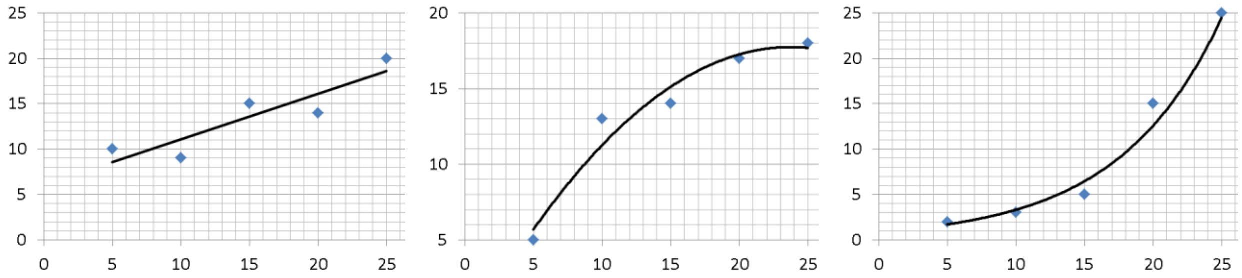
Appendix 7. Simple Regression Analysis

I. Explanation and Procedure

Simple regression analysis is a way to determine the degree to which variation in some independent variable can explain variation in a dependent variable. It is different than Pearson's and Spearman's Correlation analyses in that, rather than assuming simply a correlation between two variables, it explicitly assumes that the one variable (the dependent variable) on the y-axis is dependent upon the other variable (the independent variable) on the x-axis. Computer programs such as Microsoft Excel and SPSS employ sophisticated graphing and mathematical algorithms to do this, but the following procedure describes the fundamental approach to simple regression analysis that is easy to apply manually and to all types of linear and curvilinear or non-linear relationships that you are likely to encounter.

A. Graph your data using a scatterplot and line of best -fit

Making a scatterplot is the first step in regression analysis because this gives you a coarse understanding of the relationship between the two variables: i.e., whether or not the relationship is positive or negative, for example, and whether it is linear or some curvilinear or non-linear relationship (Fig 1). A line of best-fit (a "regression line") is then fitted to the data in the scatterplot to better illustrate this relationship (Fig 1). A line of best-fit is the relationship between x and y that is suggested by the data.



A. A positive linear relationship **B.** A positive asymptotic relationship **C.** A positive exponential relationship

Fig 1. Three of many possible relationships between two variables in biology.

B. Calculate the Coefficient of Determination, R^2

Once a best-fit line is fitted to the data, we now quantify how closely y varies with x along that line. This is done with the Coefficient of Determination (R^2, Equation 1), in which a maximum value of 1 indicates a perfect covariance and a value of 0 indicates no covariance.

$$R^2 = 1 - \frac{\Sigma\,(y_{obs} - y_{line})^2}{\Sigma\,(y_{obs} - \bar{y}_{obs})^2}$$

Equation 1, where

- **Σ** = the formulaic expression "the sum of.."
- **$(y_{obs} - y_{line})^2$** = the squared difference between one observed y value and the y value predicted by the line for a given x value.
- **$(y_{obs} - \bar{y}_{obs})^2$** = the squared difference between one observed y and the mean of all observed y's.

As an example, take the following sample graph with a best-fit line that suggests a positive, linear relationship between light bulb strength (in Watts) and photosynthetic rate (in parts per million of CO_2 consumed per minute) for some experimental setup:

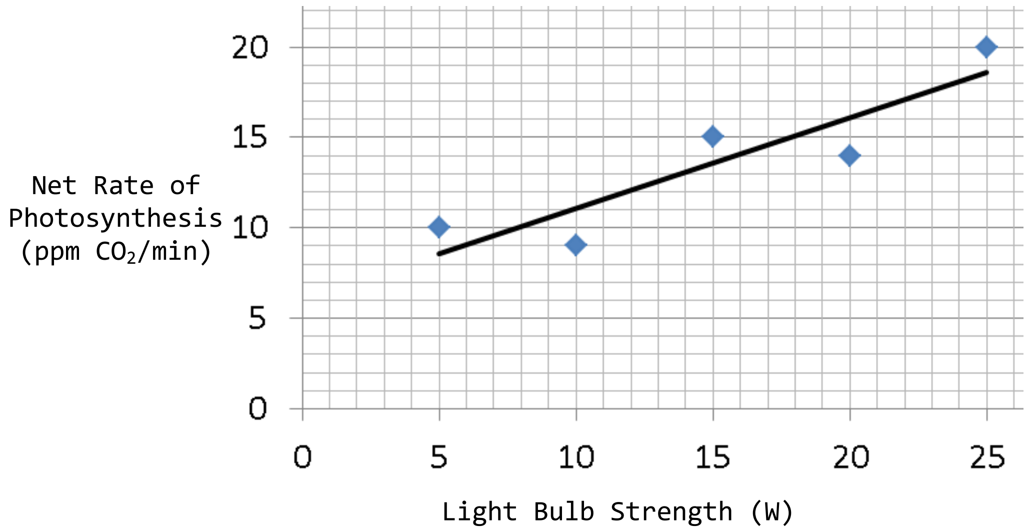


Fig 2. Sample graph of data used to demonstrate an R^2 calculation.

- Where the observed x values $(\mathbf{x_{obs}})$ were 5.0, 10.0, 15.0, 20.0, and 25.0.
- Where the observed y values $(\mathbf{y_{obs}})$ were 10.0, 9.0, 15.0, 14.0, and 20.0, with a mean of 13.6.
- Where the y values suggested by the line $(\mathbf{y_{line}})$ are 8.6, 11.1, 13.6, 16.1 and 18.6 at the observed x values.
- Then, solving for R^2, we have:

$$\boldsymbol{R^2 = 1 - \frac{\Sigma\,(y_{obs} - y_{line})^2}{\Sigma\,(y_{obs} - \bar{y}_{obs})^2}} = 1 - \frac{(10.0-8.6)^2+(9.0-11.1)^2+(15.0-13.6)^2+(14.0-16.1)^2+(20-18.6)^2}{(10.0-13.6)^2+(9.0-13.6)^2+(15.0-13.6)^2+(14.0-13.6)^2+(20.0-13.6)^2}$$

$$\boldsymbol{R^2 = 0.81}$$

C. Presenting and interpreting results

The following describes the steps you should now take in presenting and discussing your results.

1. Describing your results.

In the Results section of a report, a written description of your findings should be given that cites relevant figures or tables and might also provide the R^2 value which indicates the strength of your findings. For example,

> "Photosynthetic rate in a pea plant was found to increase in a linear fashion with increasing Wattage of the light bulb used (Fig 3, R^2 = 0.81)."

2. Visualizing your results.
Your R^2 value should be presented on your graph and your graph should be presented as formal figure that accompanies your description in the Results section of your assignment or report as follows:

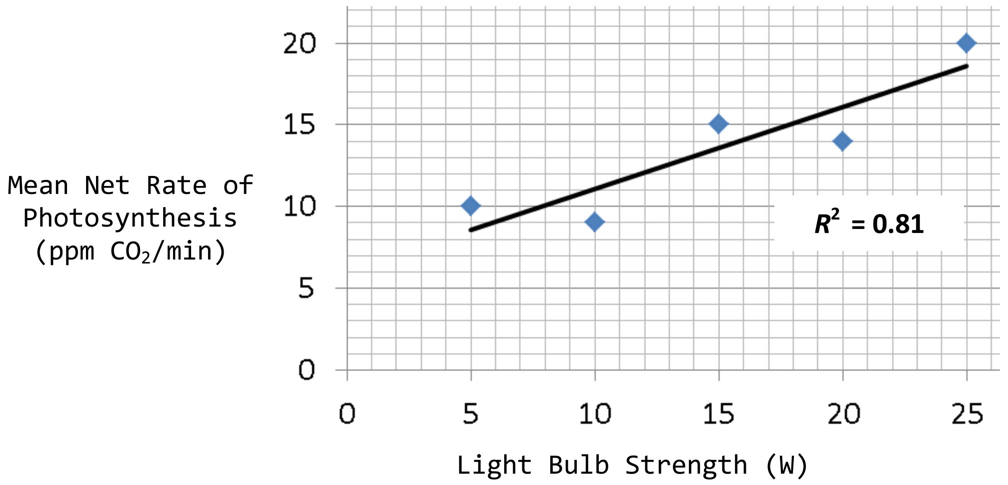


Fig 3. Photosynthetic rate in a pea plant grown using artificial light increases the Wattage of the bulb.

3. Interpreting your R^2 values.
The R^2 value of 0.81 indicates that 81% of the variation in y could be explained by variation in x. Generally, R^2 values of >0.7 indicate a strong, 0.4 - 0.7 a moderate, and <0.4 a weak if any relationship between the two variables. R^2 values of 1.0 are unlikely in an experiment due to 1) natural biological variation (e.g., the different plants used as replicates in this experiment would have had varying photosynthetic capacities), 2) experimental artifact (e.g., variation in the output of bulbs labeled as 5 vs. 10 Watts, etc.) and 3) experimental error (e.g., error in our ability to precisely measure the process of photosynthesis, or unseen damage to one or more plants during experimental setup).

4. Discussing your results.
After describing your results with both written text and a figure, you must discuss potential biological explanations for your results and make conclusions from your experiment. This discussion should again cite, where appropriate, the relevant data or figures that support your conclusion(s). Your discussion should also cite literature that contains information directly relevant to and can bolster your explanation(s). The cited literature would appear in a literature cited section (not shown here). For example,

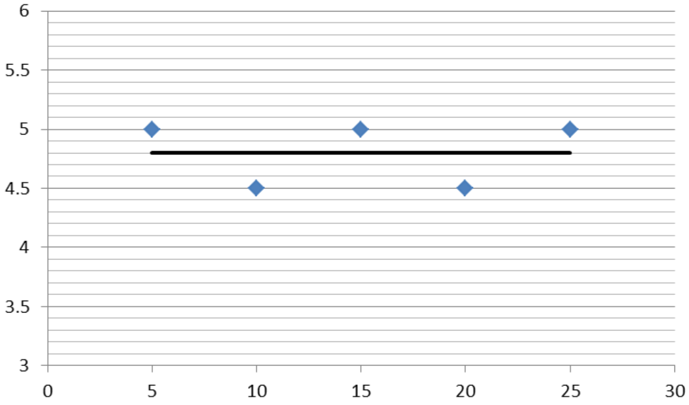
> "This study provides strong support for the hypothesis that increasing light intensity has a positive effect on the rate of photosynthesis, at least over the range of bulb Wattages for incandescent bulbs tested in this study (Fig 3). These results can be explained by the fact that as the Watts of electricity used by a bulb increases, so also increases the output of light which can then be used to drive higher rates of photosynthesis (Hoefnagels 2013). This generalization, of course, will depend on the nature of the bulb. Whereas incandescent bulbs were used in this study, fluorescent and mercury-vapor bulbs now on the market generally produce different wavelengths of visible light in different quantities and they generally utilize fewer Watts to produce that light (General Electric 2015)."

II. Practice Problems

Practice your R^2 calculations with the following sample data and graphs. You will probably want scratch paper on which to perform your calculations. Your final calculated value may not match the actual value due to you inability to precisely estimate y_{line} from the graphs. However, you should be within 0.1 to 0.01 of the actual answer. (Answers on the last page of this appendix.)

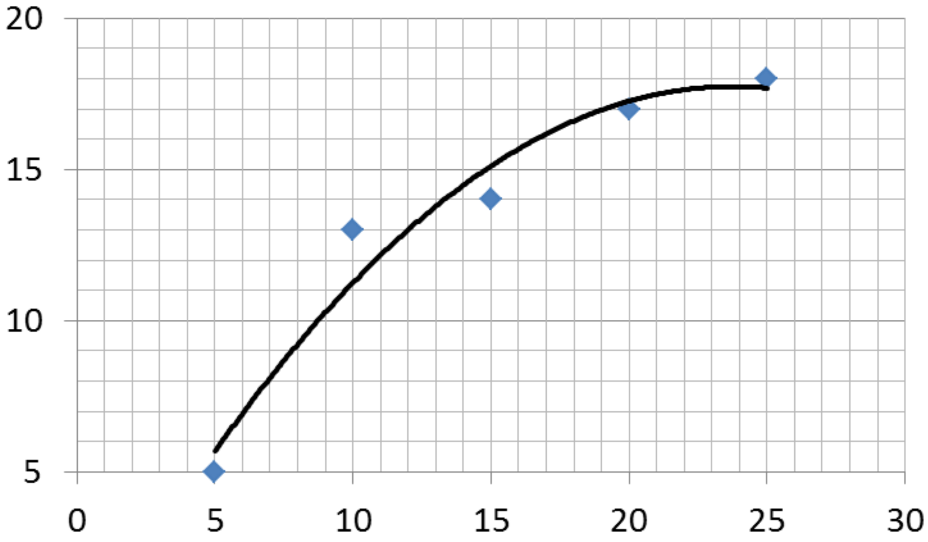
A. Sample 1. Calculate R^2 given the following data which were used to make the following graph.

x	Y
5	5
10	4.5
15	5
20	4.5
25	5



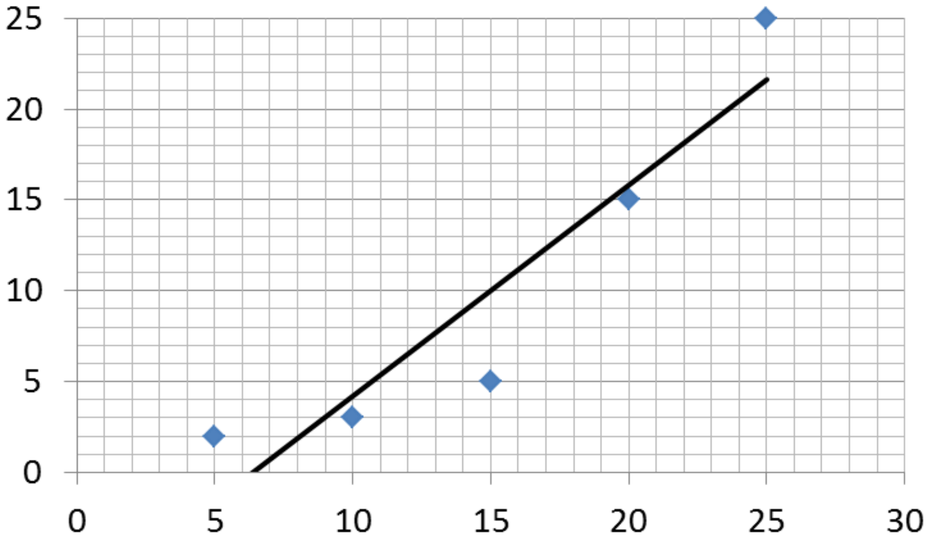
B. Sample 2. Calculate R^2 given the following data which were used to make the following graph.

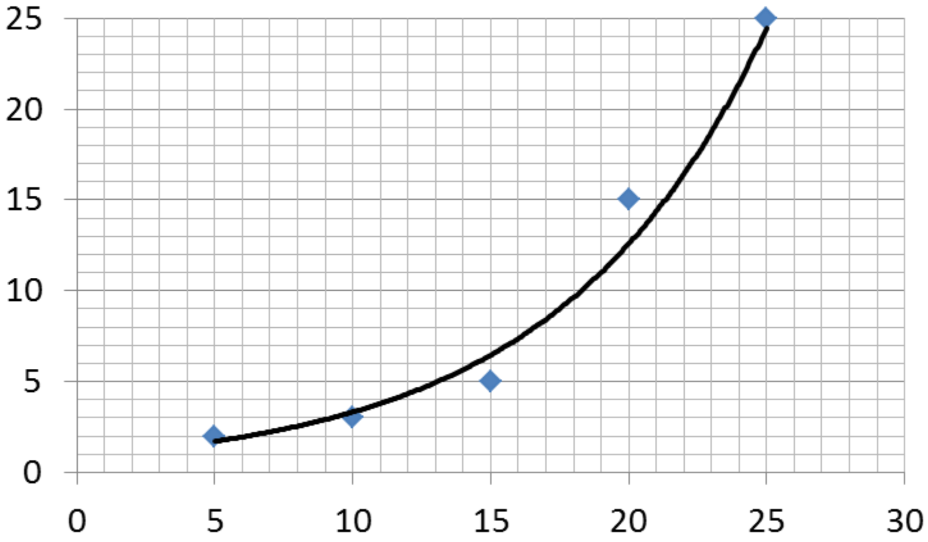
x	Y
5	5
10	13
15	14
20	17
25	18



C. Sample 3. Use R^2 to determine if the relationship between X and Y is linear (top graph) or exponential (lower graph). The line that yields the higher R^2 fits the data better.

x	Y
5	2
10	3
15	5
20	15
25	25





Credits: This appendix was developed by C.R. Hardy in January 2016.

Answers to practice problems:
A, R^2= 0.00; although the points are all very close to the line, the regression line has a slope of zero and so there is zero influence of X on Y.

B, R^2= 0.95.

C, although both graphs accurately show a positive relationship, the linear line in the top graph yields R^2= 0.87 whereas the exponential curve of the lower graph yields R^2= 0.97 with the same data. Thus, the exponential curve is a better representation of the data.