STAB22 TUT21

Chong Chen
University of Toronto, Scarborough
Department of Computer and Mathematical Sciences

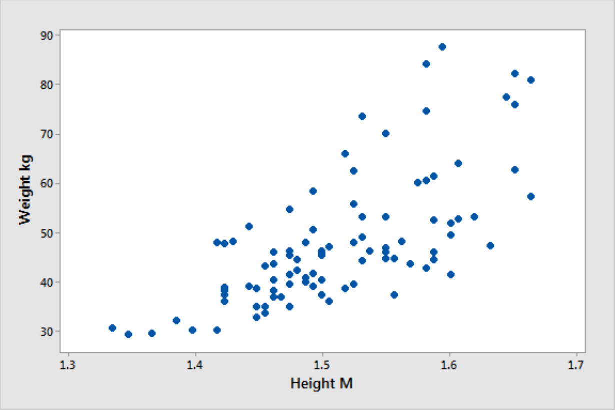
March 1, 2019

1 Understanding Correlation

A correlation between variables indicates that as one variable changes in value, the other variable tends to change in a specific direction. Understanding that relationship is useful because we can use the value of one variable to predict the value of the other variable. For example, height and weight are correlated — as height increases, weight also tends to increase.

1.1

The scatterplot below displays the height and weight of pre-teenage girls. Each dot on the graph represents an individual girl and her combination of height and weight.

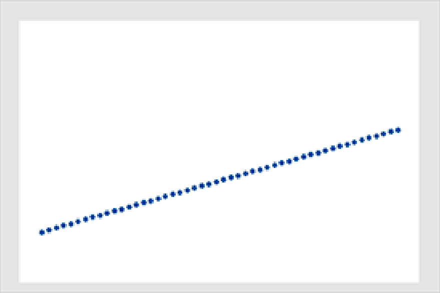


As height increases, weight also tends to increase. However, it's not a perfect relationship. If you look at a specific height, say 1.5 meters, you can see that there is a range of weights associated with it. You can also find short people who weigh more than taller people. However, the general tendency that height and weight increase together is unquestionably present.

1.2 Graphs for Different Correlations

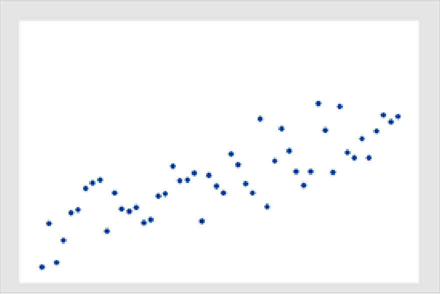
1.2.1 Correlation = $+1$

A perfect positive relationship.



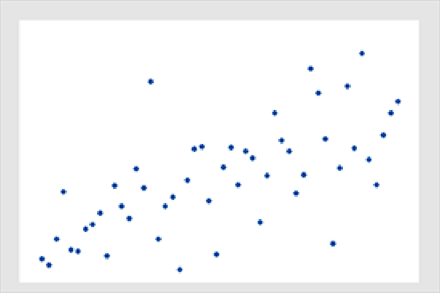
1.2.2 Correlation = 0.8

A fairly strong positive relationship.



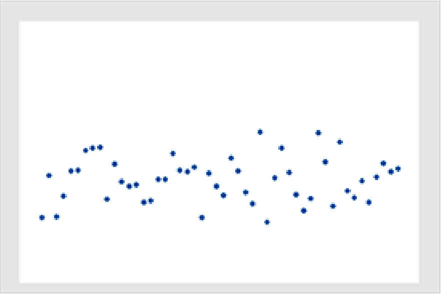
1.2.3 Correlation = 0.6

A moderate positive relationship.



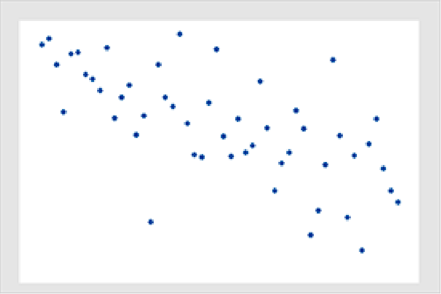
1.2.4 Correlation $= 0$

No relationship. As one value increases, there is no tendency for the other value to change in a specific direction.



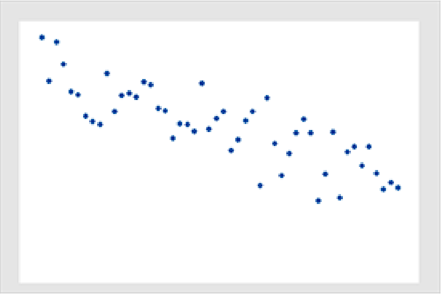
1.2.5 Correlation $= -0.6$

A moderate negative relationship.



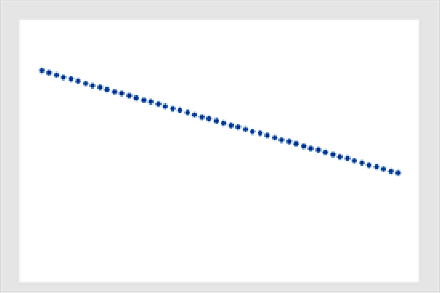
1.2.6 Correlation $= -0.8$

A fairly strong negative relationship.



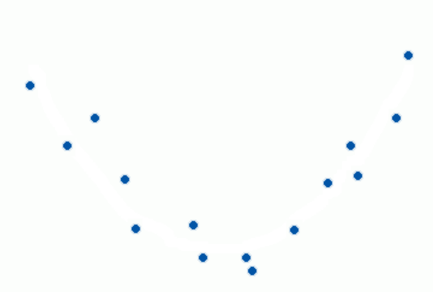
1.2.7 Correlation $= -1$

A perfect negative relationship.



1.3 Note

Correlation measures only *linear relationships*. Consequently, if your data contain a curvilinear relationship, the correlation coefficient will not detect it. For example, the correlation for the data in the scatterplot below is **zero**. However, there is a relationship between the two variables—it's just not linear.



1.4 Coefficient of determination

Coefficient of determination, r^2 is a primary measure of how well a regression model fits the data. This statistic represents the percentage of variation in one variable that other variables explain. For example, suppose the correlation r of section 1.1 is 0.694, squaring it to produce an r^2 of 0.482, or 48.2%. In other words, height explains 48.2% variability of weight in preteen girls.

2 Residuals

2.1 Recall

Last week, we talked about the linear regression, which means We can use a regression line to **predict** the dependent(response) variable y for a specific value of the independent(explanatory) variable x.

Equations:

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

where slope

$$b_1 = r \frac{s_y}{s_x}$$

and intercept

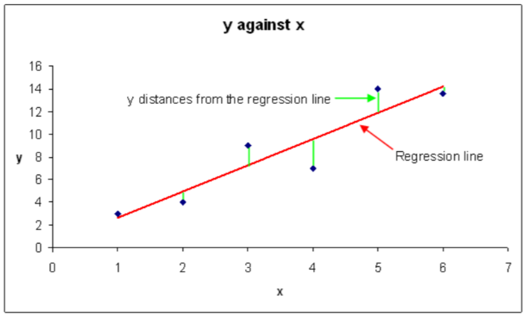
$$b_0 = \bar{y} - b_1 \bar{x}$$

2.2 Definition

A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line. That is,

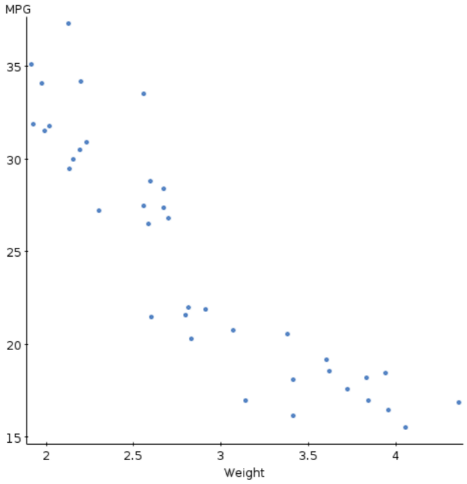
$$\mathit{Residual} = \mathit{Observed} - \mathit{Predicted} = y - \hat{y}$$

- $\mathit{Residual} > 0$: *Observed* is greater than *Predicted*, it is called **underestimate**
- $\mathit{Residual} < 0$: *Observed* is less than *Predicted*, it is called **overestimate**

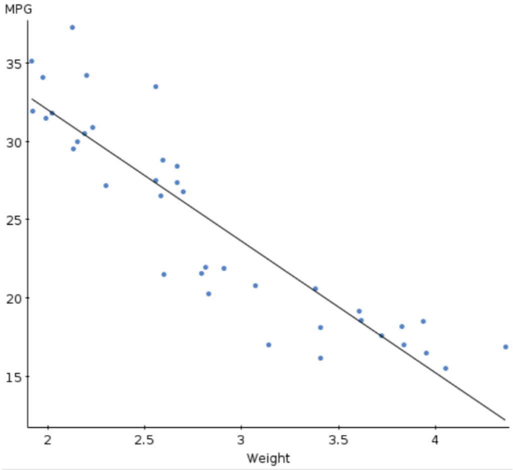


3 Residual plots

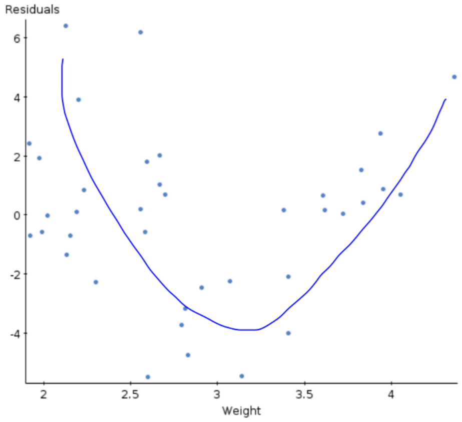
A **residual plot** is a scatterplot of the regression residuals against the explanatory variable. Residual plots help us assess the model assumptions.

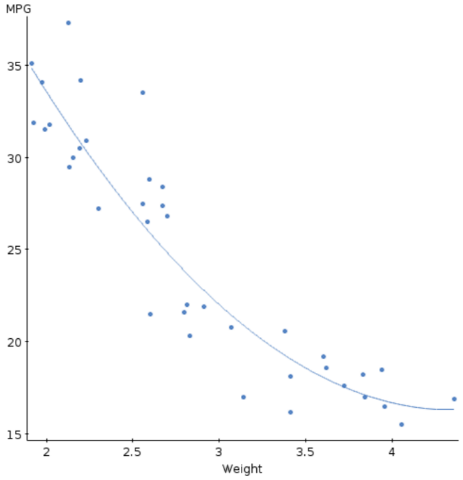


Fitted line plot



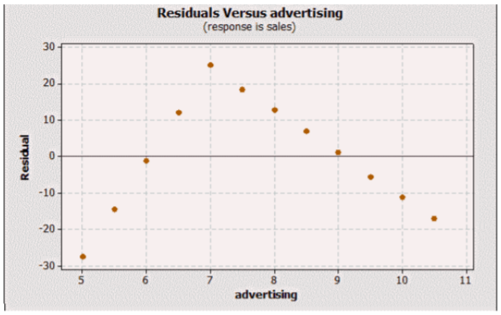
Residuals vs. Weight





4 Example

A company monitors its spending on advertising and its sales over a number of months. The regression line for predicting sales from advertising is: $sales = 260 + 23.5 \times advertising$. The plot of residuals versus advertising is given below:



(a) Calculate the predicted sales when 7 thousand dollars are spent on advertising and 450 thousand dollars are made in sales.

(b) For the observation in (a), calculate the residual, and verify that the corresponding point on the residual plot is correct.

(c) Do you think a straight line gives a reasonable description of the data? Explain briefly why or why not.