Association of Hours Studied to Exam Grade

Six students enrolled in a reading section of organic chemistry are preparing for their first exam. How are the hours each student studied and their exam grade associated?

Scatterplot

A **Scatterplot** of exam grade by hours studied variables shows the relationship on the same observation, in this case, student.

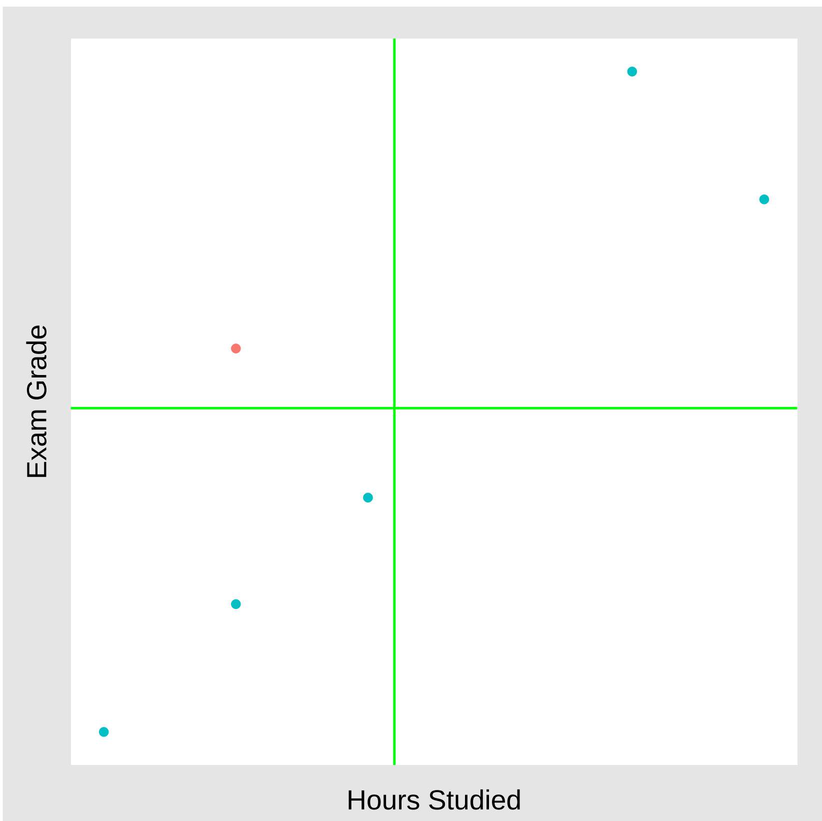

Figure 1: A scatterplot of Hours Studied v Exam Grade shows a possible linear relationship

Covariance

The **Covariance**, a measure of strength of the association between any two variables X and Y, denoted $Cov(X, Y)$ is calculated by first multiplying the deviations from their means, $Dev_{\bar{x}}$ and $Dev_{\bar{y}}$, then summing over all observations and dividing by N, the number of observations. This is very similar to the population variance calculation, and the

	examgrade	studyhours
Min.	57.0	1.0
1st Qu.	64.2	2.0
Median	71.5	2.5
Mean	72.2	3.2
3rd Qu.	80.2	4.5
Max.	88.0	6.0
Sum Sq Deviation	686.6	18.6
Variance	114.4	3.1
Standard Deviation	10.7	1.8

Table 1: Summary Statistics Hours Studied and Grades

	Exam Grade	Hours Studied	$Dev_{\bar{x}}hours$	$Dev_{\bar{y}}grade$
A	82	6	9.8	2.8
B	63	2	-9.2	-1.2
C	57	1	-15.2	-2.2
D	88	5	15.8	1.8
E	68	3	-4.2	-0.2
F	75	2	2.8	-1.2
Total	433.0	19.0	0.0	0.0
Total/N	$\bar{x} = 10.7$	$\bar{y} = 1.8$	0.0	0.0

variance can be thought of as the covariance of a variable with itself ie. $Var(X) = Cov(X, X)$.

$$Cov(X,Y) = \frac{\Sigma_{i=1}^{N} Dev_{\bar{x}} Dev_{\bar{y}}}{N}$$

The Covariance of Hours Studied with Exam Grade is 16.3 "Hours x Grade". These units make very little sense. We cannot compare covariances among variables in a data set if the units are different.

Linear Correlation

A standardized Covariance is the **Linear Correlation**, calculated by dividing each Covariance by the Standard Deviations of each of the variables:

$$Corr(X,Y) = \frac{Cov(Y,X)}{(StdDev(X)StdDev(Y))}$$

The Correlation of Hours Studied with Exam Grade is 0.84631 with **no units**, so the correlations of multiple pairs of variables can be compared.

Correlations are always between -1 and 1, and are a quantification of the linear relationship between two variables. A correlation of zero

	Exam Grade	Hours Studied	$(Dev_{\bar{x}})^2$	$(Dev_{\bar{y}})^2$	$Dev_{\bar{x}}Dev_{\bar{y}}hoursgrade$
A	82.0	6.0	96.0	7.8	27.4
B	63.0	2.0	84.6	1.4	11.0
C	57.0	1.0	231.0	4.8	33.4
D	88.0	5.0	249.6	3.2	28.4
E	68.0	3.0	17.6	0.0	0.8
F	75.0	2.0	7.8	1.4	-3.4
		Total	686.6	18.6	97.6
		Total/N	$Var(X) =$ 114.4	$Var(Y) =$ 3.1	$Cov(X,Y) =$ 16.3
		StdDev	$\sqrt{Var(X)} =$ 72.2	$\sqrt{Var(Y)} =$ 3.2	

means that there is linear relationship between two variables, although there may be a non-linear relationship. A correlation of 1 or -1 is indicates a perfect positive or negative linear relationship. $Corr(X, X) = 1$ always.

Correlation does not imply Causation! Even if two variables have a high or perfect correlation, there is not necessarily causation. Causation means X depends on Y or Y depends on X.

The Squared value of the correlation, 71.6%, called the Coefficient of Determination, and noted as R^2 is a measure of the "shared variance" of two variable, and the complement 28.4% is the proportion of variance not explained by the association.

Simple Linear Regression

When a linear correlation exists between two variables, we can explore causation using a **Simple Linear Regression**, also called Ordinary Least Squares (OLS), regressing a dependent variable, denoted Y, on an independent variable, denoted X as a line with the form:

$$Y = \alpha + \beta X + \epsilon \hat{Y} = \alpha + \beta X$$

This is very similar to the traditional algebra formula $y = mx + b$ with slope m and y-intercept b. In this case, the slope is β.

$$\beta = \frac{Cov(X, Y)}{Var(X)} = Corr(X, Y)\frac{StdDev(Y)}{StdDev(X)}$$

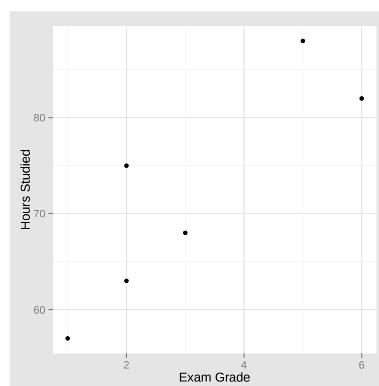

Figure 2: Green regression line with prediction error, as noted in red on the chart

Regressing exam grade on hours studied

$$\beta = \frac{16.3}{114.4} = 0.14$$

The linear regression always goes through the point $(\bar{x}, \bar{y})$, so returning to algebra, any point plus the slope determines the line:

$$\alpha = \bar{y} - \beta\bar{x}$$

$\hat{\alpha} = -6.91$ for our regression.

So,

$$\hat{y} = -6.91 + 0.14\bar{x}$$

The predicted value for any y_i is $\hat{y}_i$, and the prediction error is $\hat{\epsilon}_i = y_i - \hat{y}_i$.

Some properties of the Simple Linear Regression:

- $\Sigma_{i=1}^{N}\hat{\epsilon}_i = 0$
- $\Sigma_{i=1}^{N}x_i\hat{\epsilon}_i = 0$
- The predicted values $\hat{y}_i$ minimize the sum of the squared prediction errors, $\Sigma_{i=1}^{N}\hat{\epsilon}_i^2$, often referred to as Sum Squared Errors, or SSE.
- The regression equation is valid to predict $\hat{y}$ values in the range of X, that is, on the interval (min(X),max(X)), and any prediction will be in the range of (min(Y),max(Y))