Correlation

- Recall that correlation describes the strength of a relationship between two numeric variables, and that the ***Pearson product-moment correlation coefficient*** is a measure of the strength of the linear relationship between two variables.
- It is referred to as **Pearson's correlation** or simply as the correlation coefficient. If the relationship between the variables is not linear, then the correlation coefficient does not adequately represent the strength of the relationship between the variables.
- The symbol for Pearson's correlation is "ρ" when it is measured in the population and `r` when it is measured for a sample.
- As we will be dealing almost exclusively with samples, we will use `r` to to represent Pearson's correlation unless otherwise noted.
- Pearson's r can range from -1 to 1. An estimate of -1 indicates a perfect negative linear relationship between variables, an `r` of 0 indicates no linear relationship between variables, and an `r` of 1 indicates a perfect positive relationship between variables.
- Importantly it is assumed that the relationship in question is supposed to be linear. Some variables will in fact have a non-linear relationship (more on that later)

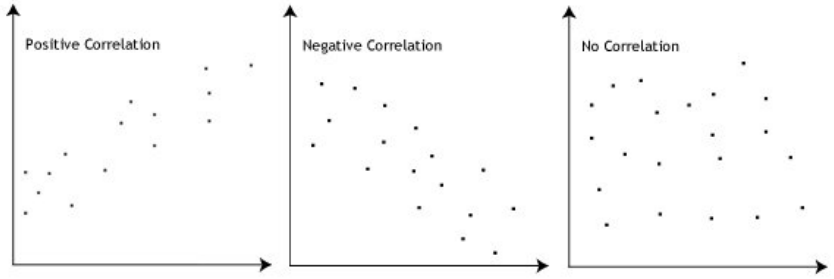


Figure 1:

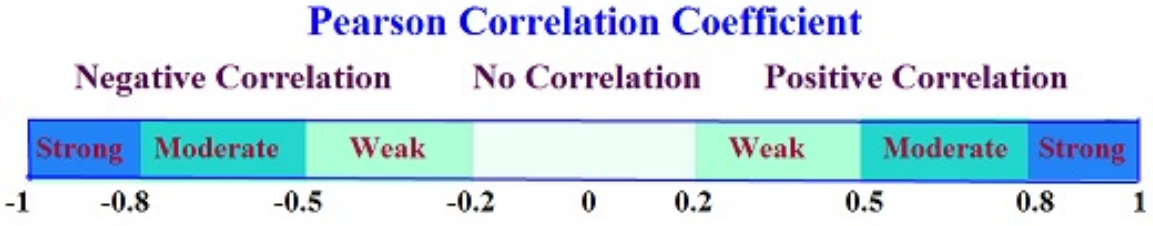
Implementation

The relevant R command to compute the correlation coefficient estimate is simply `cor()`.

- The strength of the relation is represented in a numeric value known at the **correlation coefficient**.
- This coefficient can take a value between -1 and 1. Additionally there are no units.

$$-1 \leq r \leq 1$$

- We can use the following graphic to help us interpret the correlation coefficient.




```
> X
[1] 104.40 104.14 104.84  99.34 104.13
[6] 100.93 103.85  97.16  96.18 101.42
```

```
>
> Y
[1]  98.39 106.05 111.18  97.65 104.02
[6] 100.18 106.20 101.87  92.49 101.41
>
> cor(X,Y)
[1] 0.7171676
>
```

Test for Significant Correlation

Getting a correlation coefficient is generally only half the story; you will want to know if the relationship is statistically significant. There is a more complex command called `cor.test()`. This command additionally provides a hypothesis test for the correlation estimate. The null and alternative hypotheses are as follows.

Ho : The correlation coefficient for the population of values is zero.
(i.e. No linear relationship.)

Ha : The coefficient is not zero.
(i.e. Linear relationship exists.)

```
> cor.test(X,Y)

Pearson's product-moment correlation

data:  X and Y
t = 2.9107, df = 8, p-value = 0.01957
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.1596151 0.9278331
sample estimates:
cor
0.7171676
```

- A confidence interval for the coefficient is provided for in the R output.
- If the interval includes 0 then we fail to reject the null hypothesis.

Spearman and Kendall Correlation Coefficients

- Non-parametric statistics are statistics that do not require any special assumptions (i.e. Assumption of normality).
- The **Spearman's rank-order** and **Kendall Tau** correlation coefficients are the **nonparametric** version of the Pearson product-moment correlation.
- Both methods measure the strength of association between two **ranked** (ordinal) variables.
- The coefficients are interpreted the same way as Pearson's Correlation Coefficient.

```
### Spearman Correlation Coefficient

cor(X,Y, method="spearman")

### Kendall Correlation Coefficient

cor(X,Y, method="kendall")

## [1] 0.6242424
## [1] 0.5111111
```

The Coefficient of Determination

- The coefficient of determination R^2 is the proportion of variability in a data set that is accounted for by the linear model.
- Equivalently R^2 provides a measure of how well future outcomes are likely to be predicted by the model.
- For simple linear regression, it can be computed by squaring the correlation coefficient. It is not specifically defined that way.
- This relationship is co-incidental when there are just two variables.

```
> summary(lm(Y~X))

Call:
lm(formula = Y ~ X)
....

Coefficients:
            Estimate Std. Error t value
(Intercept) -18.5506    41.4156  -0.448
X             1.1855     0.4073   2.911
Pr(>|t|)
(Intercept)   0.6661
X             0.0196 *
....
Residual standard error: 3.884 on 8 degrees of freedom
Multiple R-squared:  0.5143,    Adjusted R-squared:  0.4536
F-statistic: 8.472 on 1 and 8 DF,  p-value: 0.01957
```

The Adjusted R-square value

The adjusted R-square value is found on the summary output for a fitted model. It is called ***adjusted*** because it takes into account the number of predictor variables being used. The law of parsimony states the simplest model that adequately explains the outcomes is the best. The candidate model with the higher adjusted R squared is considered preferable.

The Akaike Information Criterion

The AIC is a model selection metric often used in statistics. It is computed using the R command `AIC()`. The candidate model with the smallest AIC value is considered preferable.