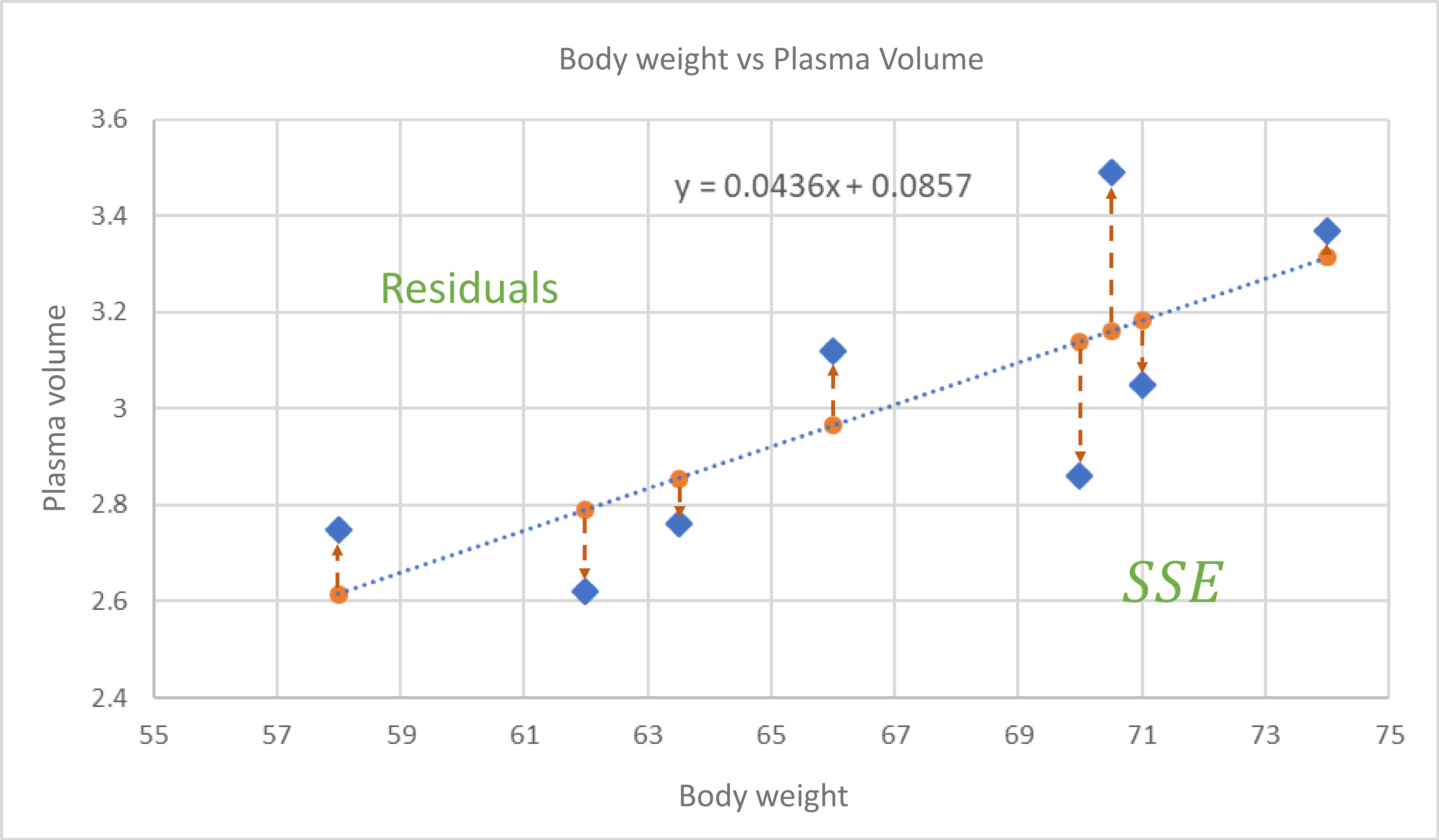
$SST - \mathrm{SSE} = SSR$

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.

Estimated regression values

Subject	Body Weight in Kg	Plasma volume in lt	$\hat{y}_i = 0.044x_i + 0.0857$	$\hat{y}_i$ (predicted plasma volume)
	x	y		
1	58	2.75	$\hat{y}_i = 0.044(58) + 0.0857$	2.61566
2	70	2.86	$\hat{y}_i = 0.044(70) + 0.0857$	3.1391
3	74	3.37	$\hat{y}_i = 0.044(74) + 0.0857$	3.31358
4	63.5	2.76	$\hat{y}_i = 0.044(63.5) + 0.0857$	2.85557
5	62	2.62	$\hat{y}_i = 0.044(62) + 0.0857$	2.79014
6	70.5	3.49	$\hat{y}_i = 0.044(70.5) + 0.0857$	3.16091
7	71	3.05	$\hat{y}_i = 0.044(71) + 0.0857$	3.18272
8	66	3.12	$\hat{y}_i = 0.044(66) + 0.0857$	2.96462
	$\bar{x} = 66.875$	$\bar{y} = 3.0025$	Observed vs Predicted	


Body weight vs Plasma Volume
y = 0.0436x + 0.0857
Plasma volume
3.6
3.4
3.2
3
2.8
2.6
2.4
Body weight
55
57
59
61
63
65
67
69
71
73
75
2.61566
2.79014
2.85557
2.96462
3.1391
3.16091
3.18272
3.31358



Residuals

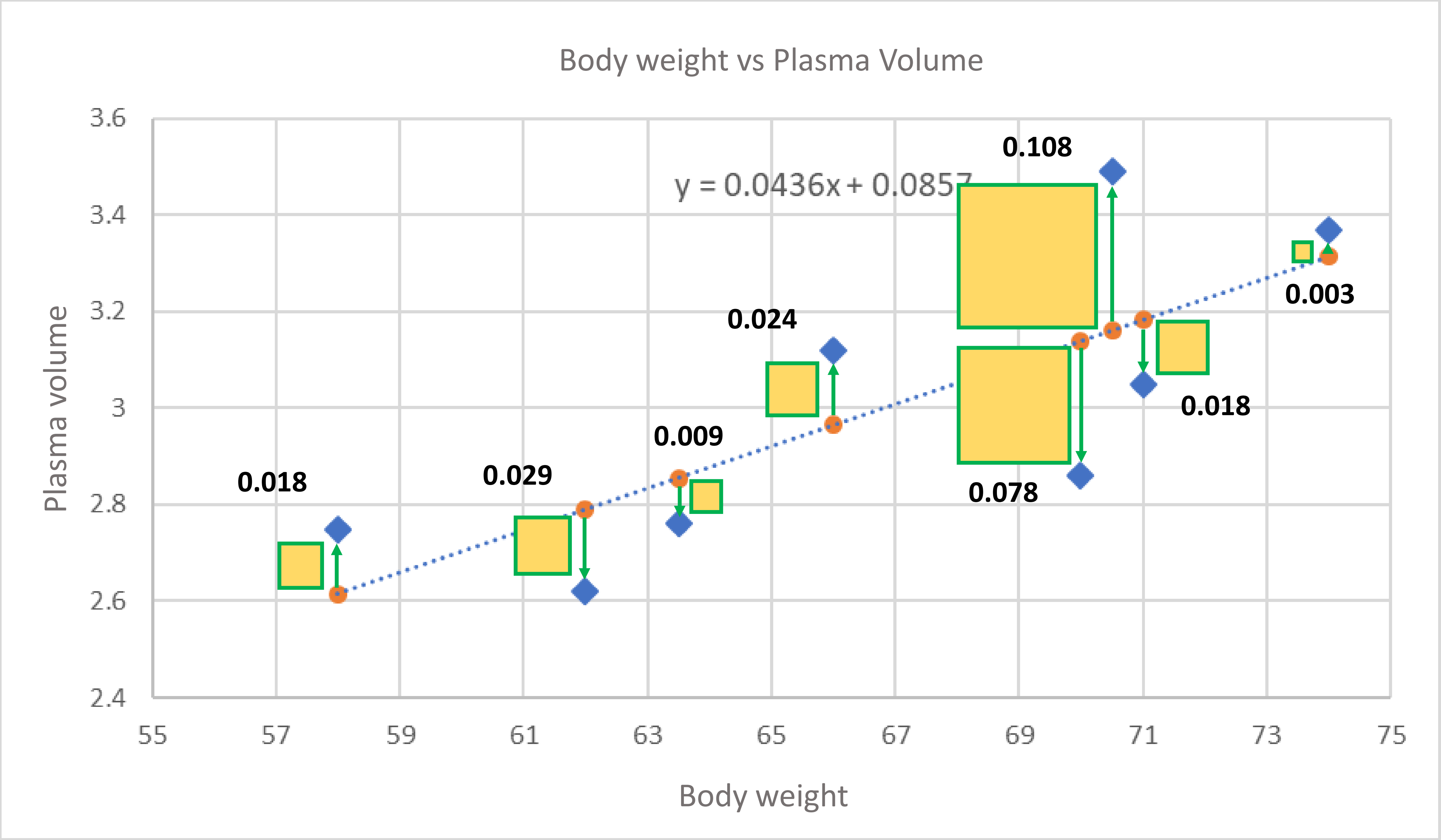
SSE

Regression error (residuals)

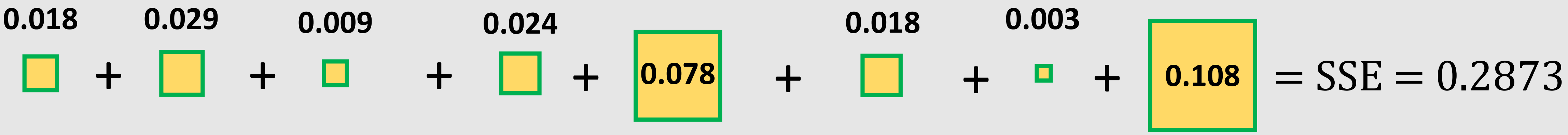
Subject	Body weight	Plasma volume	$\widehat{y_i}$ (predicted plasma volume)	Error: (observed - predicted) $(y_i - \widehat{y_i})$
	x	y		$(y_i - \widehat{y_i})$
1	58	2.75	2.61566	$2.75 - 2.61566 = 0.13434$
2	70	2.86	3.1391	$2.86 - 3.1391 = -0.2791$
3	74	3.37	3.31358	$3.37 - 3.31358 = 0.05642$
4	63.5	2.76	2.85557	$2.76 - 2.85557 = -0.09557$
5	62	2.62	2.79014	$2.62 - 2.79014 = -0.17014$
6	70.5	3.49	3.16091	$3.49 - 3.16091 = 0.32909$
7	71	3.05	3.18272	$3.05 - 3.18272 = -0.13272$
8	66	3.12	2.96462	$3.12 - 2.96462 = 0.15538$
	$\bar{x} = 66.875$	$\bar{y} = 3.0025$		

Regression squared error (residuals)

Subject	Body weight	Plasma volume	$\hat{y}_i$ (predicted plasma volume)	Error	Squared error
	x	y		$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
1	58	2.75	2.61566	0.13434	0.01804724
2	70	2.86	3.1391	−0.2791	0.07789681
3	74	3.37	3.31358	0.05642	0.00318322
4	63.5	2.76	2.85557	−0.09557	0.00913362
5	62	2.62	2.79014	−0.17014	0.02894762
6	70.5	3.49	3.16091	0.32909	0.10830023
7	71	3.05	3.18272	−0.13272	0.0176146
8	66	3.12	2.96462	0.15538	0.02414294
	$\bar{x} = \mathbf{66.875}$	$\bar{y} = \mathbf{3.0025}$		**SSE =**	$\sum = \mathbf{0.2873}$


Body weight vs Plasma Volume
y = 0.0436x + 0.0857
Plasma volume
Body weight
3.6
3.4
3.2
3
2.8
2.6
2.4
55
57
59
61
63
65
67
69
71
73
75
0.018
0.029
0.009
0.024
0.108
0.078
0.018
0.003

Dependent variable and independent variable (plasma volume as a function of body weight)

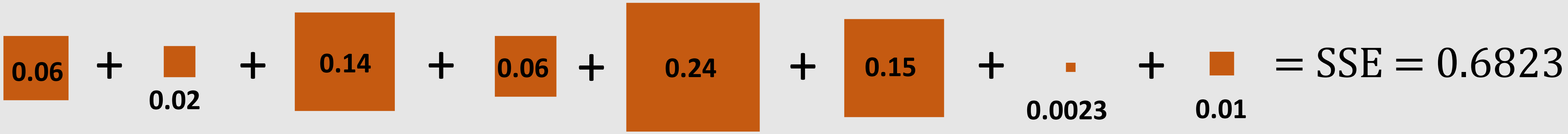


0.018 + 0.029 + 0.009 + 0.024 + 0.078 + 0.018 + 0.003 + 0.108 $= \text{SSE} = 0.2873$

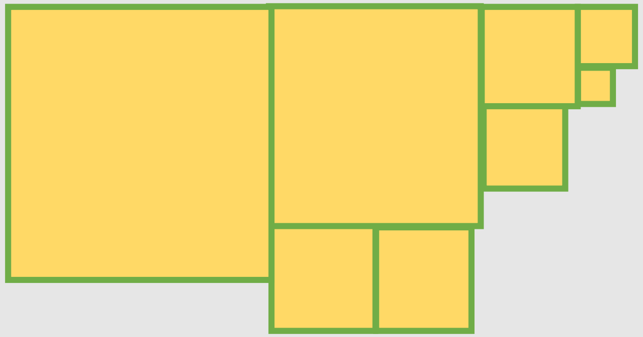
Sum of squared errors comparison

Dependent variable only (plasma volume)

$$SSE = SST$$

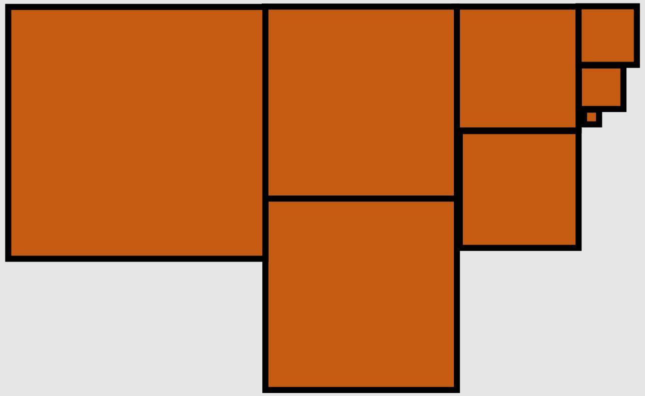


0.06 + 0.02 + 0.14 + 0.06 + 0.24 + 0.15 + 0.0023 + 0.01 $= \text{SSE} = 0.6823$



$= 0.2873$

Sum of squared errors comparison



$= 0.6823$

So when we conducted the regression, the SSE decreased from 0.68 to 0.2873. That is, 0.2873 of the sum of squares was explained or allocated to ERROR.

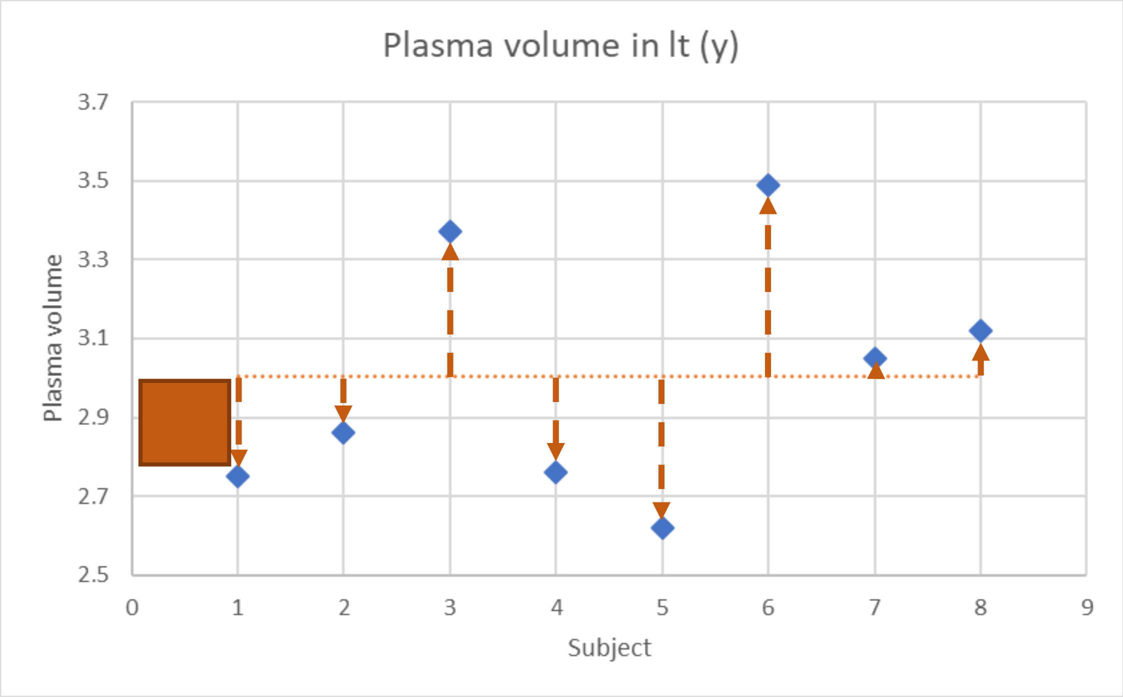
Where did the other 0.3927 go?

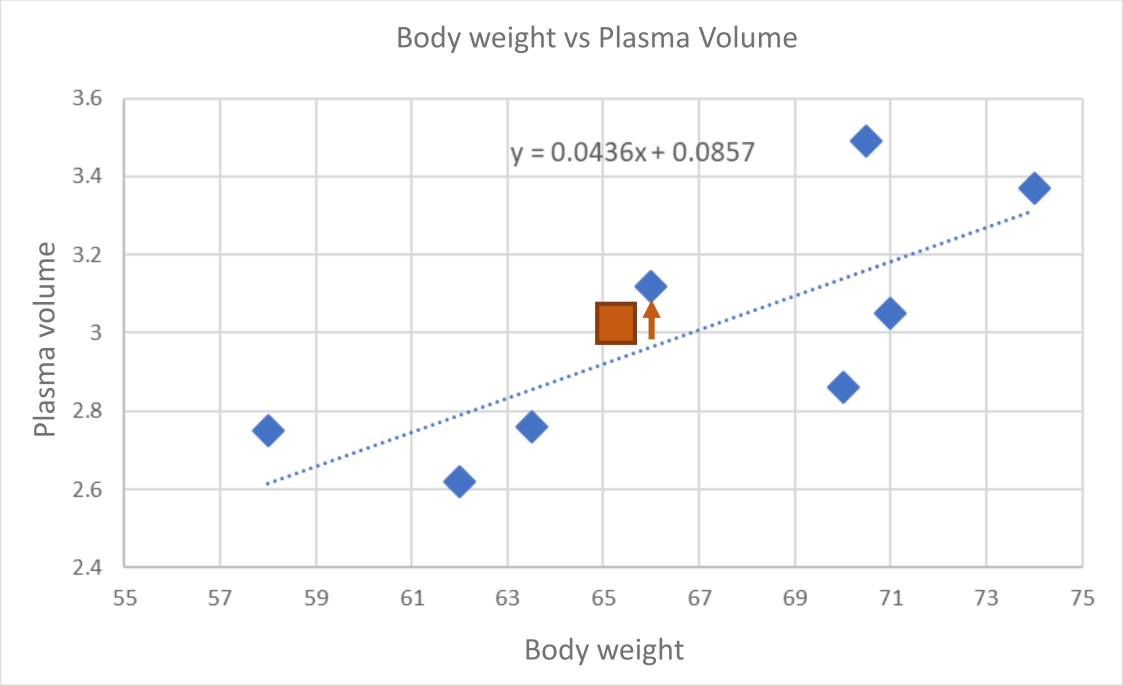
The 0.3927 is the sum of squares due to regression (SSR).

$$SST = SSR + SSE$$

$$0.6823 = 0.3927 + 0.2873$$

Regression lines





$SSE = 0.6823$

$SSE = SST$

$SST = 0.6823$

$SST = 0.6823$

$0.6823 - 0.2873 = SSR$

$0.6823 - 0.2873 = 0.3927$

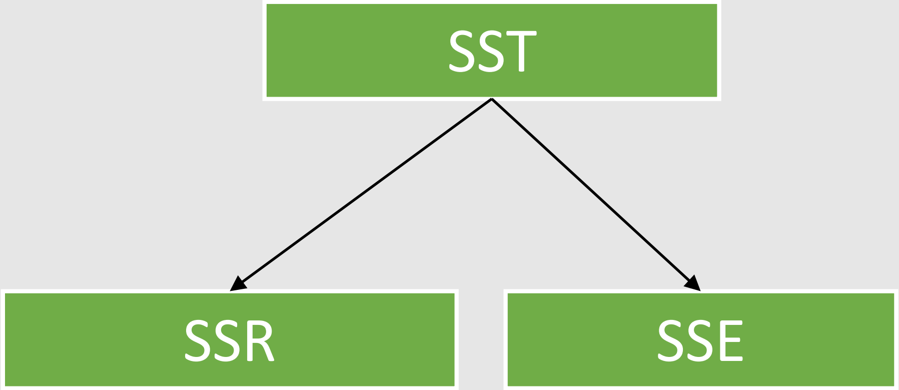
$\text{SSE} = 0.2873$

$SSR = 0.3927$

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 SSR, and SSE.



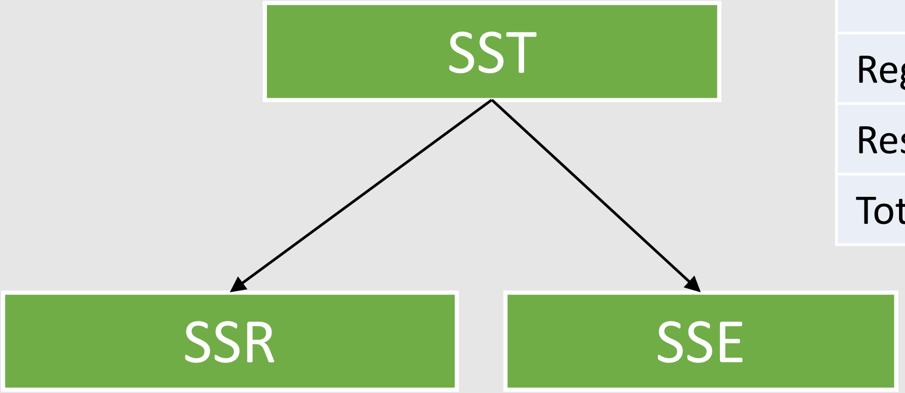
If SSR is large, it uses up more of SST and therefore SSE is smaller relative to SST. The coefficient of determination quantifies this ratio as a percentage.

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

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 SSR, and SSE.



ANOVA					
	df	SS	MS	F	Significance F
Regression	1	0.390684388	0.39068439	8.160065927	0.028930913
Residual	6	0.287265612	0.0478776		
Total	7	0.67795			

r^2 interpretation

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

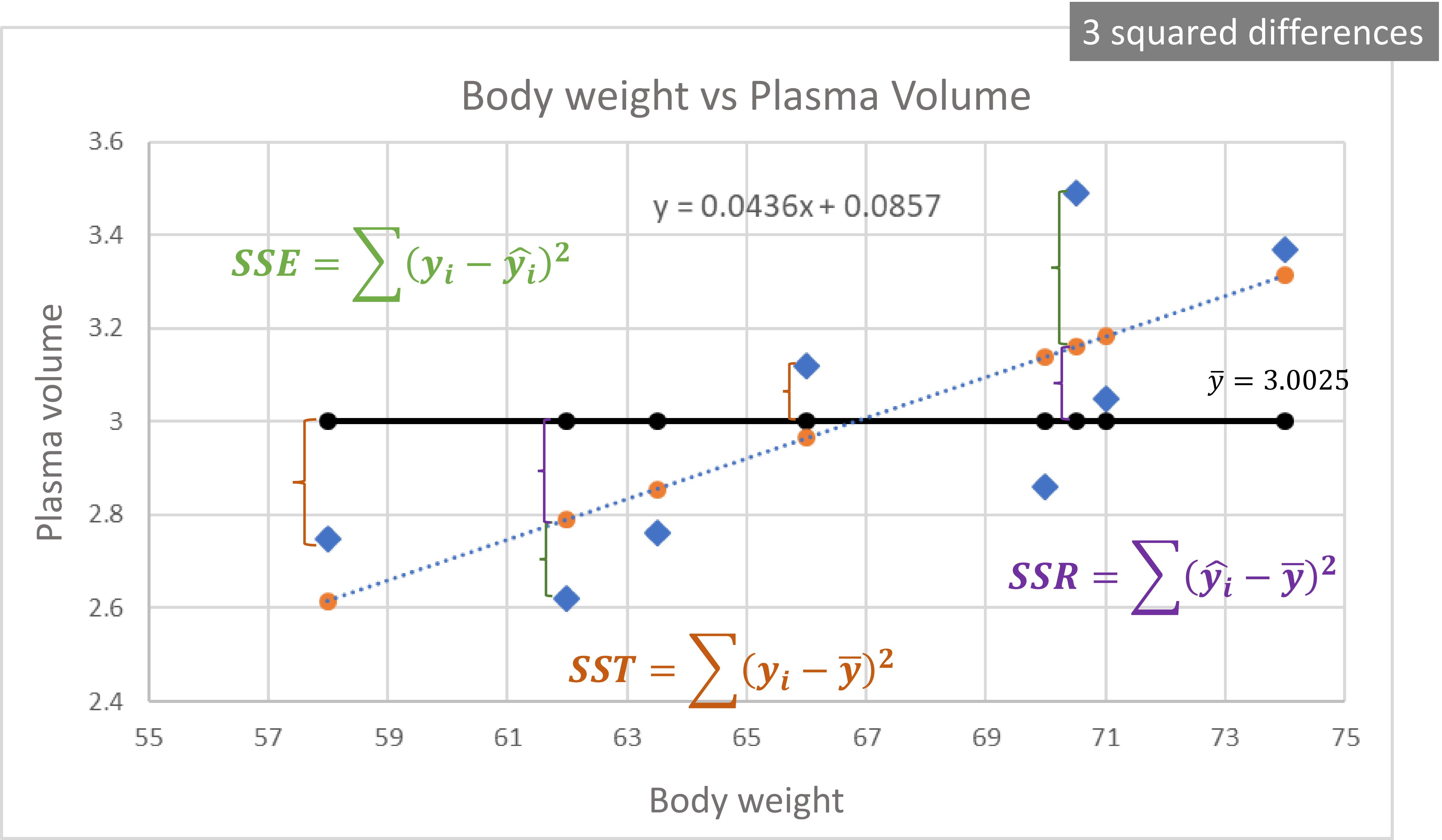
GOOD FIT!

$$\text{Coefficient of determination} = r^2 = \frac{0.3927}{0.68}$$

$$\text{Coefficient of determination} = r^2 = 0.5775 \text{ or } 57.75\%$$

We can conclude that 57.75% of the total sum of squares can be explained using the estimated regression equation to predict the plasma volume. The remaining, 42.25%, is error.

3 squared differences


Body weight vs Plasma Volume
y = 0.0436x + 0.0857
$SSE = \sum(y_i - \hat{y}_i)^2$
$\bar{y} = 3.0025$
$SSR = \sum(\hat{y}_i - \bar{y})^2$
$SST = \sum(y_i - \bar{y})^2$
Plasma volume
Body weight