MDM4U Linear and Non-Linear Regression

Linear regression analyzes the relationship between two variables X and Y, and determines the best straight line through the data.

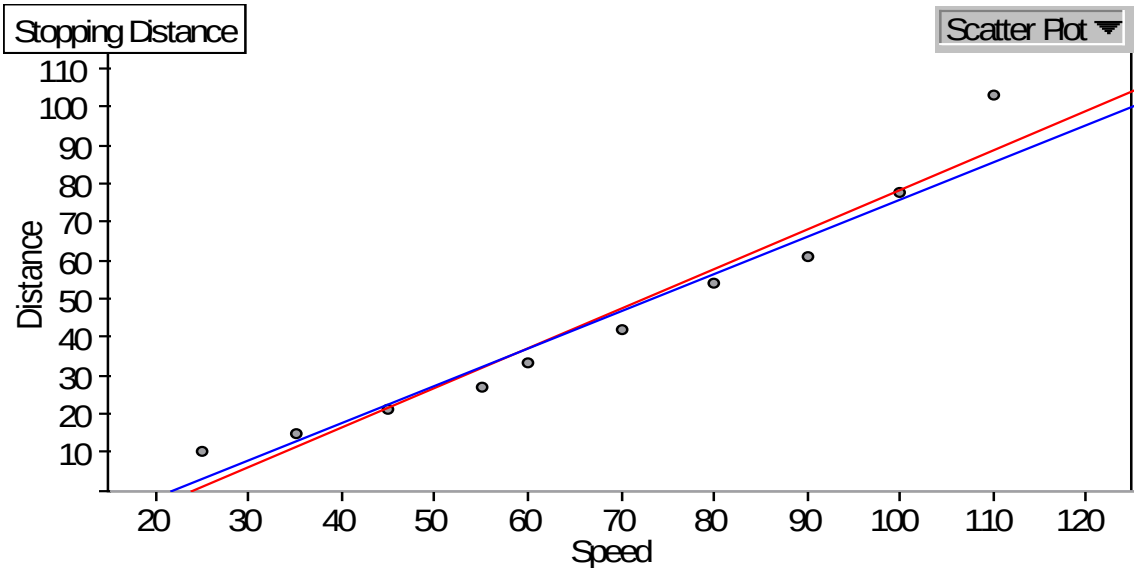
The term “regression”, like many statistical terms is used in statistics quite differently than it is used in other contexts. The method was first used to examine the relationship between the heights of fathers and sons. They were related, of course, but the slope was less than 1. Why? The height of sons regressed to the mean. The term **“regression”** is now used for many sorts of **curve fitting.**

In general, the goal of linear regression is to find the line that best predicts Y from X. Linear regression does this by finding the line that minimizes the sum of the squares of the vertical distances of the points (actual data) from the line (estimated data).

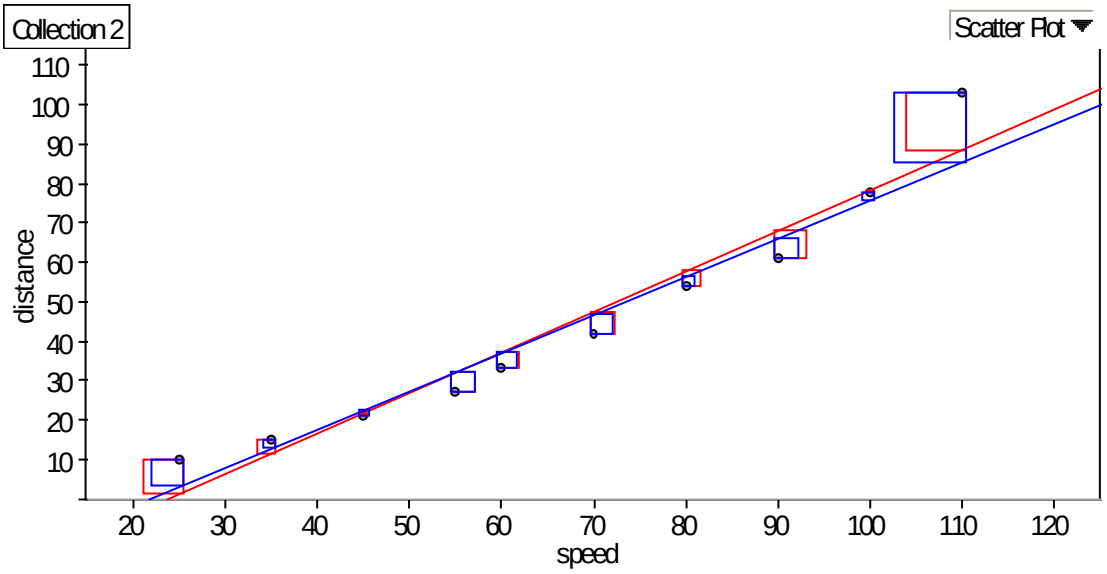
Note that linear regression assumes that data are linear, and finds the slope and intercept that make a straight line best fit the data.

The data in this scatter plot relates driving speed and stopping distance.

Speed of Car (km/h)	25	35	45	55	60	70	80	90	100	110
Stopping Distance (m)	10	15	21	27	33	42	54	61	78	103



The **least squares regression line** (red) minimizes the total distance of the points from the line and is very tedious to create by hand (so we won’t). As you can see from the graph, it is pretty close to the median-median line (blue).



For the line of best fit in the least-squares method,

1- the sum of the **residuals** is zero (the positive and negative residuals cancel out)
2- the sum of the squares of the residuals has the least possible value

Residual value – vertical distance between a point and the regression line.

Recall**: Correlation coefficient** (r) – used to measure the strength and direction of the relationship modelled by the least squares line. It is a measure of how well a regression line fits a set of data. The sign of r indicates the slope, while a number close to ± 1 indicates a strong correlation and a number close to zero indicates a weak correlation.

Coefficient of determination (r^2) – used to measure the strength of the relationship modelled by the least squares line. An r^2 value of 0.8 means that 80% of the change in the dependent variable is due to changes in the independent variable.

Predictive model – the median-median line and the least squares line are examples of **linear regression** models for the data. It is also possible to model data using other equations (quadratic, exponential).