

The background of the slide is a dark teal color with a complex, semi-transparent financial chart overlay. The chart includes candlestick patterns, several overlapping line graphs in shades of blue and orange, and some numerical data points. The overall aesthetic is technical and data-driven.

Linear Regression

Machine Learning Types

Supervised Learning

Classification

Logistic Regression
K-Nearest Neighbors (K-NN)
Support Vector Machine (SVM)
Kernel SVM
Naïve Bayes
Decision Tree Classification
Random Forest Classification

Predicting a categorical variable
Input: labeled data set
Output: Discrete values

Regression

Linear Regression
Multiple Linear Regression
Polynomial Regression
Support Vector Regression (SVR)
Decision Tree Regression
Random Forest Regression

Predicting a numeric variable
Input: Label data set
Output: continuous values

Unsupervised Learning

Clustering

K-Means Clustering
Hierarchical Clustering

Identify a pattern or
groups of similar
objects

Reinforcement Learning

Decision Making

Upper Confidence (UCB)
Thompson Sampling

Artificial Intelligence (AI):
Q-learning
R learning

solve interacting problems where the data observed up to time t is considered to decide which action to take at time $t + 1$. It is also used for Artificial Intelligence

Linear Regression

- *Linear Regression is a commonly used supervised Machine Learning algorithm that predicts continuous values.*
- *It assumes that there is a linear relationship present between dependent and independent variables.*
- *In simple words, it finds the best fitting line/plane that describes two or more variables.*

Linear Regression

Coefficient Equations

Prediction Equation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Sample Slope

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Sample Y-intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Example:

Parameter estimation solution:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n X_i\right)\left(\sum_{i=1}^n Y_i\right)}{n}}{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}} = \frac{16204 - \frac{(391)(200)}{5}}{31219 - \frac{(391)^2}{5}} = 0.877$$

tempreature	icre-cream sold
92	50
85	47
60	20
71	40
83	43

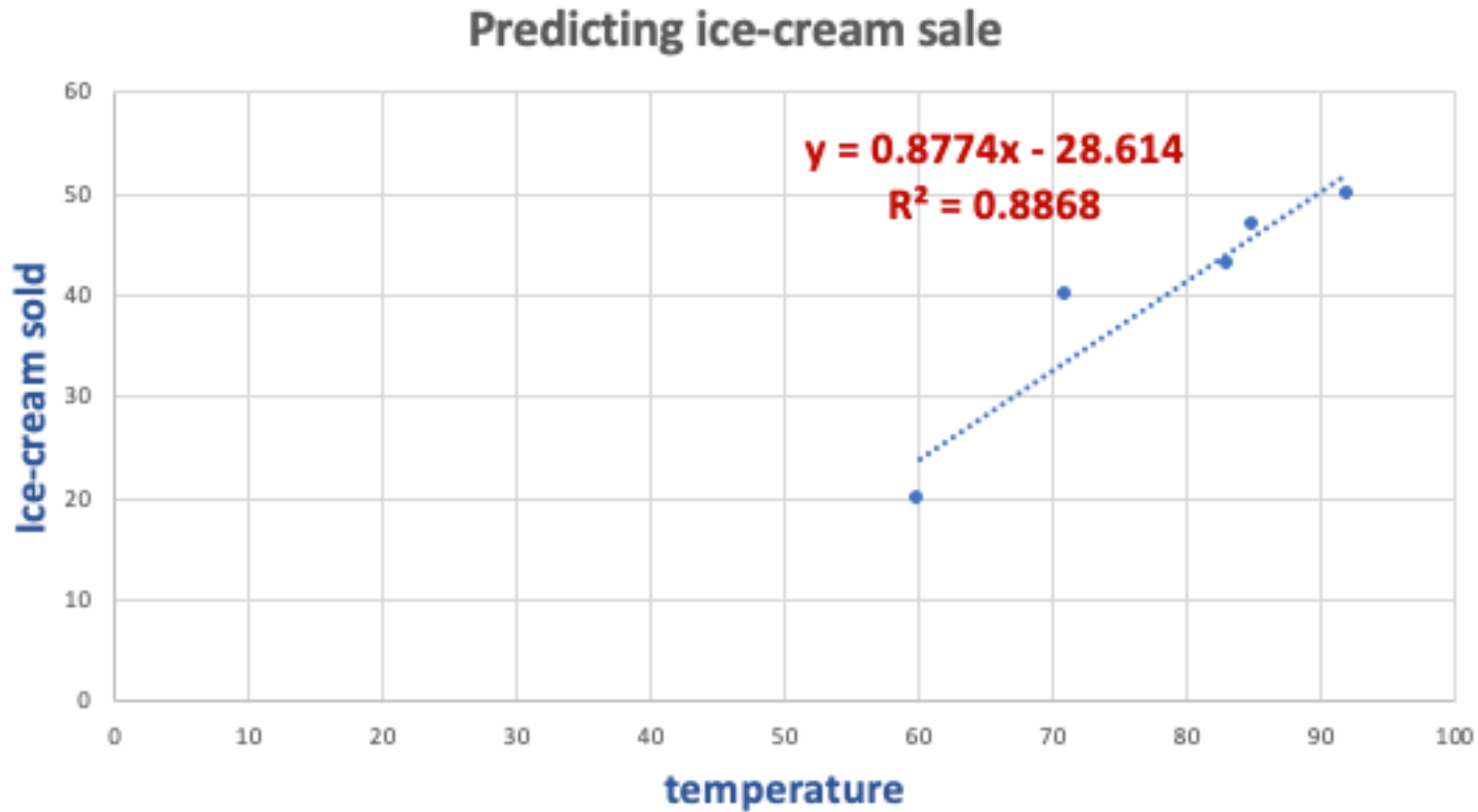
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 40 - (0.877 * 78.2) = -28.61$$

	Xi	Yi	(Xi)^2	(Yi)^2	XiYi
	92	50	8464	2500	4600
	85	47	7225	2209	3995
	60	20	3600	400	1200
	71	40	5041	1600	2840
	83	43	6889	1849	3569
sum:	391	200	31219	8558	16204

Prediction equation:

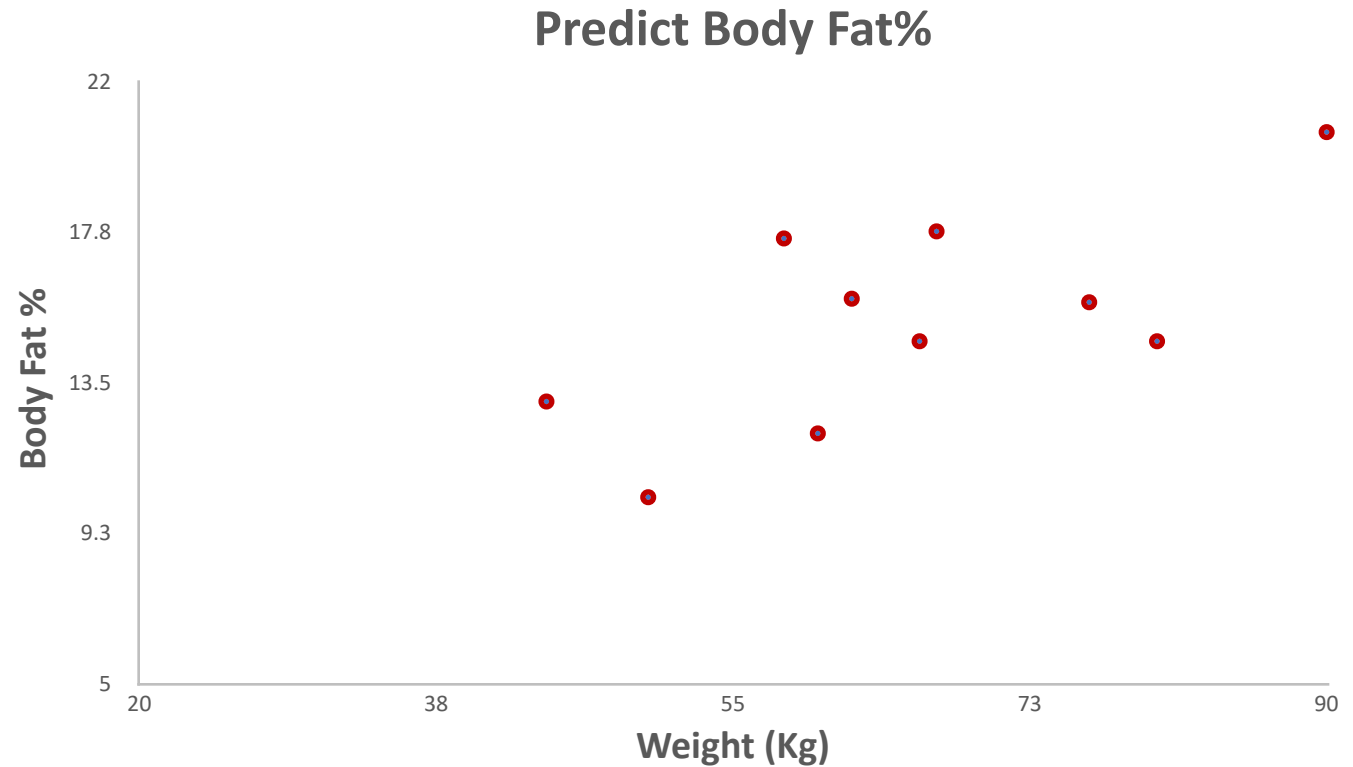
$$Y = 0.877 X_1 - 28.61$$

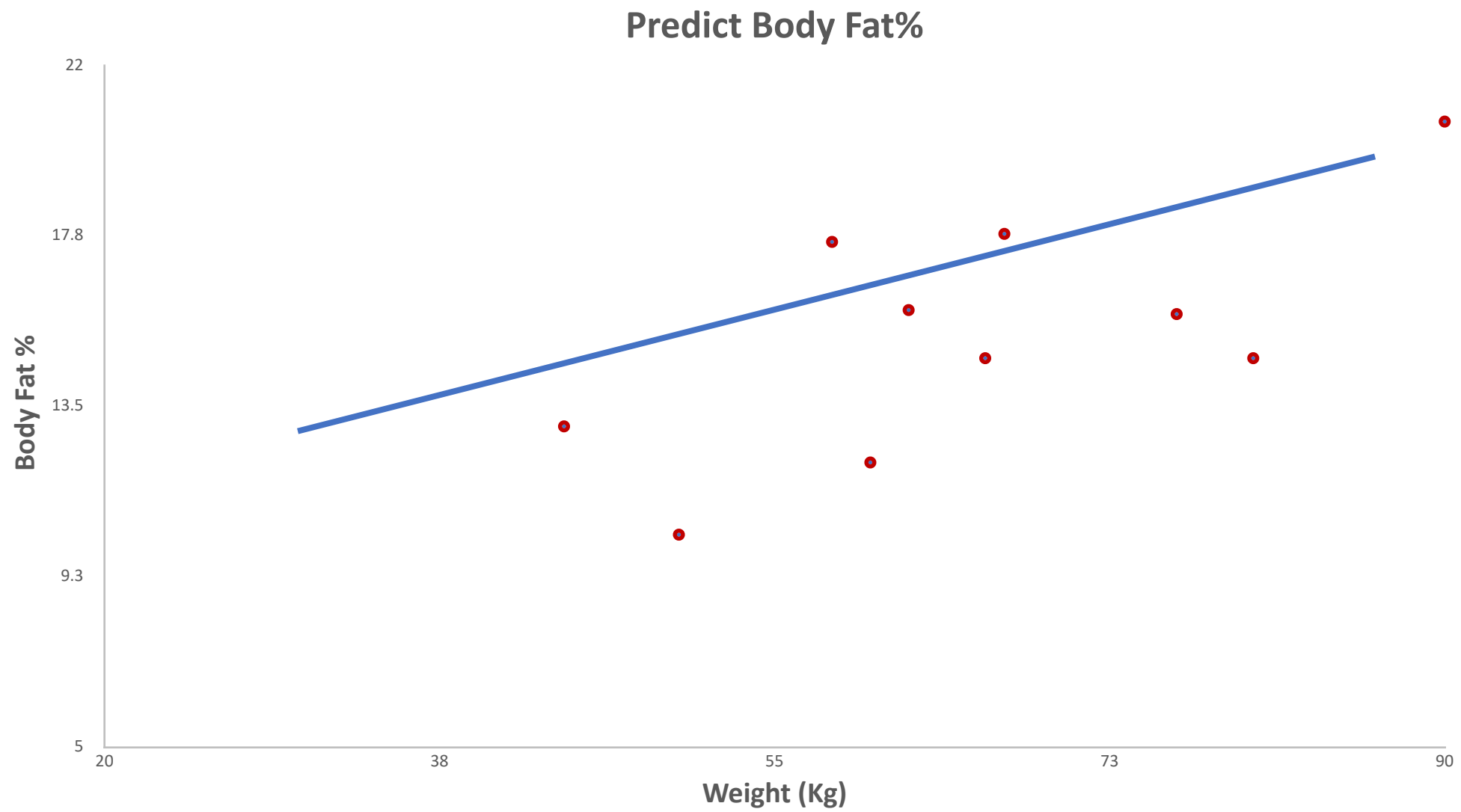
Example continue:

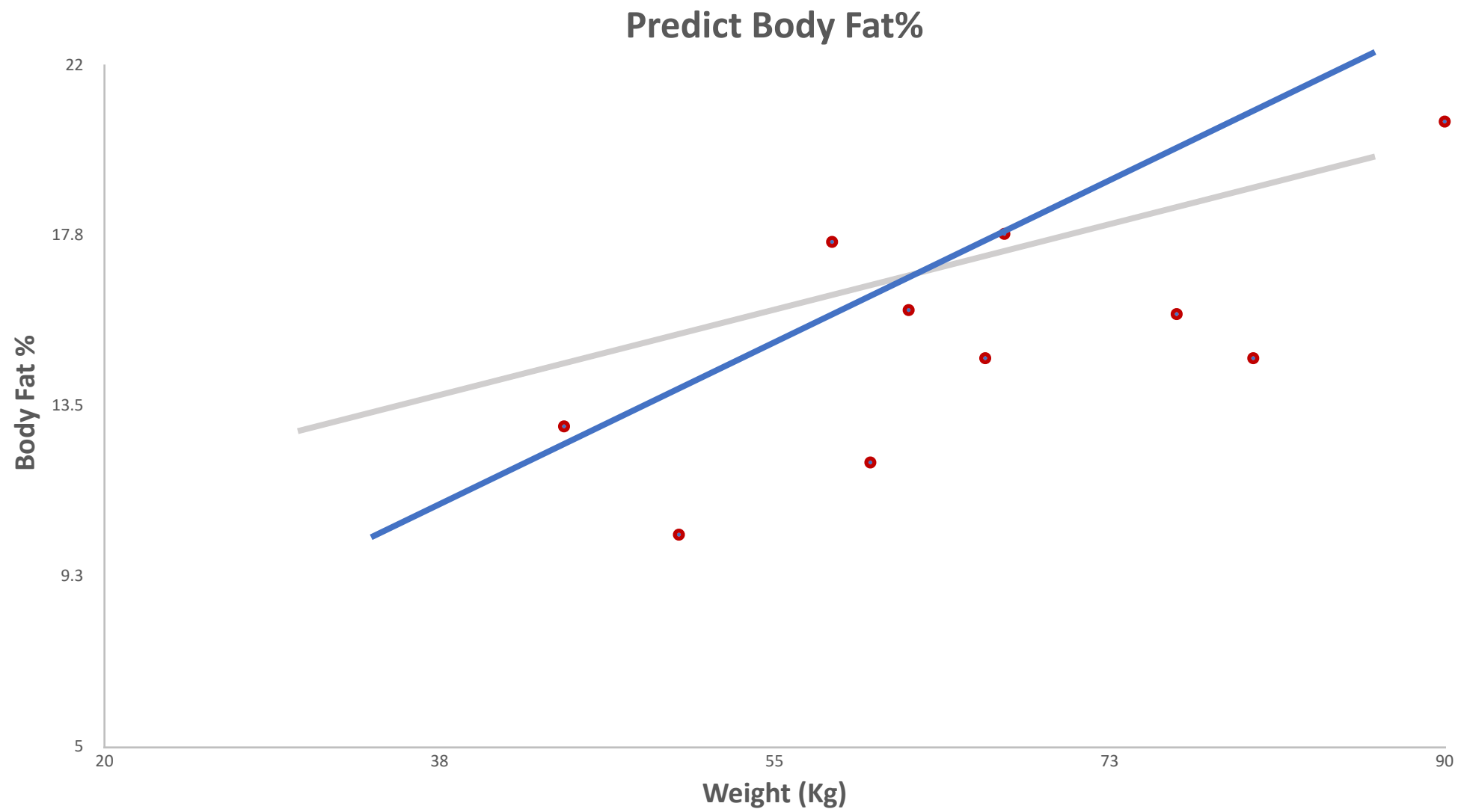


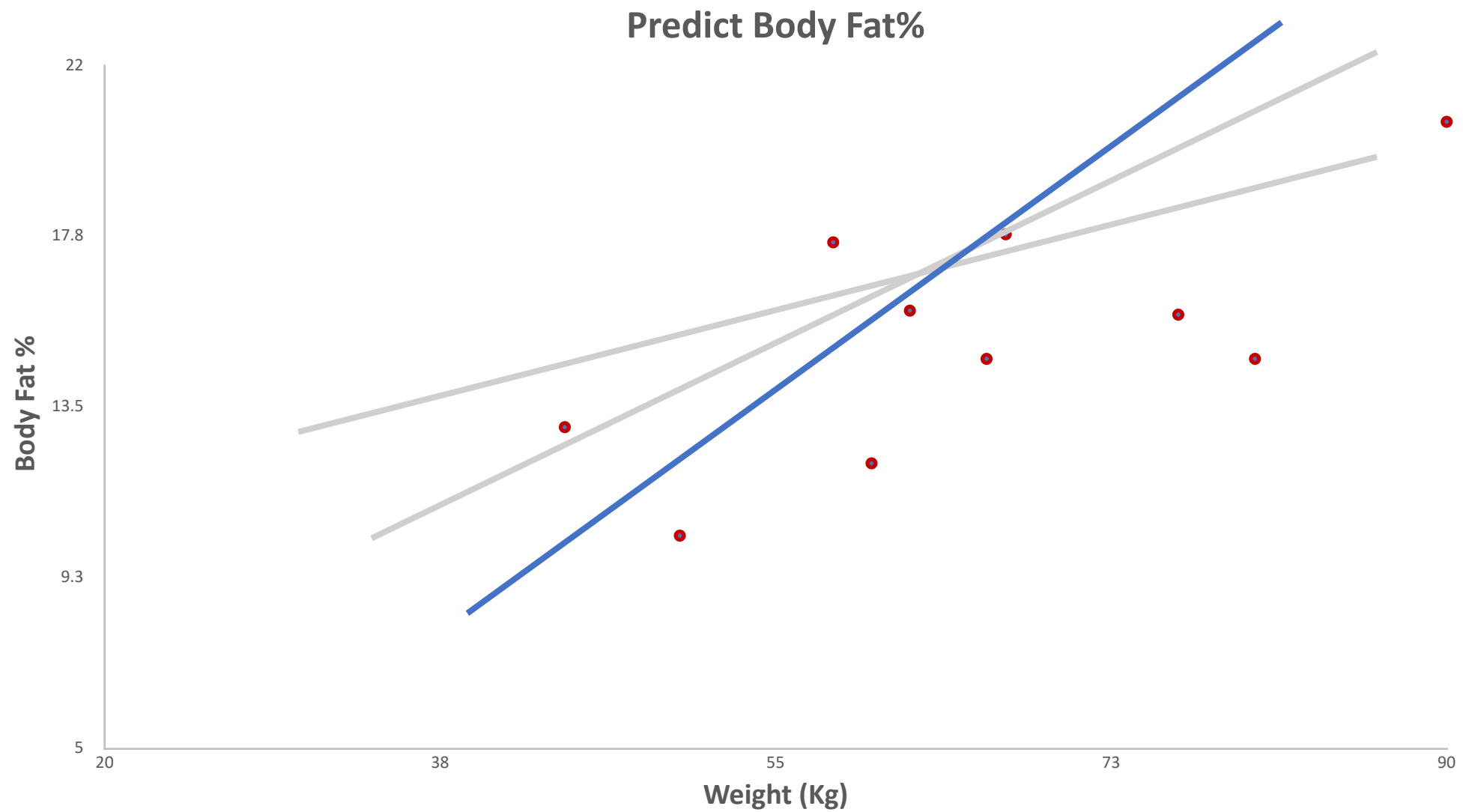
Example: predicting Body Fat Percent

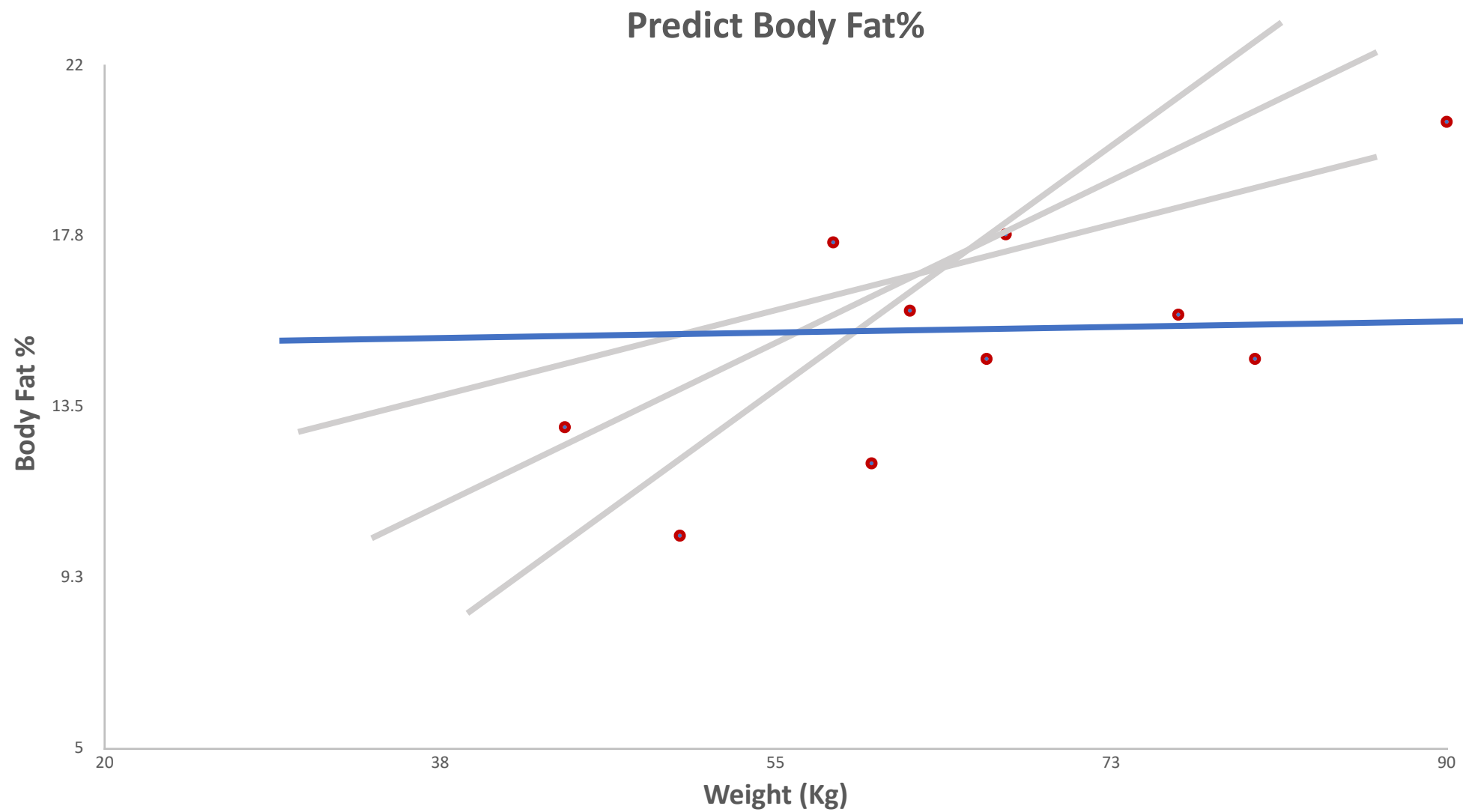
weight(kg)	Body Fat%
44	13
50	10.3
58	17.6
60	12.1
62	15.9
66	14.7
67	17.8
76	15.8
80	14.7
90	20.6



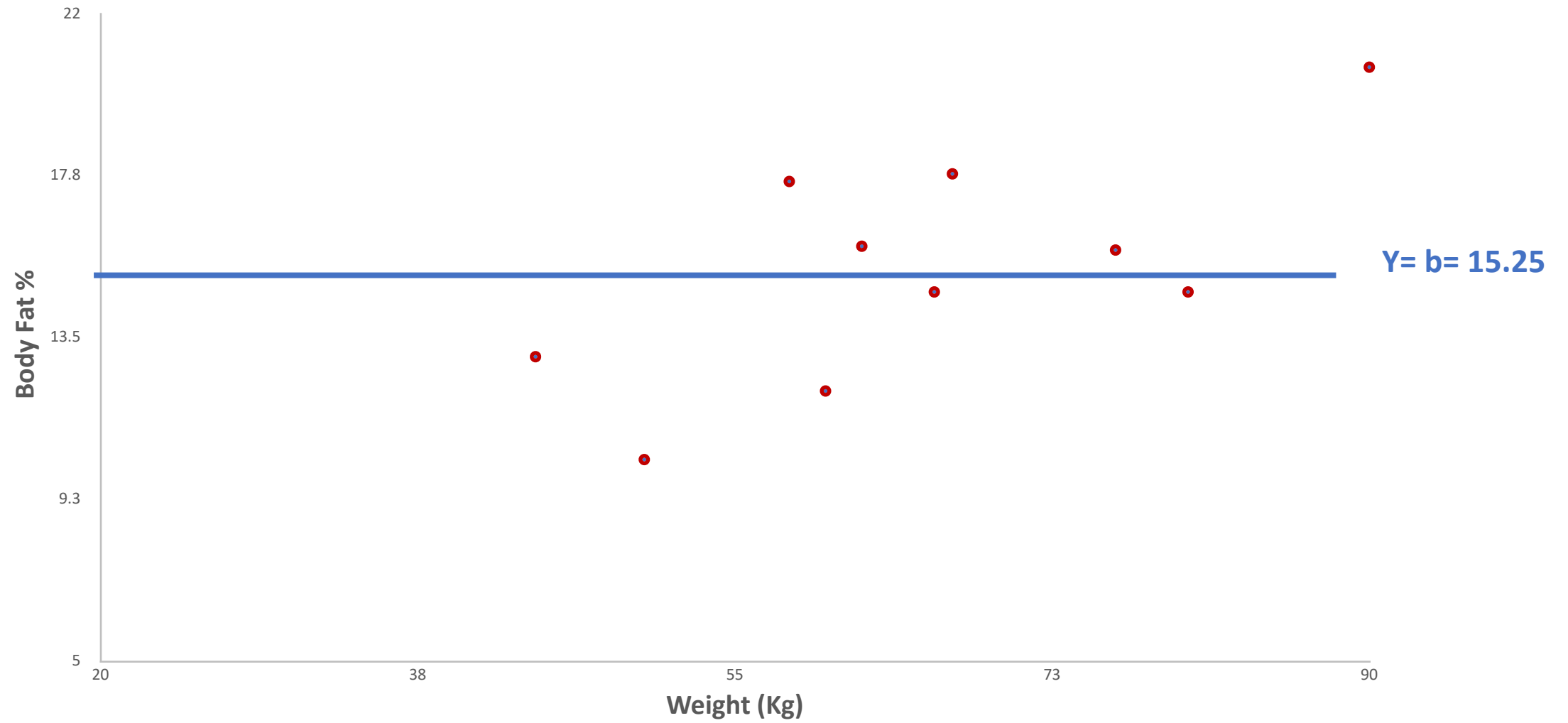




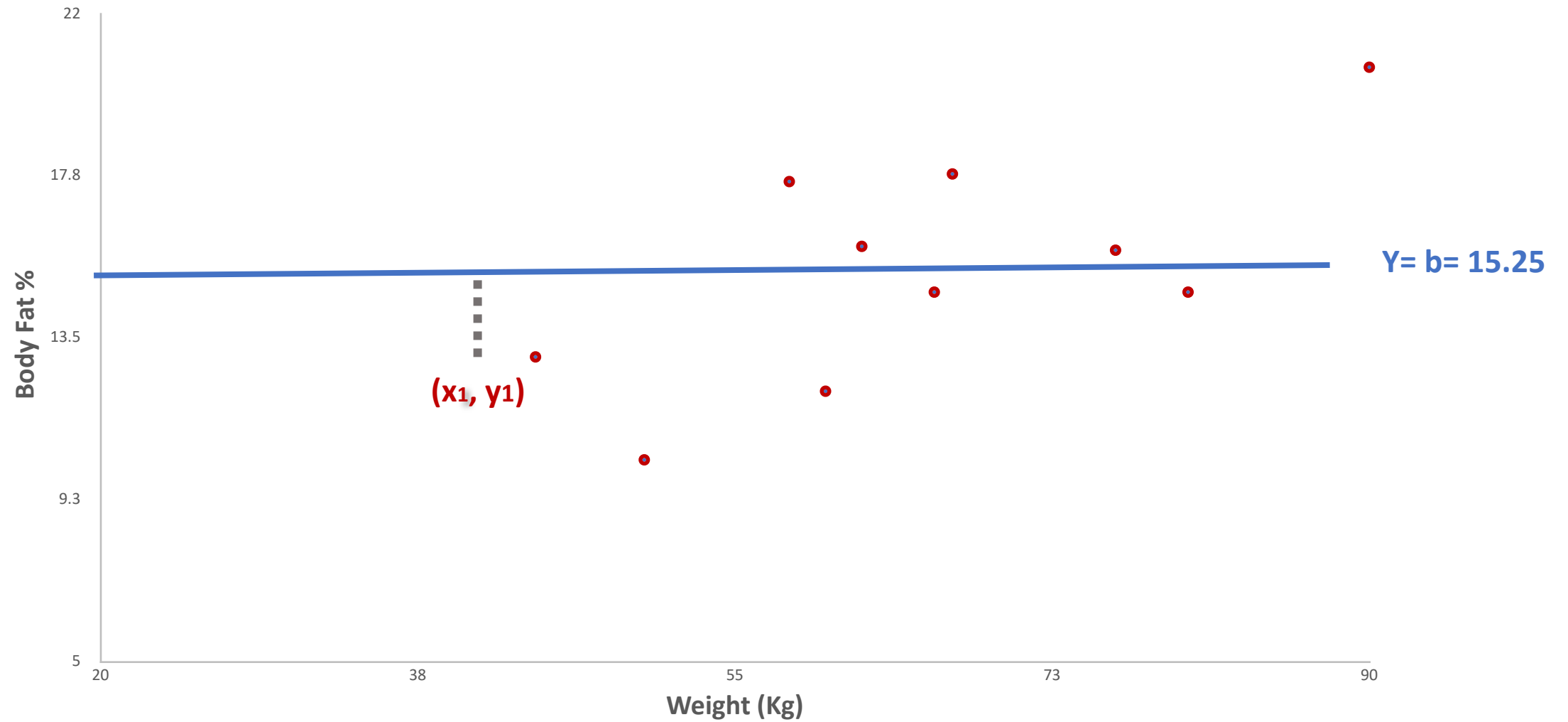




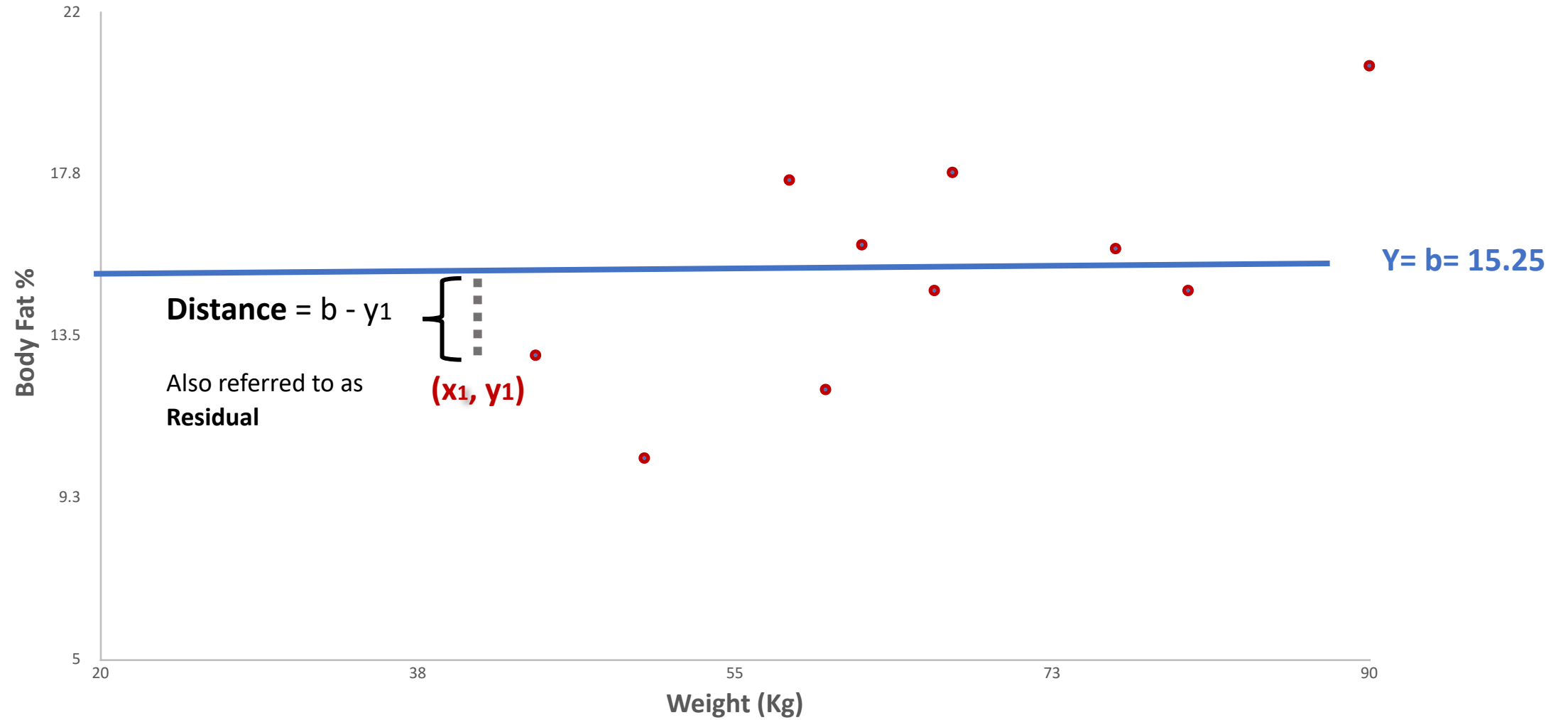
Predict Body Fat%



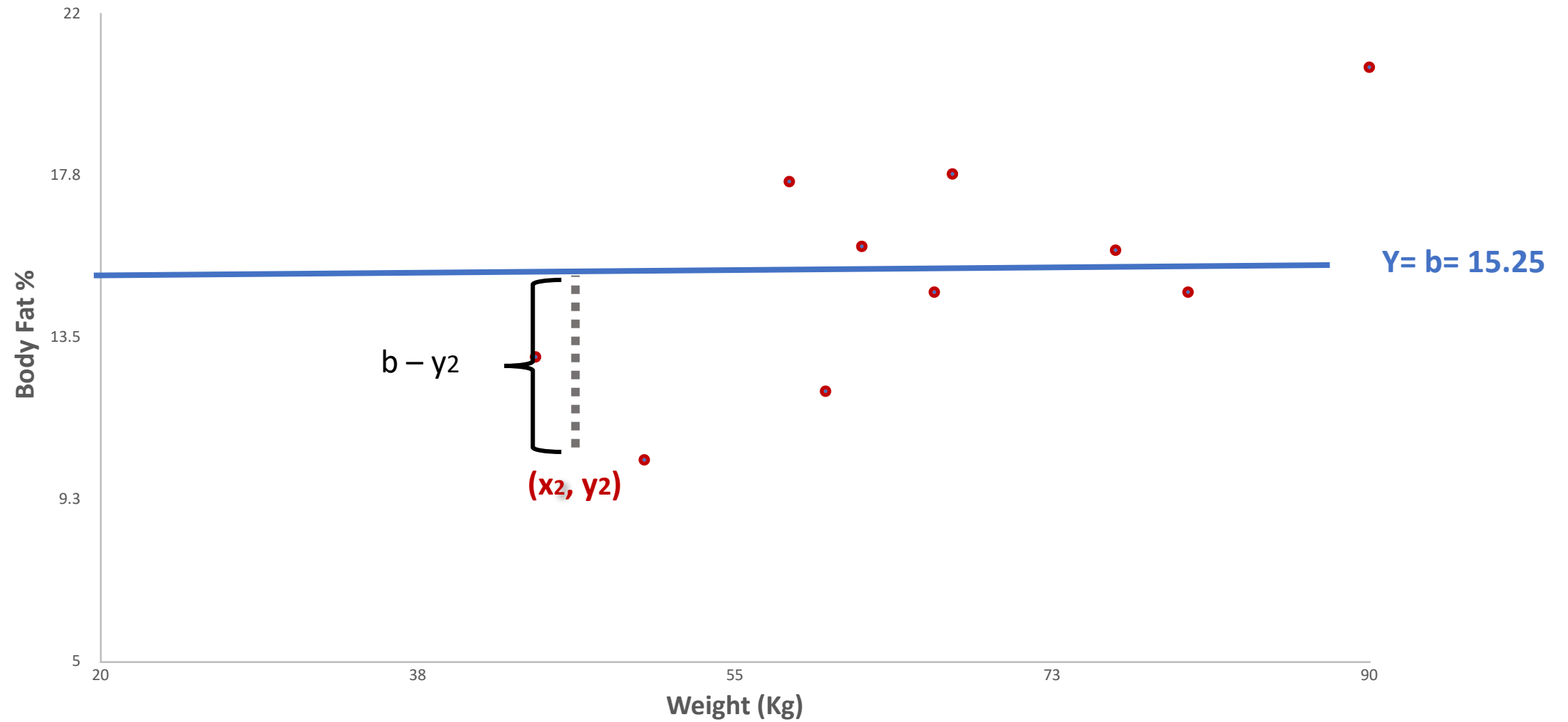
Predict Body Fat%



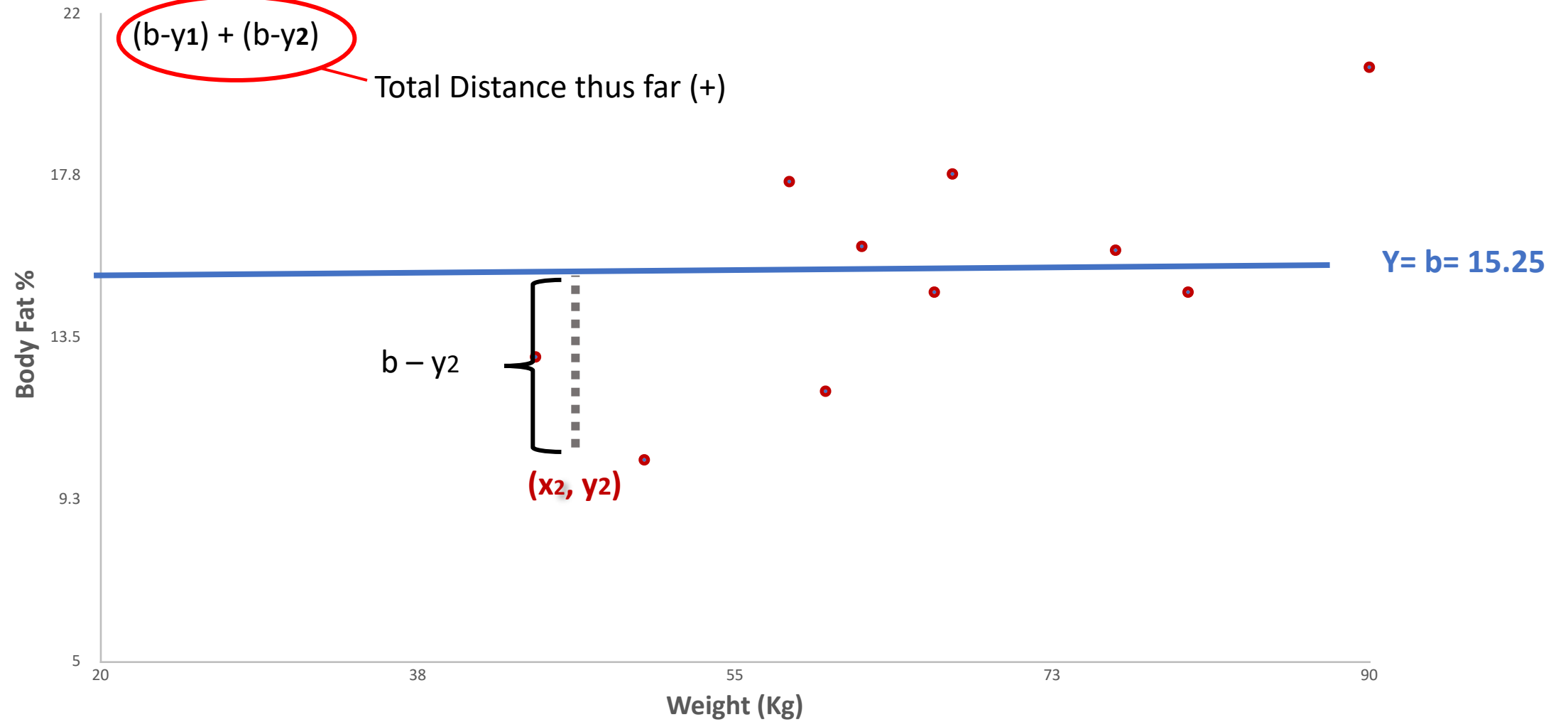
Predict Body Fat%



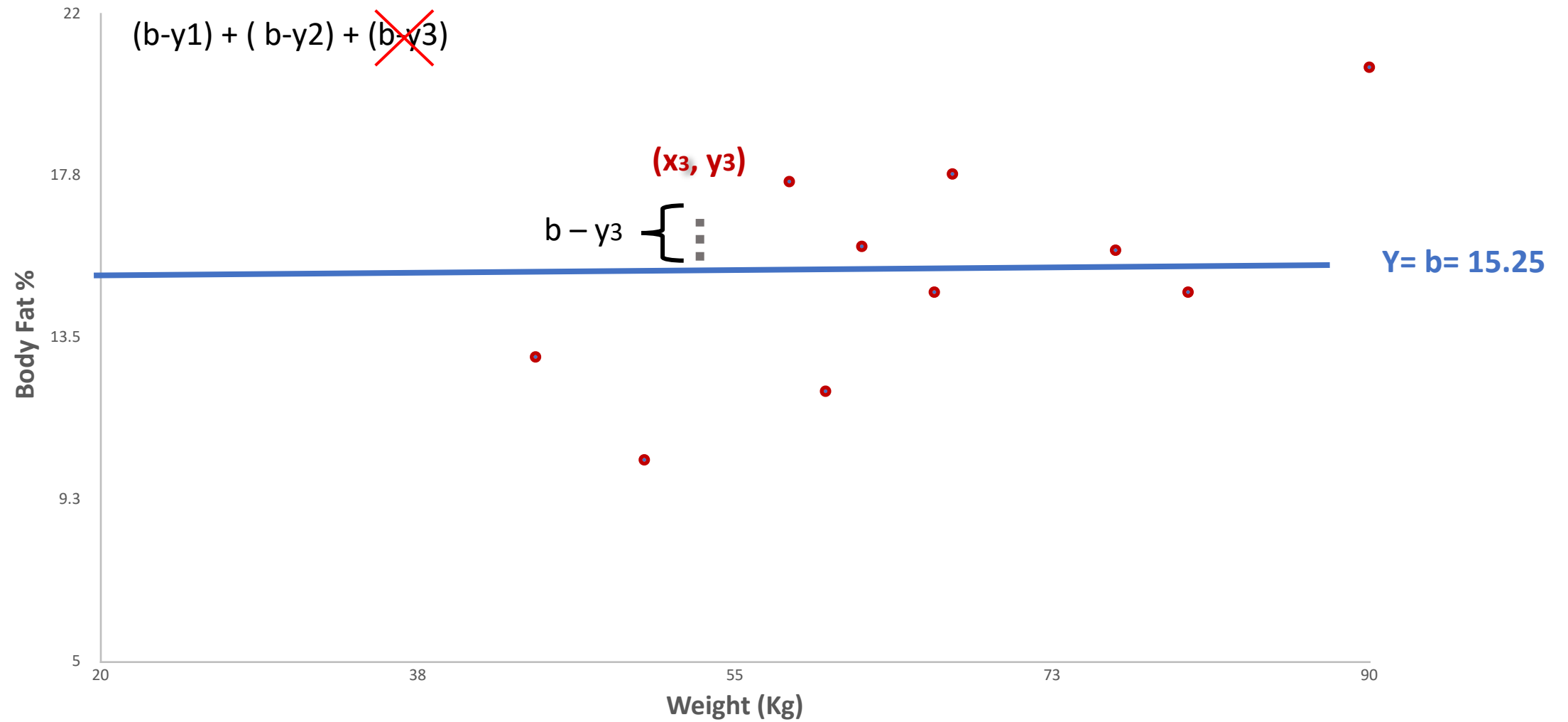
Predict Body Fat%



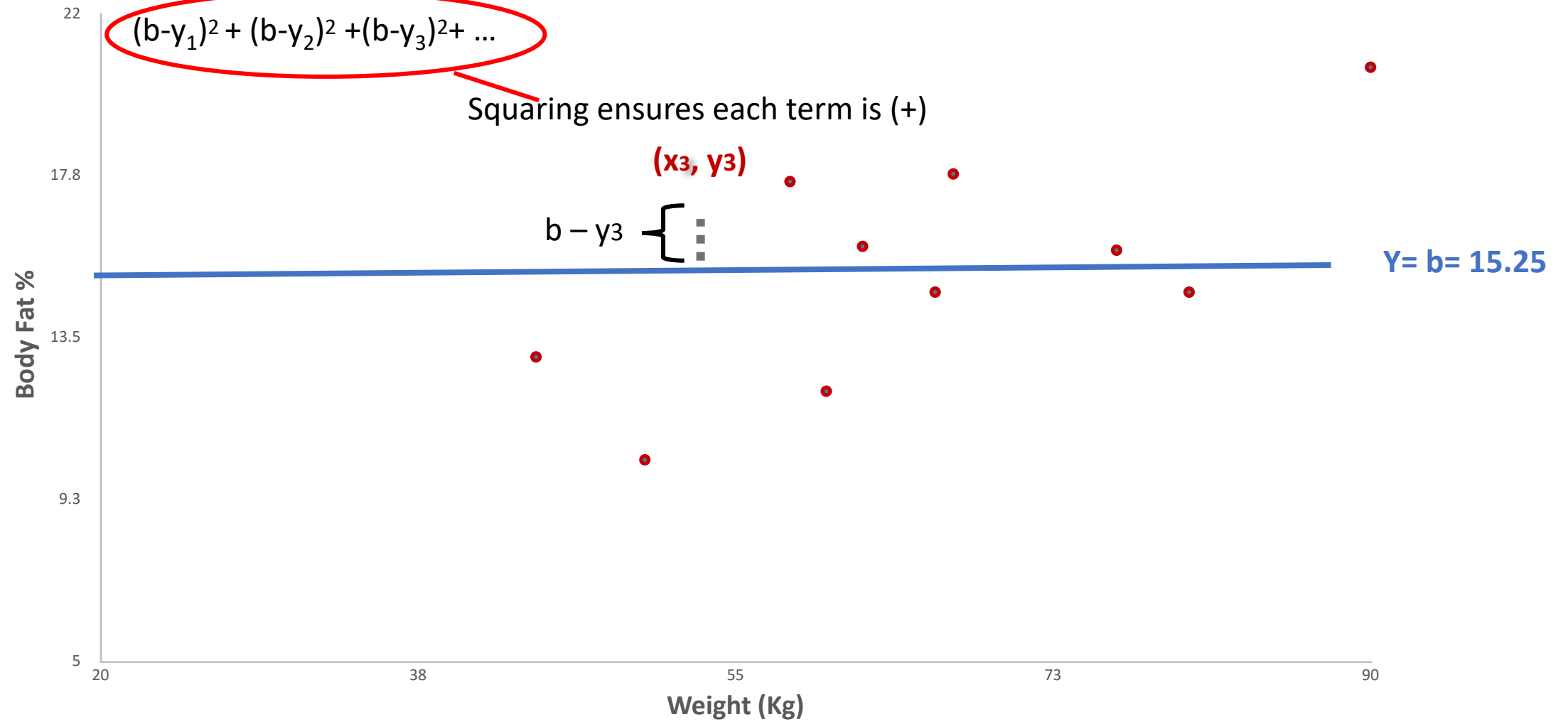
Predict Body Fat%



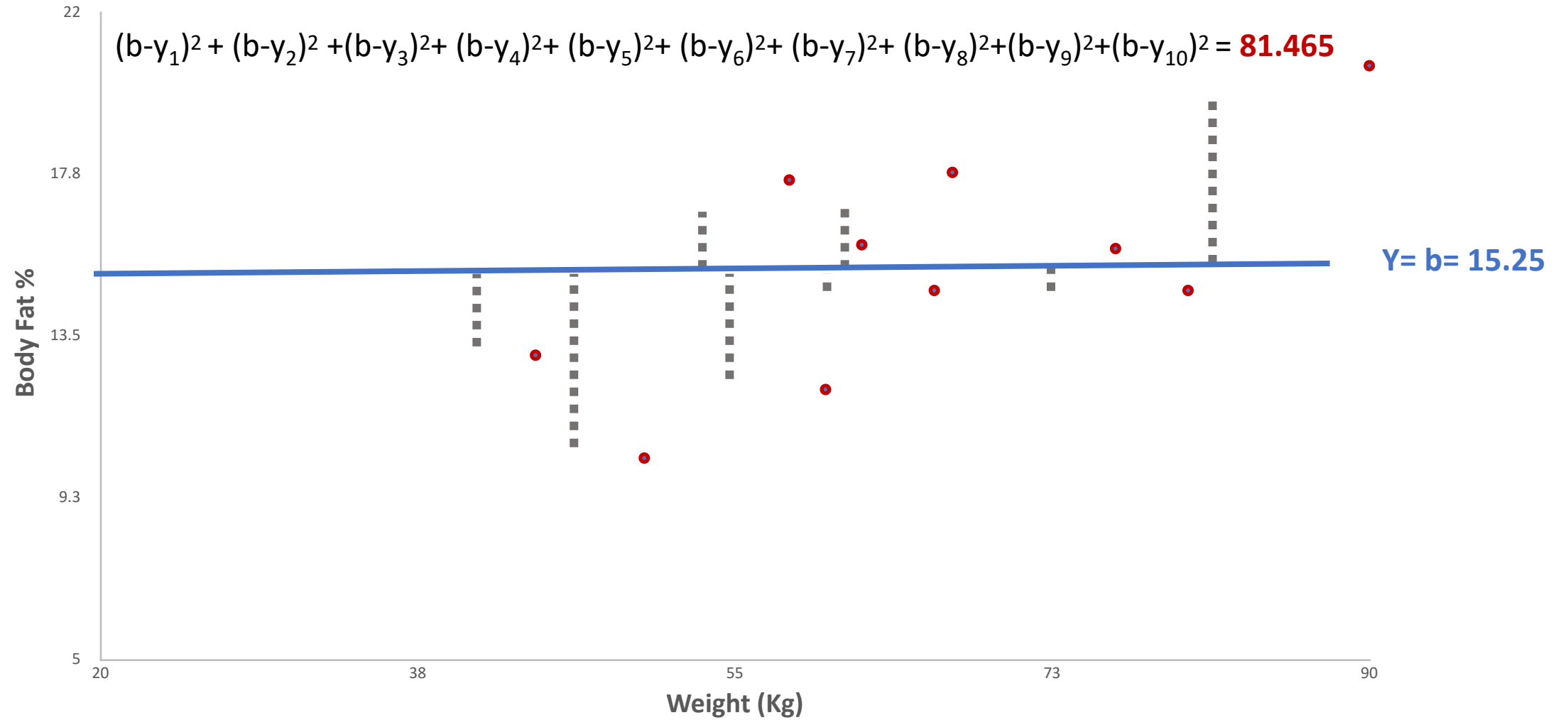
Predict Body Fat%



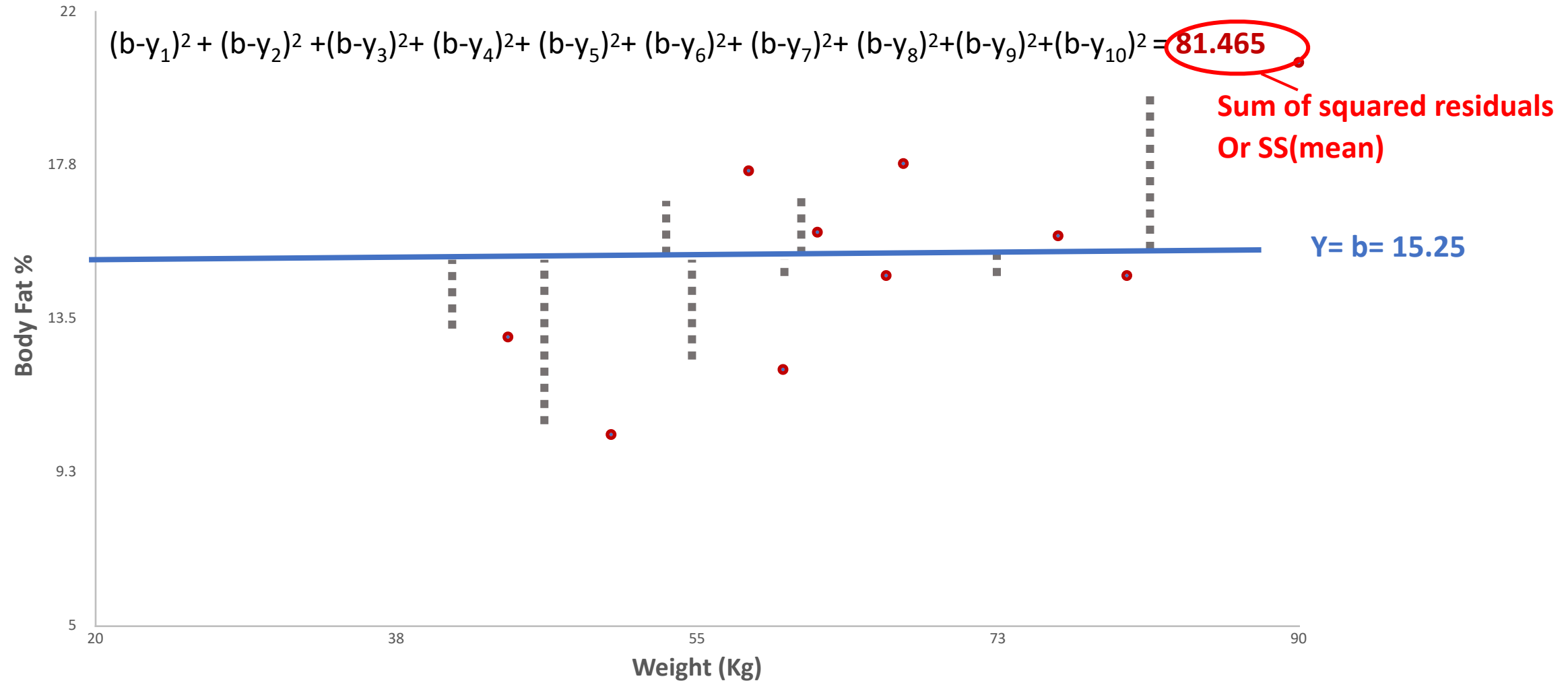
Predict Body Fat%

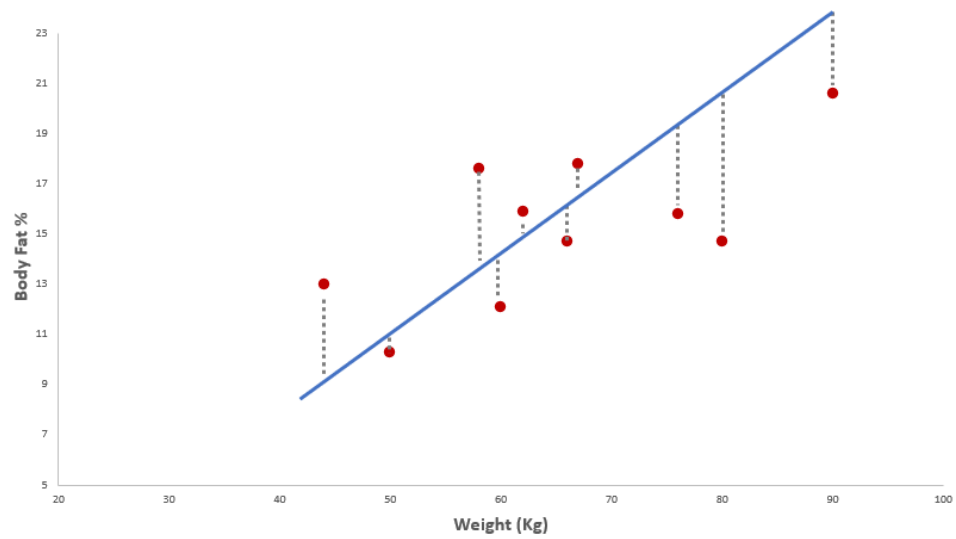
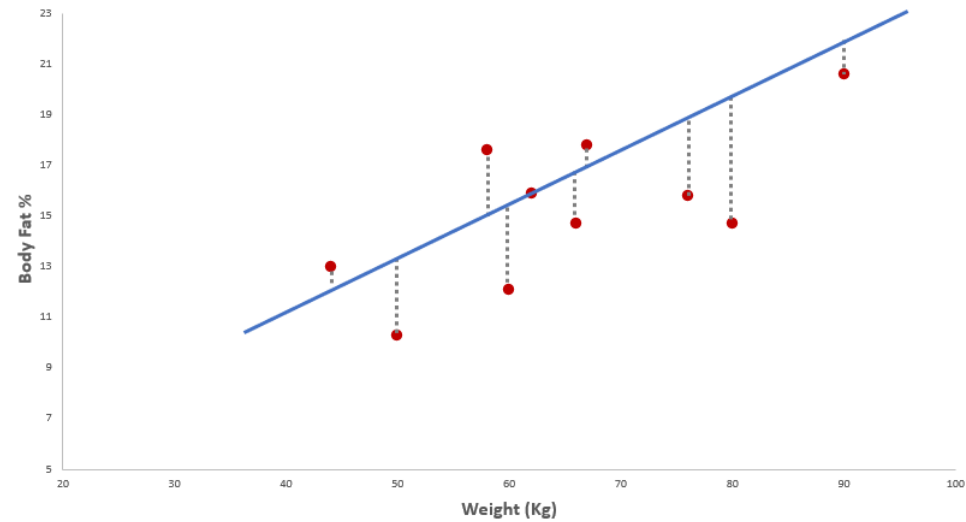
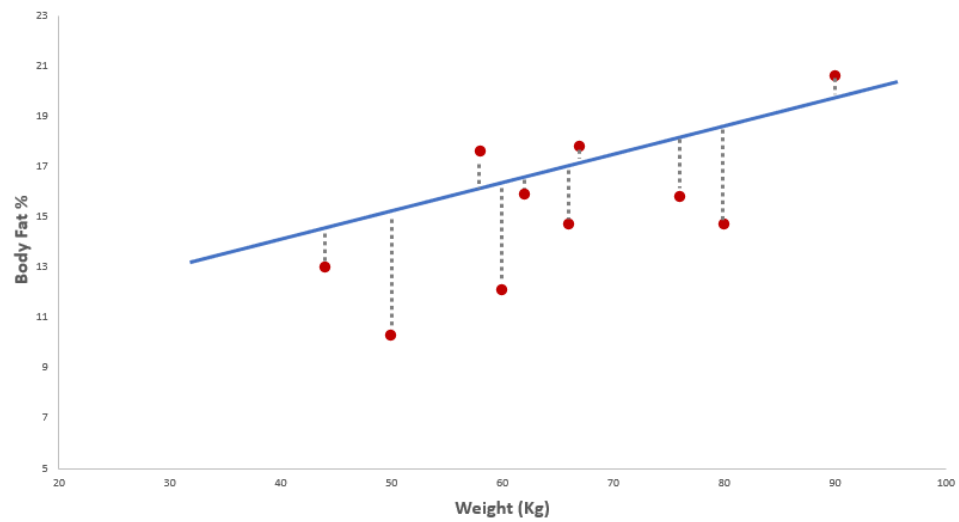


Predict Body Fat%

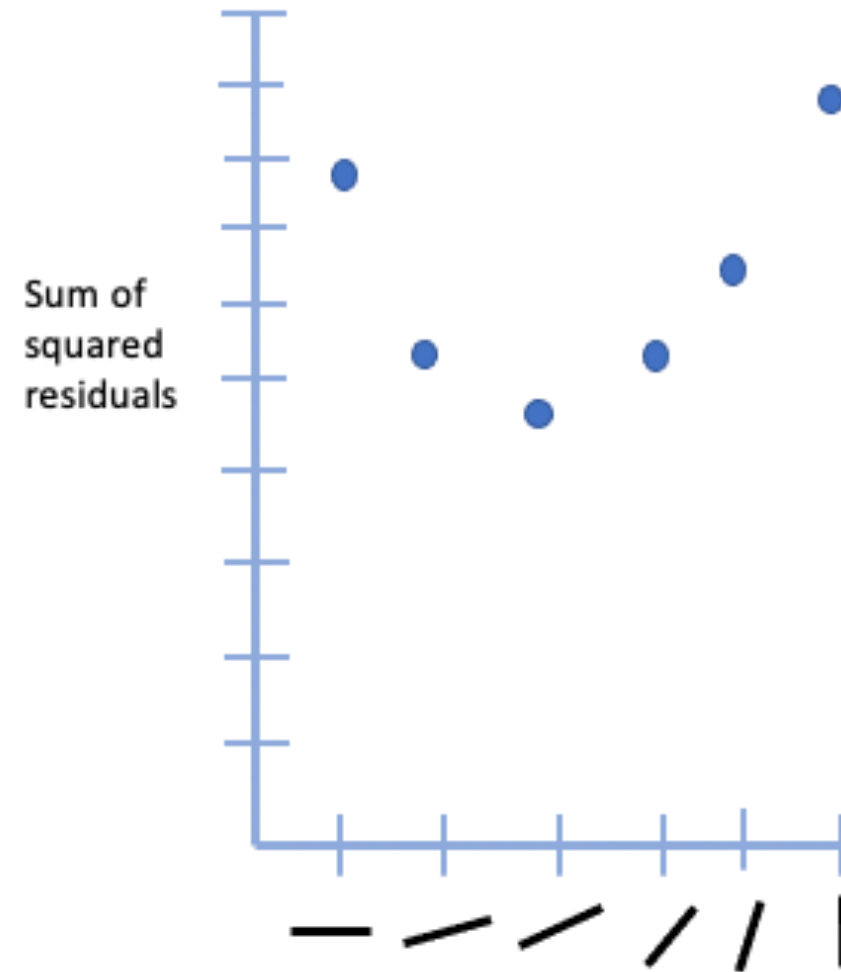


Predict Body Fat%

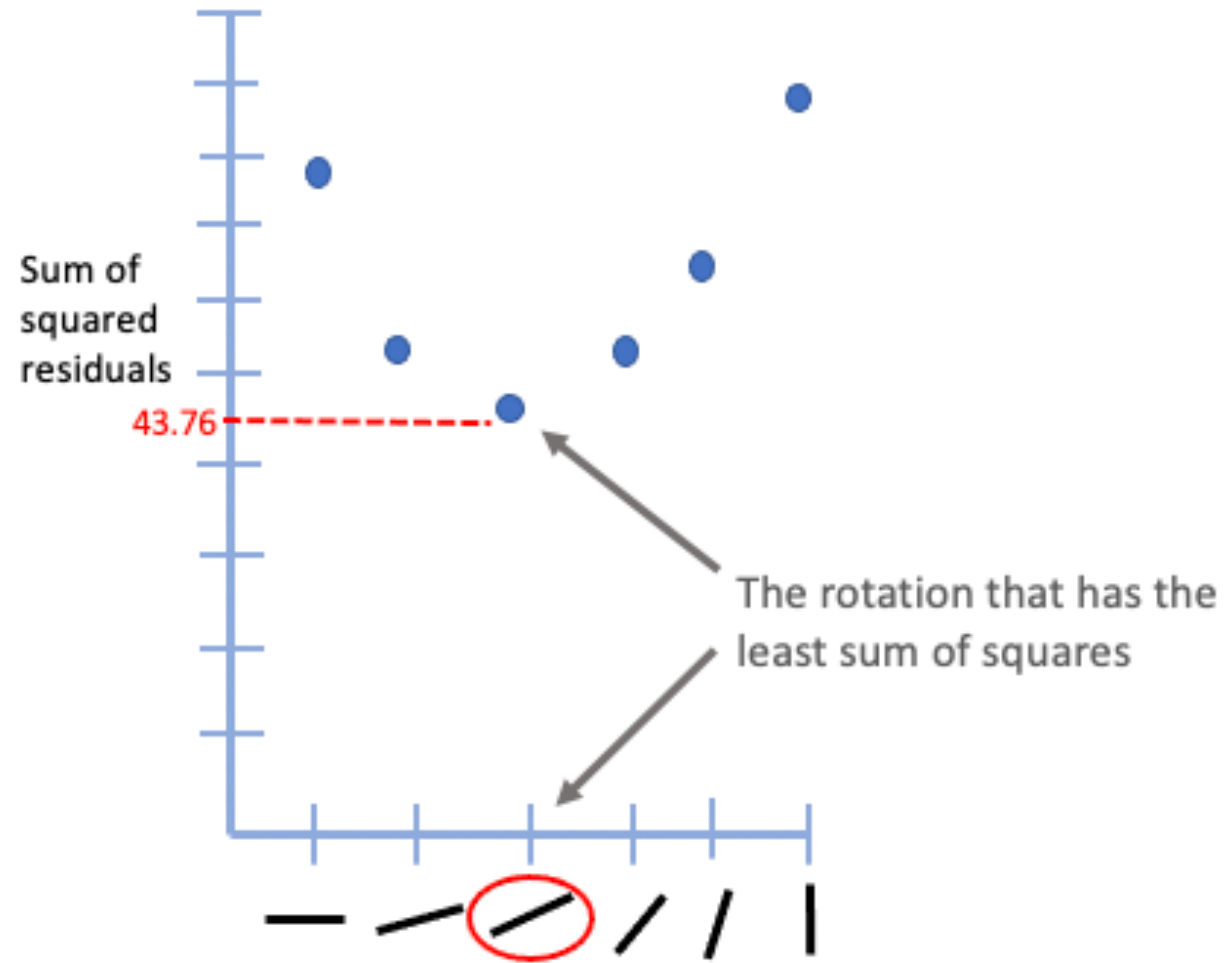


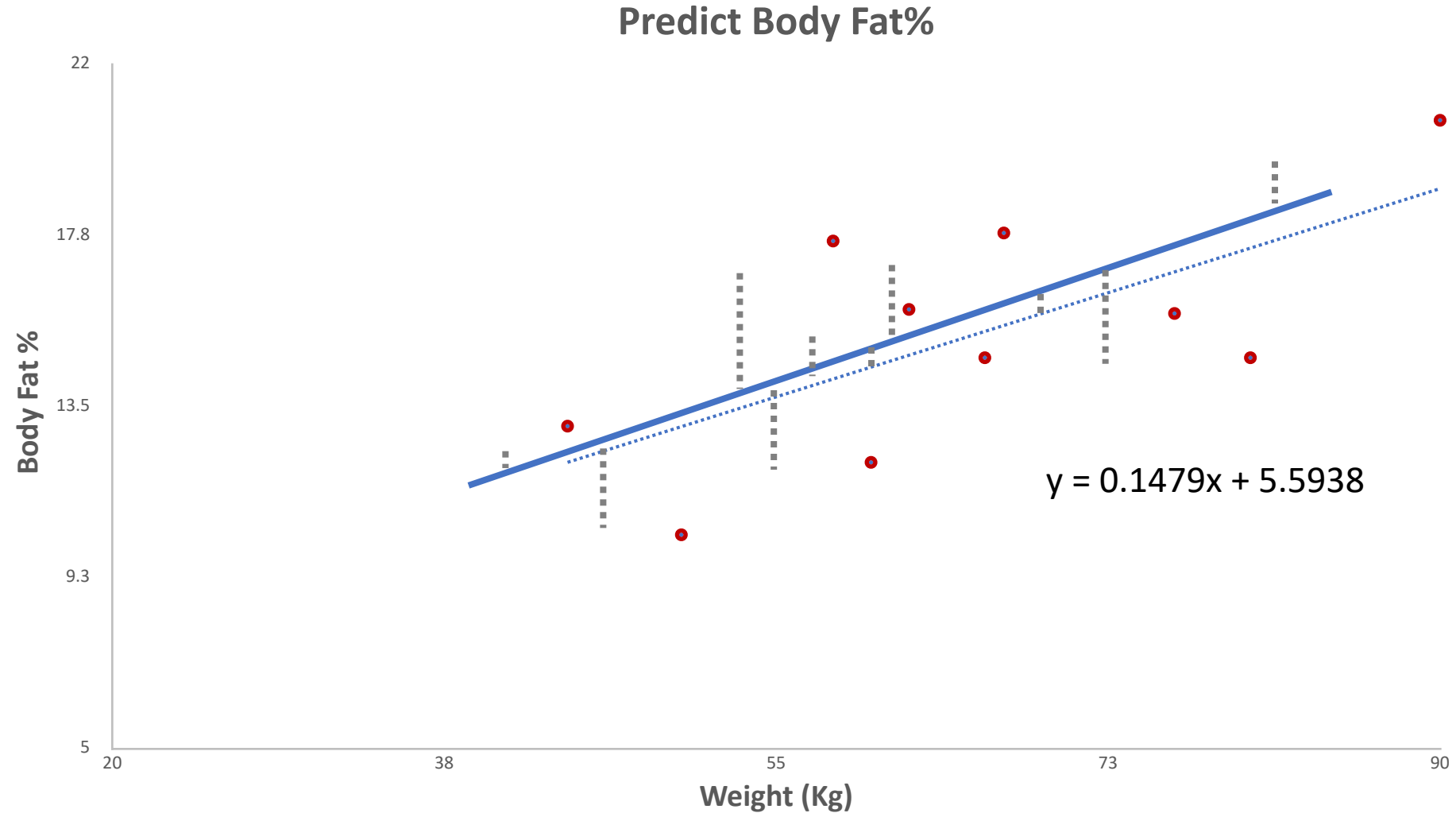


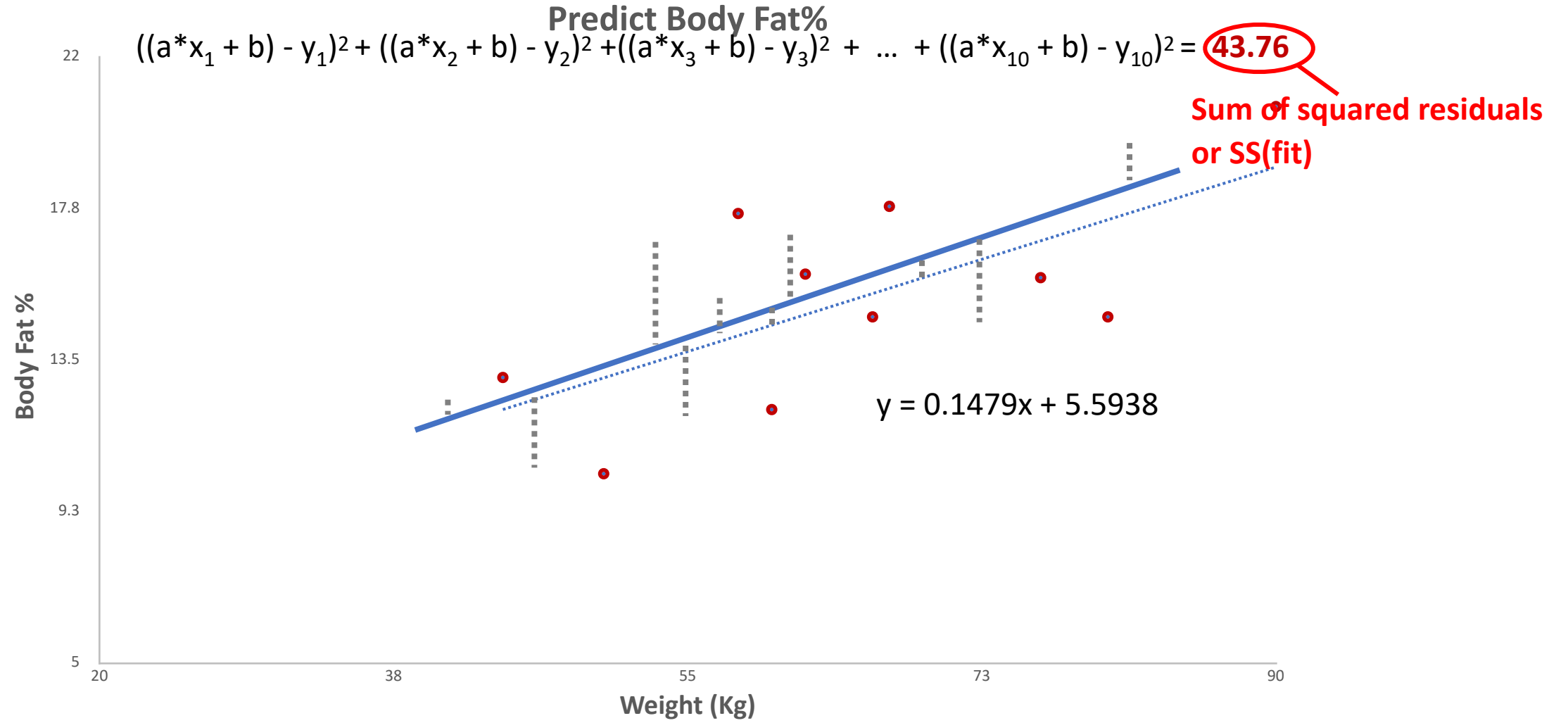
Sum of Squared Residuals and Corresponding Rotation



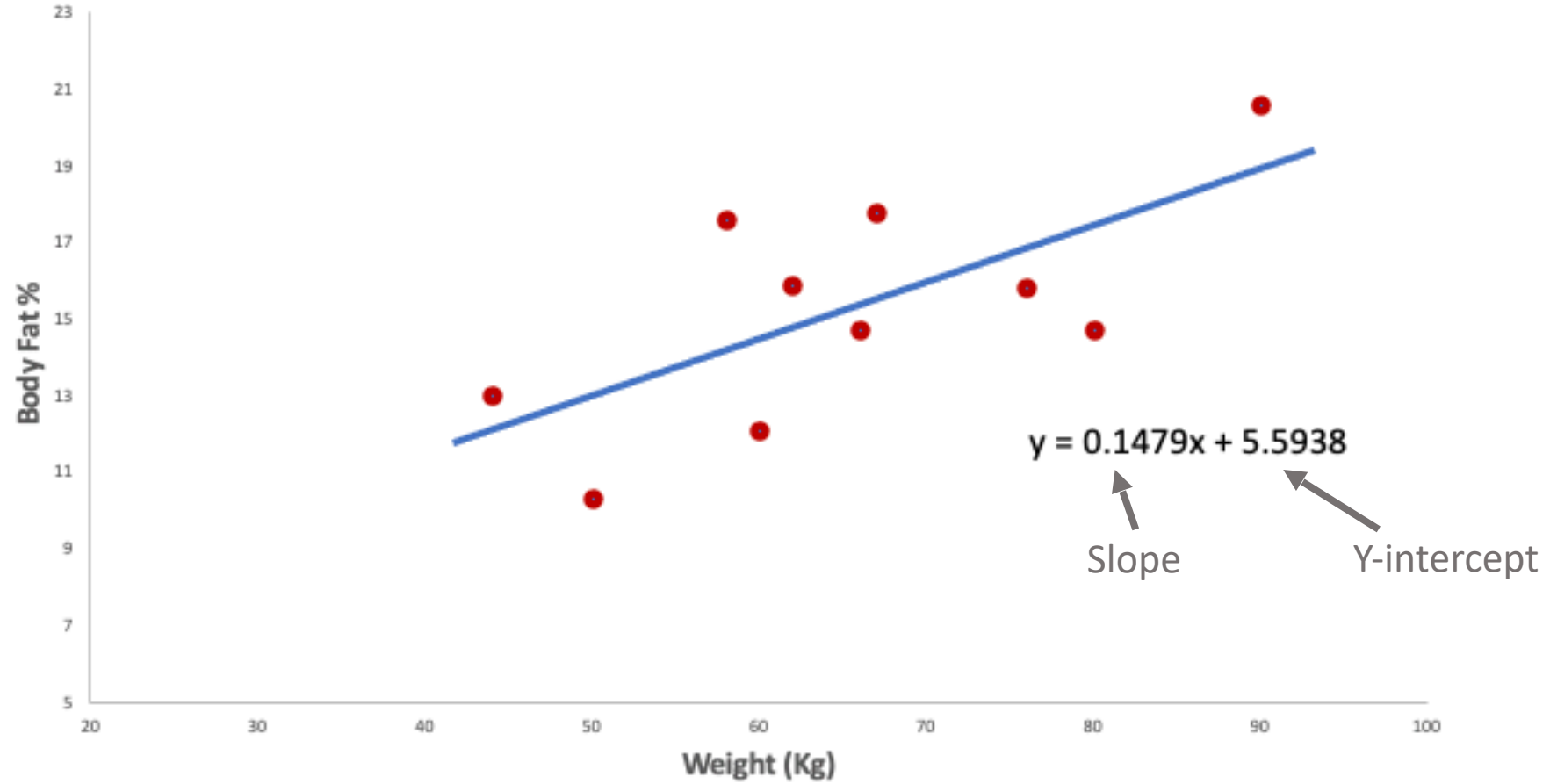
Sum of Squared Residuals and Corresponding Rotation





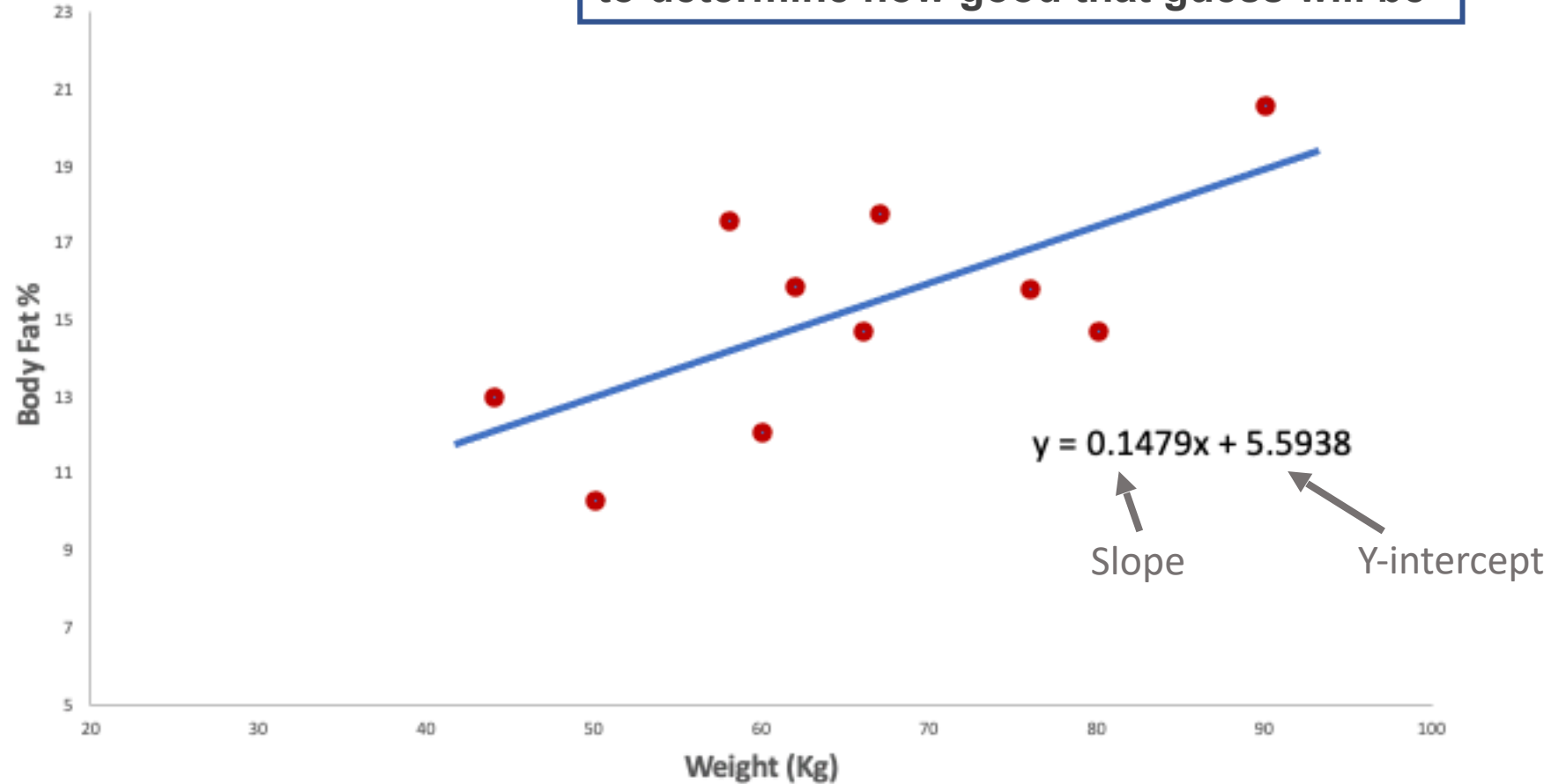


But how good is that guess?!



But how good is that guess?!

Calculating “**R-Squared**” is the first step to determine how good that guess will be



So how to
calculate
“R-
Squared”



Steps to Calculate R-Squared:

1) Calculate the variance(mean):

$$\text{Var}(\text{mean}) = \frac{\sum (\text{mean} - \text{data})^2}{n} = \frac{\text{SS}(\text{mean})}{n}$$

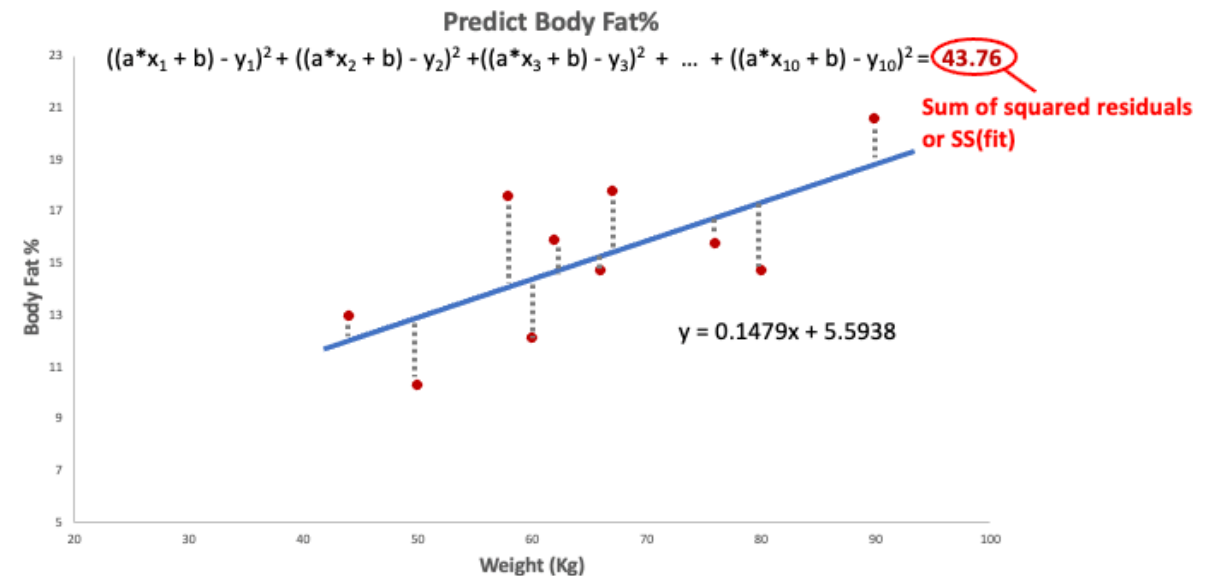
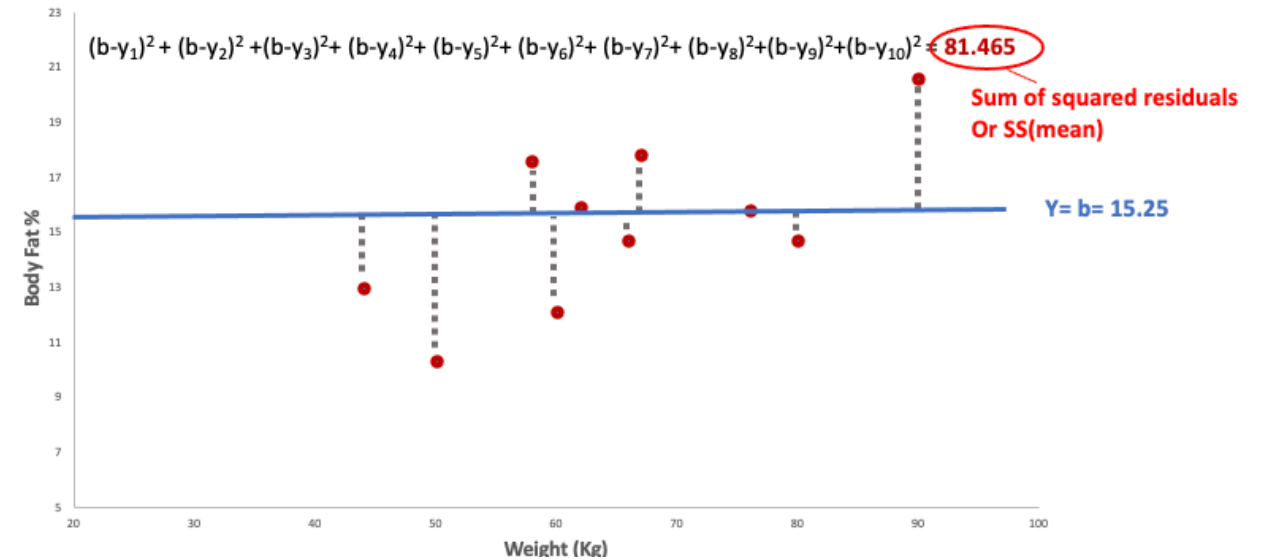
The size, (n=10 in this case)

$$= \frac{81.465}{10} = 8.1465$$

2) Calculate the variance(fit):

$$\text{Var}(\text{fit}) = \frac{\sum (\text{line} - \text{data})^2}{n} = \frac{\text{SS}(\text{fit})}{n}$$

$$= \frac{43.76}{10} = 4.376$$



Steps to Calculate R-Squared:

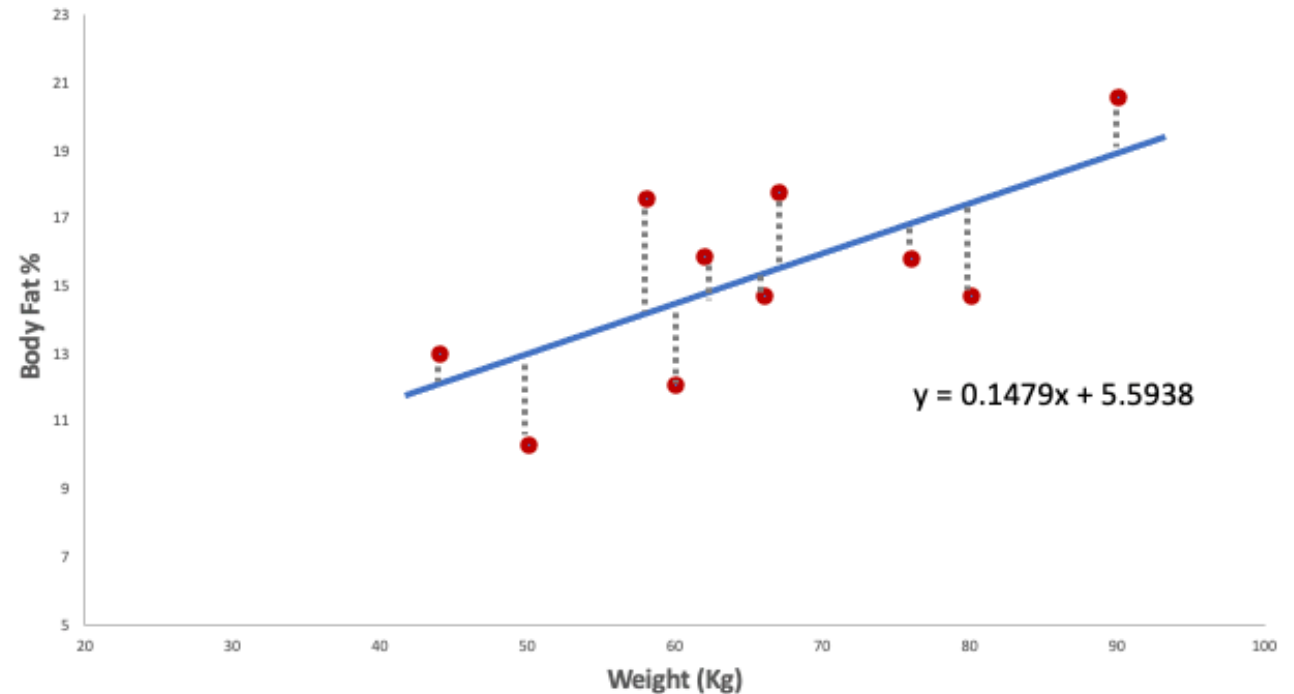
3) Calculate R-squared:

$$R^2 = \frac{Var(mean) - Var(fit)}{Var(mean)}$$

$$R^2 = \frac{8.1465 - 4.376}{8.1465}$$

$$R^2 = 0.4628 = 46.28\%$$


Thus, we can say that human weight “explains” 46.28% of the variation in human body fat%.



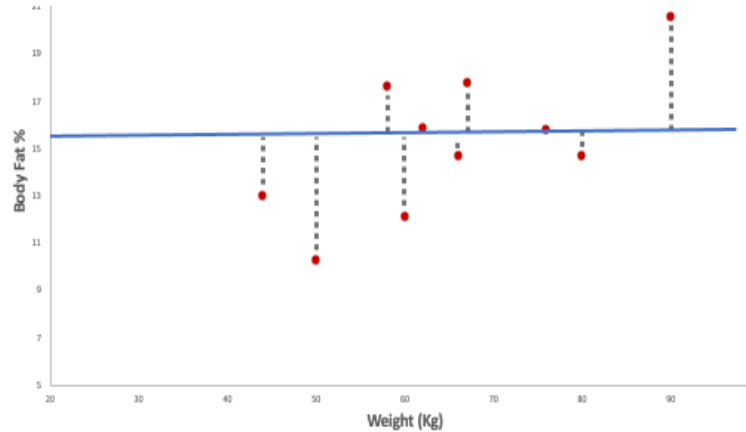
The p-value:

- The p-value helps in determining if the R-Squared value is statistically significant.
- A low p-value (< 0.05) indicates that the value we have is significant.
- In other words, a predictor that has a low p-value is likely to be a meaningful addition to the model because changes in the predictor's value are related to changes in the response variable.
- Conversely, a larger (insignificant) p-value suggests that changes in the predictor are not associated with changes in the response.

$$F = \frac{SS(\text{mean}) - SS(\text{fit}) / (P_{\text{fit}} - P_{\text{mean}})}{SS(\text{fit}) / (n - P_{\text{fit}})}$$


Degrees of Freedom

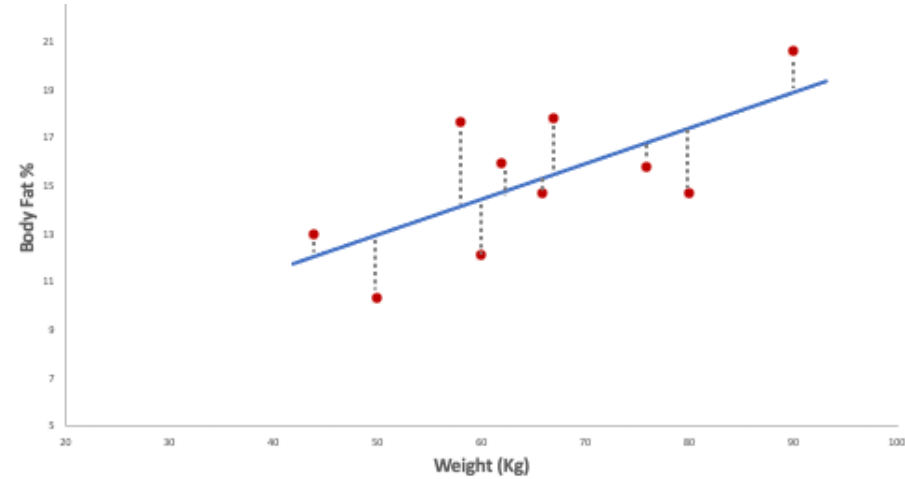
The P-value:



$y = y\text{-intercept}$

1 parameter

$$P_{\text{mean}} = 1$$



$y = y\text{-intercept} + \text{slope } x$

2 parameters

$$P_{\text{fit}} = 2$$

$$F = \frac{SS(\text{mean}) - SS(\text{fit}) / (P_{\text{fit}} - P_{\text{mean}})}{SS(\text{fit}) / (n - P_{\text{fit}})}$$

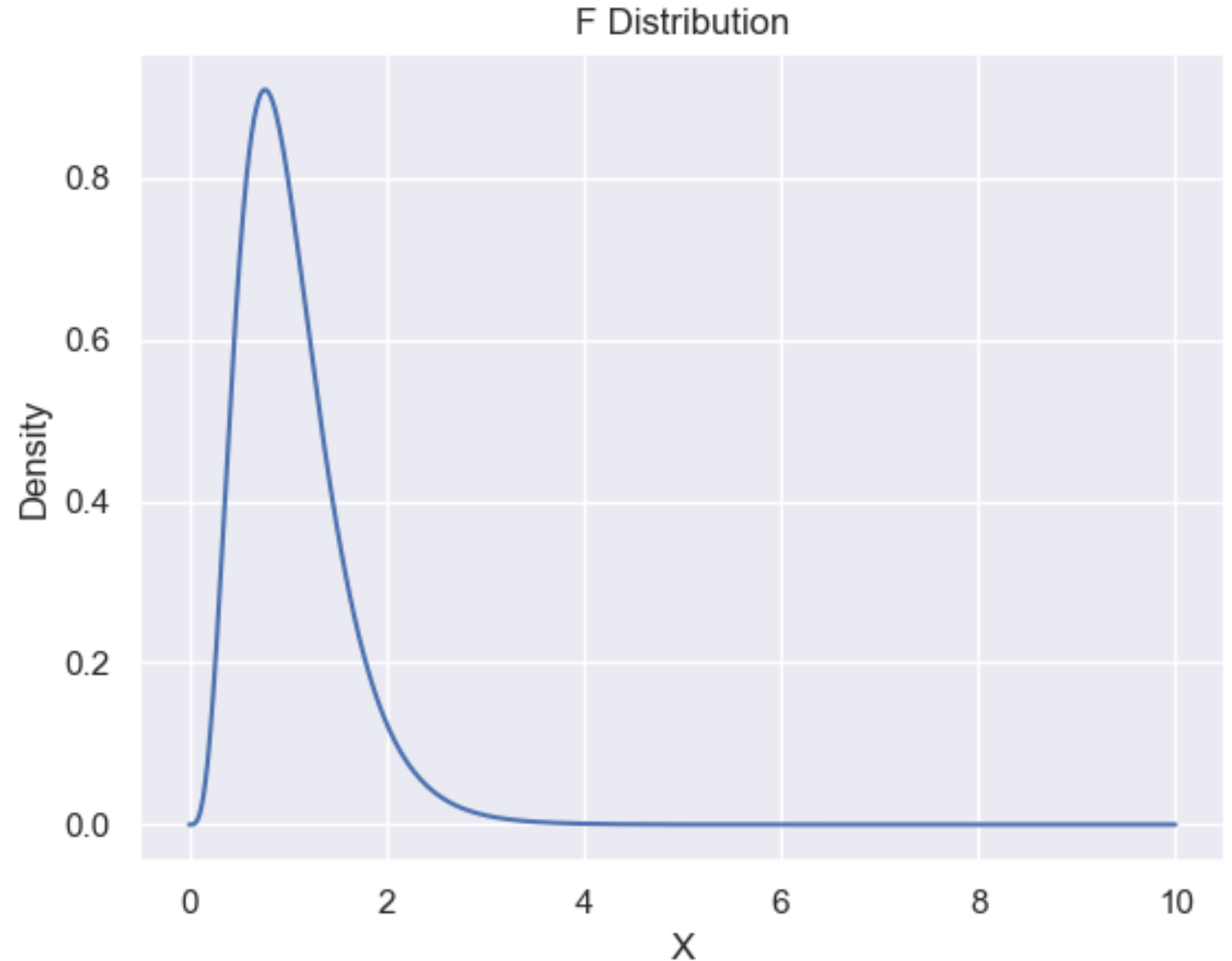
The F-Statistic:

The F value for the best fitted line:

$$F = \frac{SS(\text{mean}) - SS(\text{fit}) / (P_{\text{fit}} - P_{\text{mean}})}{SS(\text{fit}) / (n - P_{\text{fit}})}$$

$$= \frac{(81.465) - (43.76)}{(2 - 1)} \\ = \frac{(43.76)/(10 - 2)}{}$$

$$= 6.89$$



The p-value:

```
from scipy.stats import f
numerator = (81.465 - 43.76) / (2 - 1)
denominator = 43.76 / (10 - 2)
f_stat = numerator / denominator
print('F-statistic = ', f_stat)
p_value = 1 - f.cdf(f_stat, 1, 8)
print('p_value = ', p_value)
```

```
F-statistic = 6.893053016453384
p_value = 0.030391556165255018
```

- H_0 = coefficients are 0 and only the y-intercept is statistically significant.
- The p-value is less than our Significance Level (0.05) and we can reject the null hypothesis.

The F value for the best fitted plane (using weight and height features):

$$F = \frac{SS(\text{mean}) - SS(\text{fit}) / (P_{\text{fit}} - p_{\text{mean}})}{SS(\text{fit}) / (n - p_{\text{fit}})}$$
$$= \frac{\frac{(81.465) - (25.741)}{(3 - 1)}}{(25.741) / (10 - 3)}$$
$$= 7.576$$

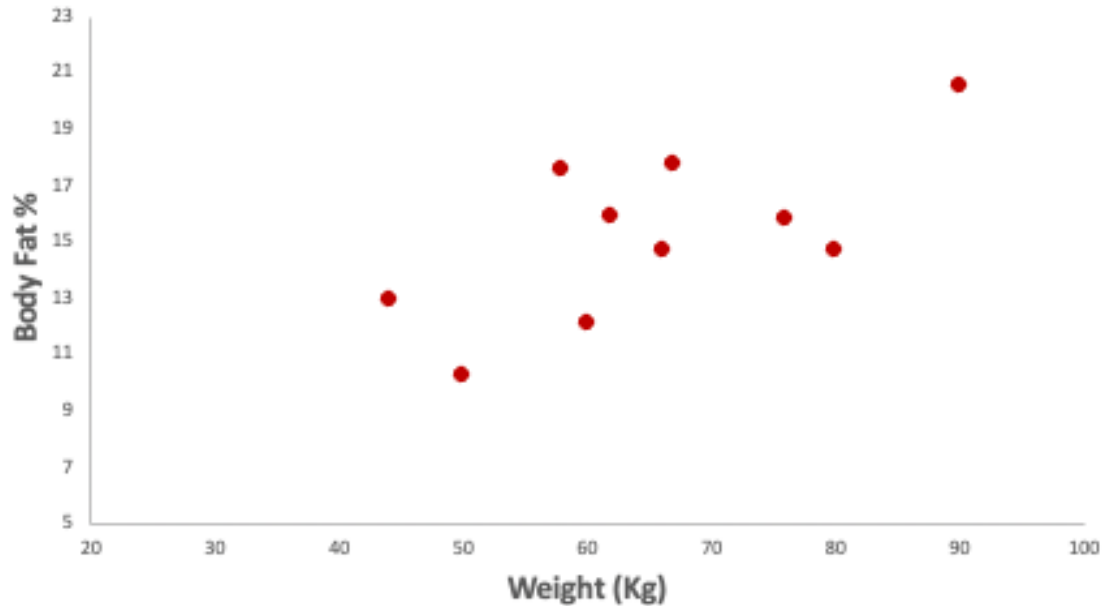
```
from scipy.stats import f
numerator = (81.465 - 25.741) / (3 - 1)
denominator = 25.741 / (10 - 3)
f_stat = numerator / denominator
print('F-statistic = ', f_stat)
p_value = 1 - f.cdf(f_stat, 2, 7)
print('p_value = ', p_value)
```

```
F-statistic = 7.576784118721108
p_value = 0.01773332749628509
```

Adding a 2nd feature improved the performance of our model

In conclusion:

When given some data that we think are related...



Linear regression:

1) Find R^2

- Quantifies the relationship in the data
- Needs to be large

2) Find the p-value (Calculated with F)

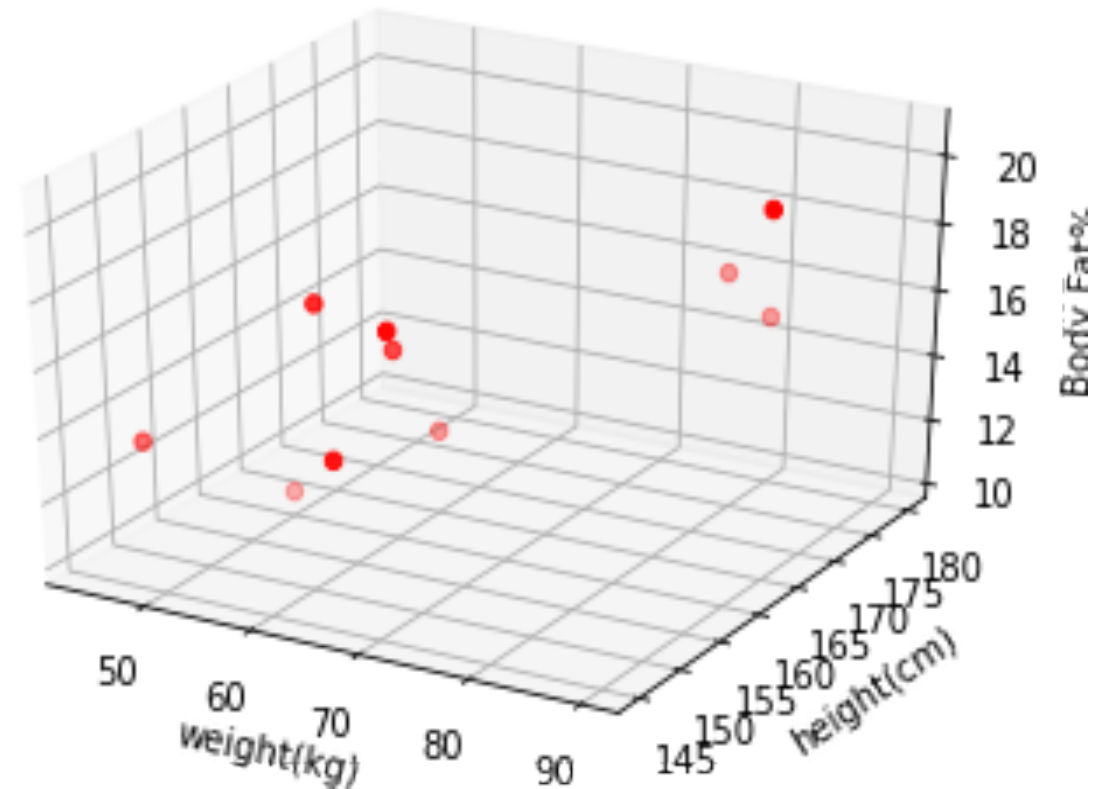
- Determines how reliable this relationship is
- Needs to be small

We need both to have an interesting result!!!

Multiple Linear regression (MLR)

- A statistical technique that uses two or more independent variables to predict an output variable, which is an extension of linear regression that uses just one independent variable.

weight(kg)	height(cm)	Body Fat %
44	150	13
50	160	10.3
58	153	17.6
60	165	12.1
62	157	15.9
66	145	14.7
67	150	17.8
76	180	15.8
80	180	14.7
90	165	20.6



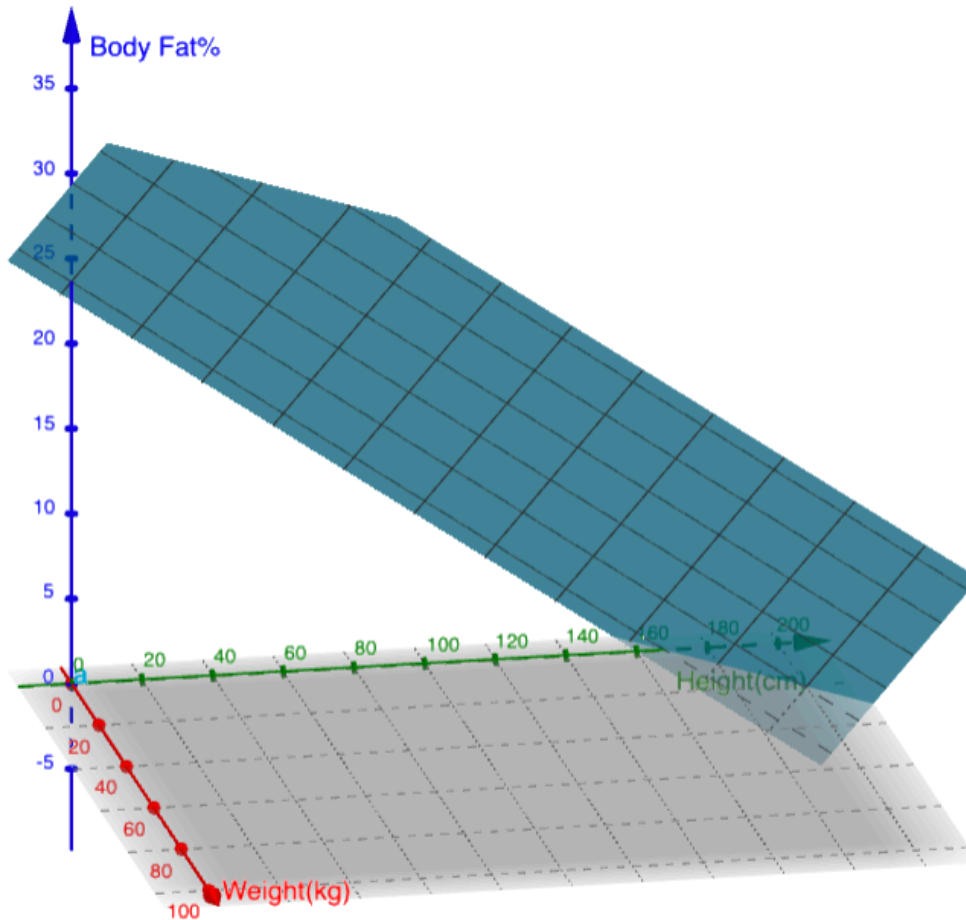
$$\text{Regression Equation} = \hat{y} = b_1X_1 + b_2X_2 + a$$

$$y = 0.21738X_1 - 0.14074X_2 + 23.64325$$

$$\text{Sum}(\text{fit}) = \text{least squared} = 25.741$$

$$R^2 = \frac{SS(\text{mean}) - SS(\text{fit})}{SS(\text{mean})} = \frac{81.465 - 25.741}{81.465}$$

$$R^2 = 0.6840 = 68.4\%$$



We can say that human weight and height “explain” 68.4% of the variation in human body fat%.

Therefore, R^2 value increases when adding more significant features (independent Variables), thus better prediction Result!

Recap: Multiple Linear Regression

Simple
Linear
Regression

$$y = b_0 + b_1 * x_1$$

Multiple
Linear
Regression

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \dots + b_n * x_n$$

- Multiple regression is an extension of simple linear regression
- Two or more independent variables are used to predict/ explain the variance in one dependent variable
- In multiple regression, each coefficient is interpreted as the estimated change in y corresponding to one unit change in a variable, when all other variables are held constant.

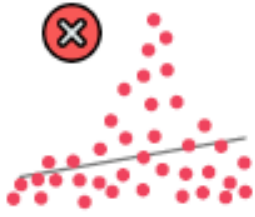
Linear/multiple regression continued

Assumptions of Linear Regression



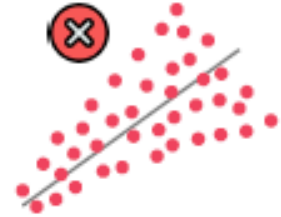
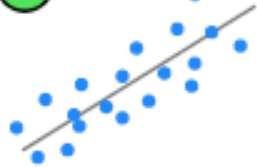
1. Linearity

(Linear relationship between Y and each X)



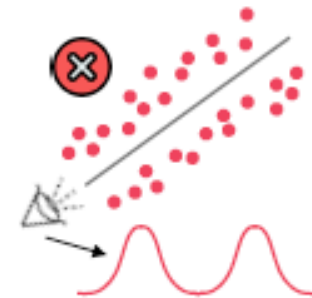
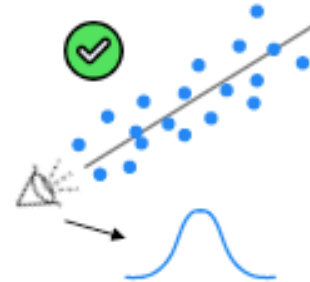
2. Homoscedasticity

(Equal variance)



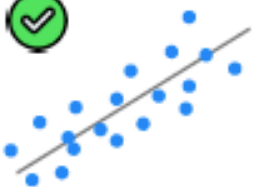
3. Multivariate Normality

(Normality of error distribution)



4. Independence

(of observations. Includes "no autocorrelation")



5. Lack of Multicollinearity

(Predictors are not correlated with each other)



$X_1 \not\sim X_2$

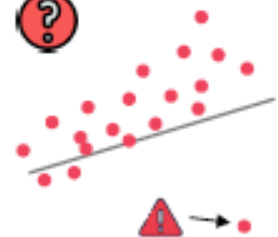
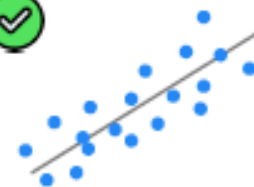


$X_1 \sim X_2$



6. The Outlier Check

(This is not an assumption, but an "extra")



Multiple Linear Regression: things to consider

Two problems may arise:

- Adding more independent variables to a multiple regression procedure doesn't mean the regression will be "better" or offer better predictions; in fact it can make things worse. This is called **Overfitting**.
- The addition of more independent variables create more relationships among them. So not only are the independent variables potentially related to the dependent variable, they are also potentially related to each other. When this happens, it is called **Multicollinearity**.
- The ideal is for all of the independent variables to be correlated with the dependent variable but **Not** with each other.

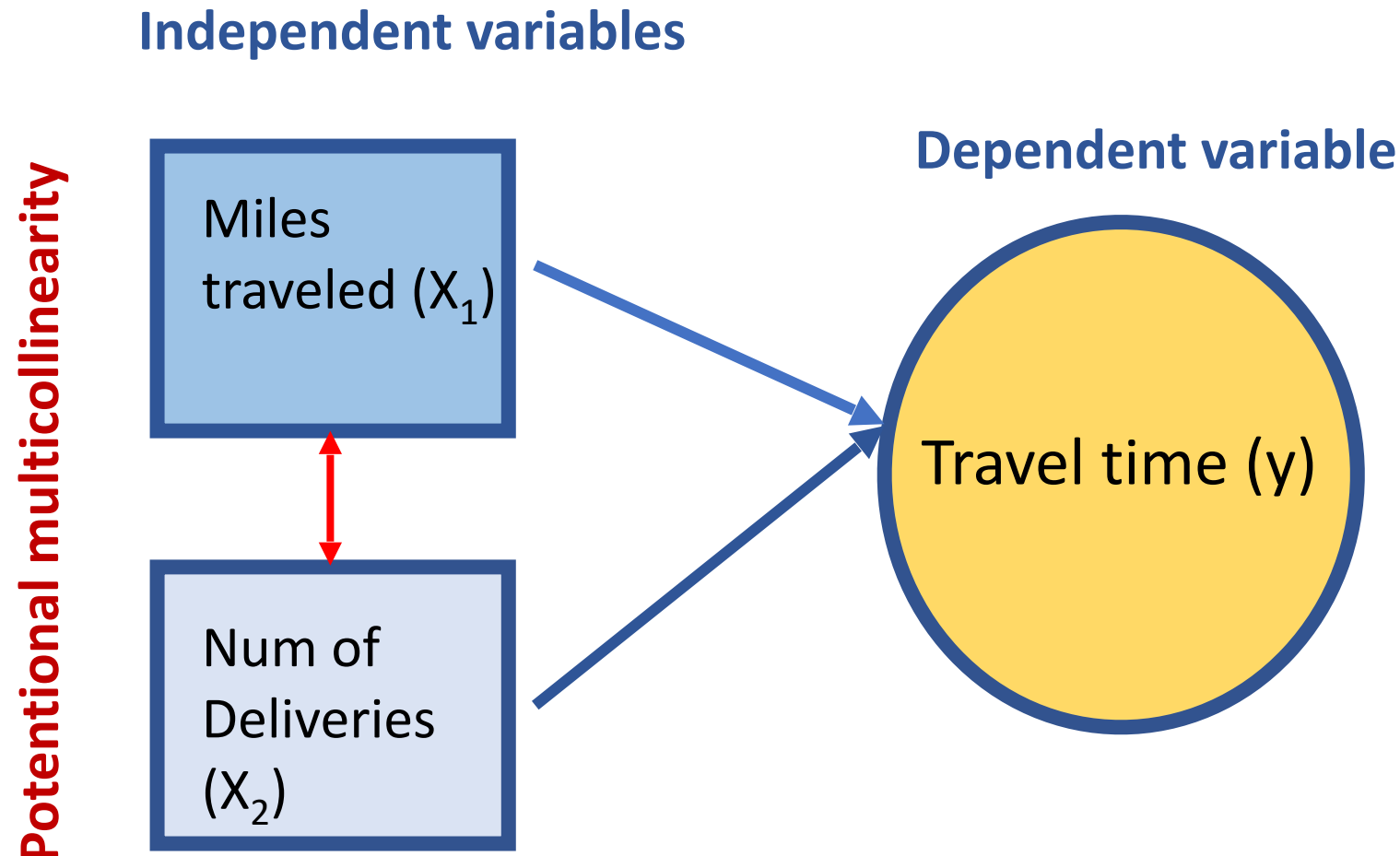
Multiple Linear Regression: things to consider

Because of **multicollinearity** and overfitting, there is a fair amount of prep-work to do before conducting multiple regression analysis if one is to do it properly:

- Correlations
- Scatter plots
- Simple regressions.

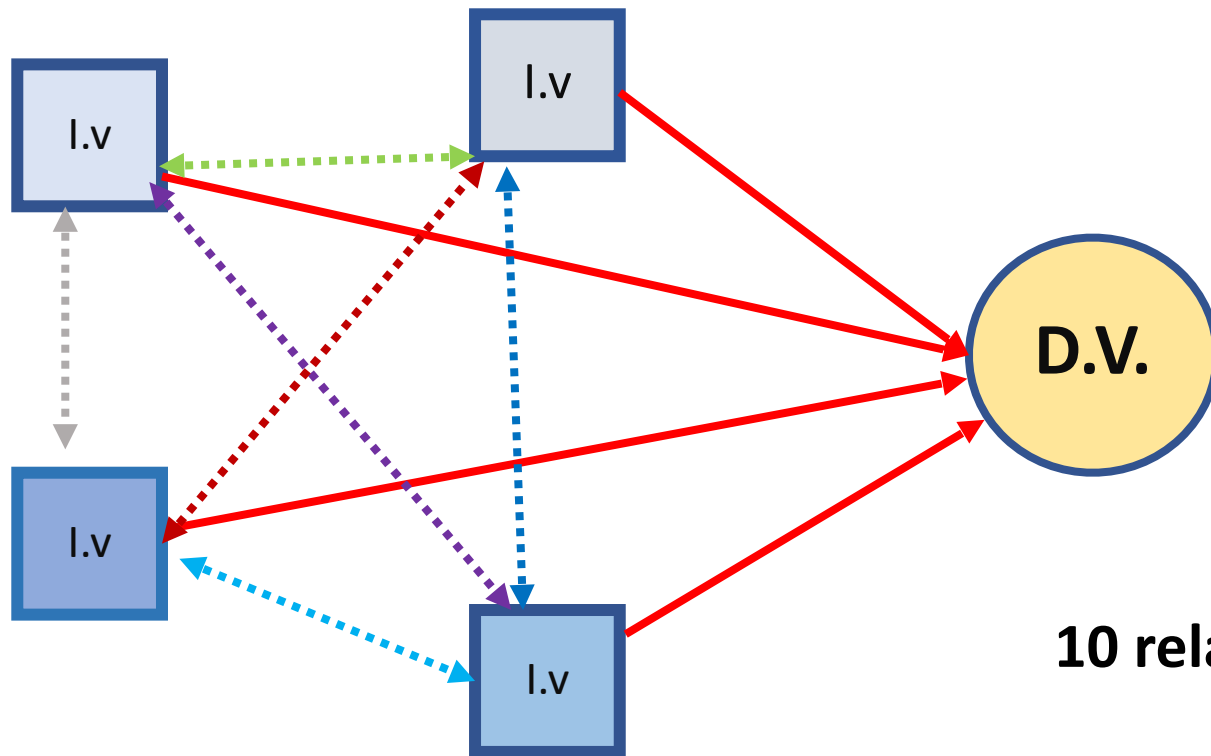
Multiple Linear Regression: things to consider

- Example: many-to-one relationship



Multiple Linear Regression: things to consider

- Many to one relationships:



10 relationships to consider!

Multiple Linear Regression: things to consider

Conclusion:

- Some independent variables, or sets of independent variables, are better at predicting the dependent variable than others. some contribute nothing.
- The ideal is for all of the independent variables to be correlated with the dependent variable but **Not** with each other.

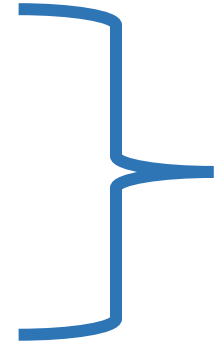
Building a model: Multiple Linear Regression

which features (independent variables) should we consider?



5 methods of building models:

1. All-in
2. Backward Elimination
3. Forward Selection
4. Bidirectional Elimination
5. Score Comparison



Stepwise Regression

Building A Model:

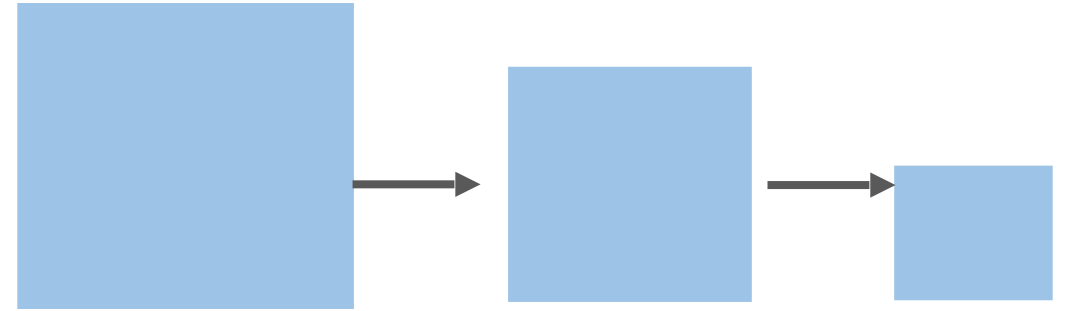
1) All-in cases:

- Prior Knowledge
- Or, preparing for Backward Elimination

Including all the
independent variables

Building A Model

2) Backward Elimination



Step 1: select a significance level to stay in the model (e.g. $SL = 0.05$)

Step 2: fit the full model with all possible predictors

Step 3: consider the predictor with the highest P-value. If $P > SL$, continue on to Step 4. Otherwise your model is Ready.

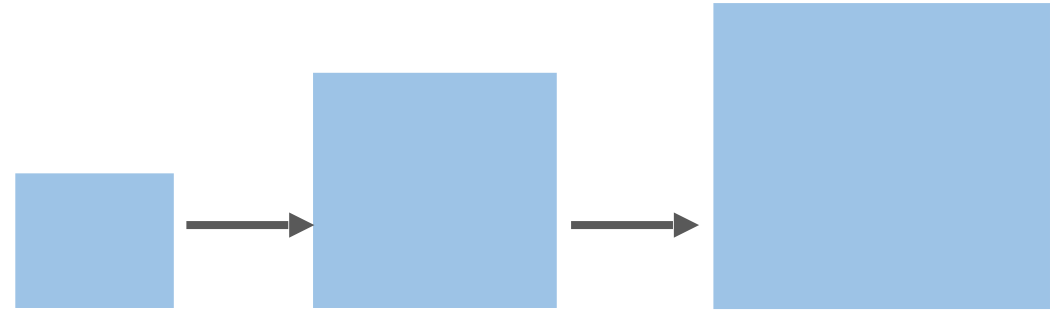
Step 4: Remove the predictor

Step 5: fit model with this variable




Building A Model

3) Forward Selection



Step 1: Select a significance level to enter the model (e.g. $SL = 0.05$)

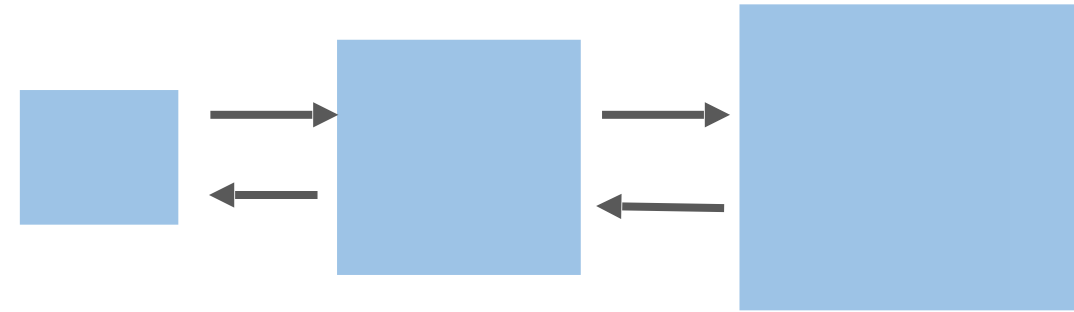
Step 2: Fit all simple regression models $y \sim x_n$. Select the one with the lowest P-value

 Step 3: Keep this variable and fit all possible models with one extra predictor added to the one(s) you already have

Step 4: Consider the predictor with the lowest P-value. If $P < SL$, go to step 3, otherwise your model is ready.

Building A Model

4) Bidirectional Elimination



Step 1: Select a significance level to enter and to stay in the model
e.g.: SLENTER = 0.045, SLSTAY = 0.05

Step 2: perform the next step of Forward Selection (new variables must have: $p < \text{SLENTER}$ to enter)

Step 3: Perform All steps of Backward Elimination (old variables must have $p < \text{SLSTAY}$ to stay)

Step 4: no new variables can enter and no old variables can exit

Building A Model

5) All possible Models

Step 1: Select a criterion of goodness of fit (e.g., R-squared)

Step 2: construct all possible regression models: $2^N - 1$ total combination

Step 3: select the one with the best criterion

**For example, 1,023 models
Will be generated if you have
10 columns!!**