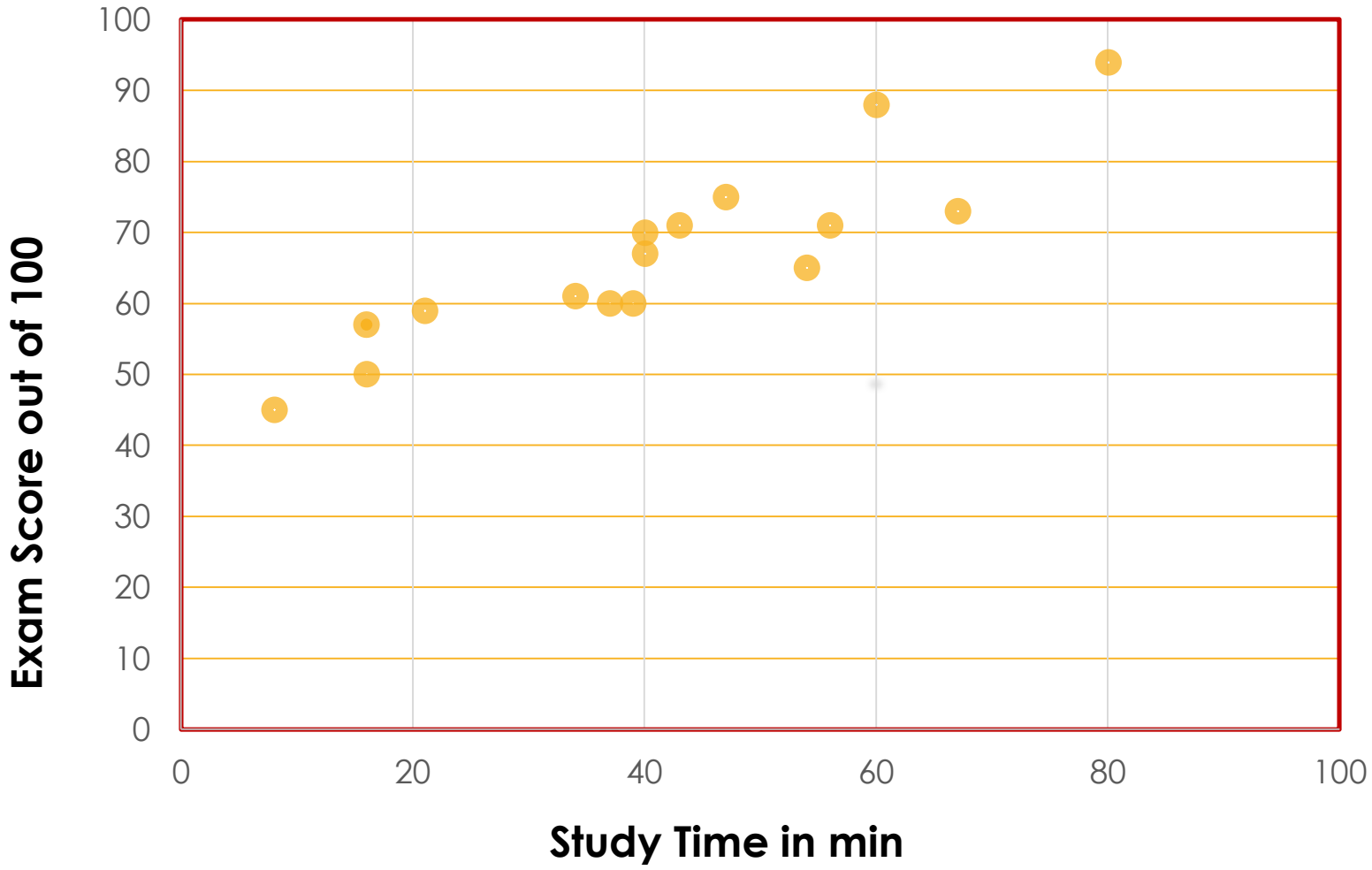
LINEAR REGRESSION

AN EXAMPLE

Does study hours have influence on exam score?

Study time (min)	Score out of 100
16	50
34	61
8	45
38	60
39	60
40	67
54	65
21	59
16	57
67	73
40	68
43	71
47	75
56	71
60	88
80	94

SCATTERPLOT



WHAT CAN WE SAY?

From the plot it appears that:

Study hours and exam score are positively related. That means, the more time someone spends studying the more score they will get.

We want to derive a linear mathematical equation that can be used to predict exam score if study time is given.

FORMAL DEFINITION

Linear regression is used to make predictions on continuous data.

It is used to relate one dependent variable with one or more independent variables.

$y = b_0 + b_1 x$ is the equation of linear regression with one independent variable. This is known as simple linear regression.

Here b_1 is known as the coefficient of x or slope of the line and b_0 is the intercept where the line intersects the y-axis, b_0 is also known as bias.

If we have more than one independent variable then we will have more

WHAT'S THE EQUATION?

Let y represent exam score and let x represent study time.

Computing the linear regression will give us

Intercept = 42.944, slope = 0.575, R-squared = 0.8069.

How do we put them together?

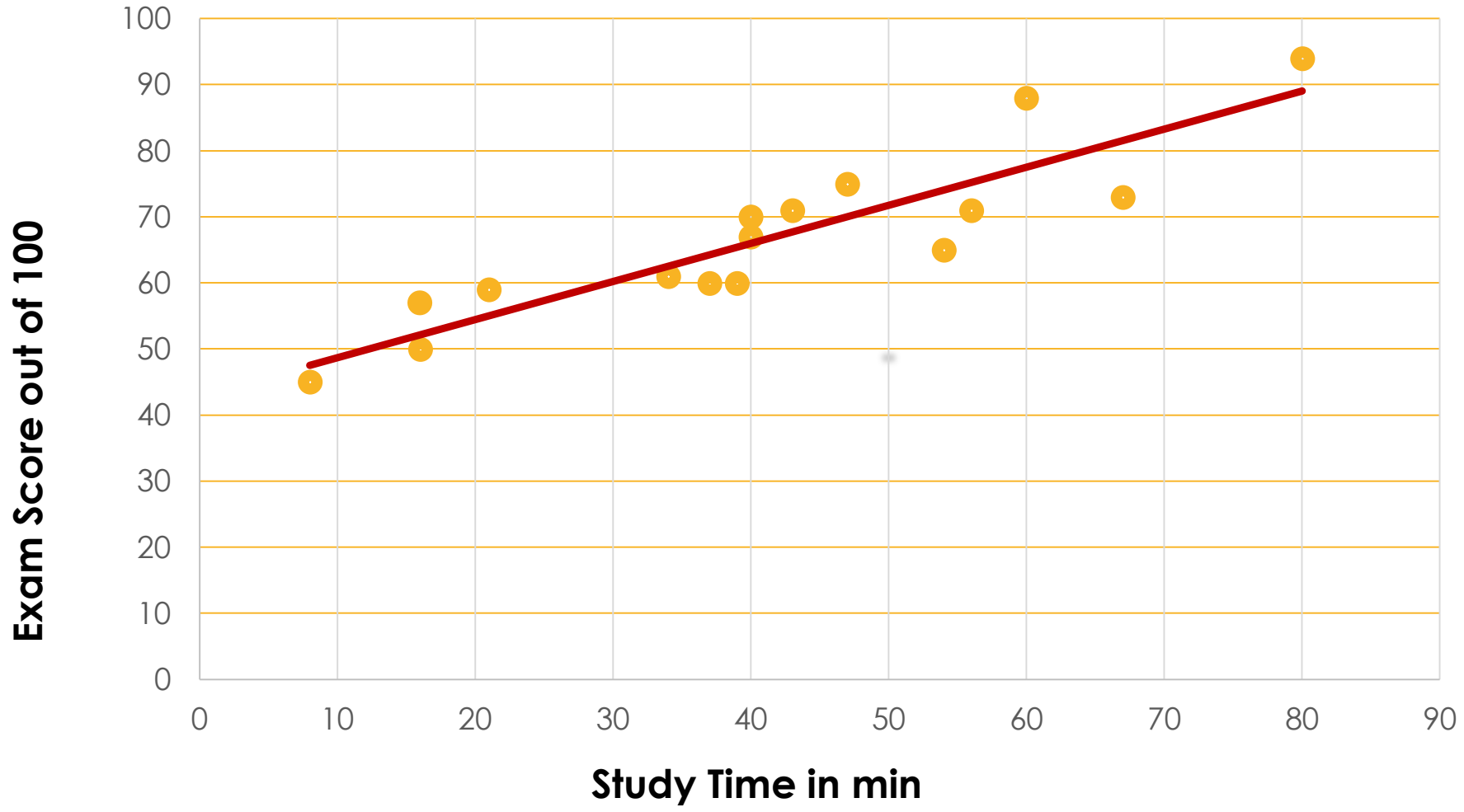
$$y = intercept + slope \ * x$$

$$y = 42.944 + 0.575\ x$$

So, to predict the exam score for study time 75 min, substitute 75 for x in the above equation

$$y = 42.944 + 0.575 * 75 = 86.069$$

CAN WE FIT THE LINE?



WHAT IS R-SQUARED?

R-squared determines goodness of fit and values range from 0 to 1.

R-squared value closer to 1 indicate that the regression line perfectly fits the data.

Coefficient of Determination,

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \overline{y})^2}$$

$\sum(y_i - \hat{y}_i)^2$ = Sum Squared Regression Error

$\sum(y_i - \overline{y})^2$ = Sum Squared Total Error

MULTI LINEAR REGRESSION

If we had three independent variables x_1, x_2 and x_3 and one dependent variable, y then the multi linear regression will be

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$

REFERENCE

Links

https://machinelearningmastery.com/linear-regression-for-machine-learning/

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html