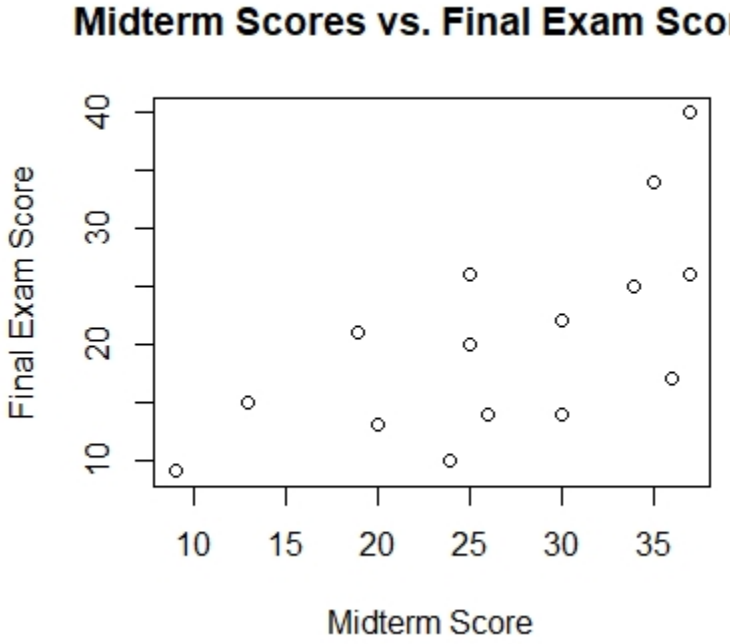
Stat 260 Lecture Notes
Set 3 - The Correlation Coefficient

When data arise in pairs, such as (x_1, y_1), (x_2, y_2), (x_3, y_3) $\ldots$ (x_n, y_n) the structure is called **bivariate**.

Here we may want to see if there is a relationship between the x and y values. To do this we can use a **scatterplot**.



Be careful with **extrapolating** (making predictions). This data only shows what happens on the final exam for students with midterm scores between 9 and 37. The data included here would not be useful for making final exam predictions for midterm scores of, say, 80. (In other words, your data is only useful in making predictions for other data values close to your collection.)

Recall: variance $= s^2 = \dfrac{\sum (x_i - \overline{x})^2}{n-1}$

When we work with bivariate data we can calculate the **covariance**, s_{xy}.

$$s_{xy} = \frac{\sum\ (x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

Example 1: Calculate the covariance for the data $(3, 4)$, $(8, 7)$, $(10, 8)$, $(11, 8)$.

The **correlation coefficient**, r, measures the strength of the linear relationship between x and y values.

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}} = \frac{s_{xy}}{s_x s_y}$$

where s_{xy} = the covariance of x and y, s_x = the standard deviation of x values, s_y = the standard deviation of y values

Example 2: Calculate the correlation coefficient using the data from Example 1.

Rule: No matter what we have for x and y values, $-1 \leq r \leq +1$.

The closer that that correlation coefficient r is to $+1$ or -1, the stronger the linear relationship there is. A positive value of r indicates a positive linear relationship, and a negative value of r indicates a negative linear relationship. A value of 0 indicates no linear relationship.

An exact linear relationship occurs when $r = 1$ or $r = -1$ and in this case we can represent the data as a straight line in the form $y_i = ax_i + b$.

Be careful! Correlation $\neq$ causation.