Objectives:

- To develop an awareness of ***least squares regression*** as a technique for modeling the relationship between two variables
- To learn to use ***regression lines*** to make ***predictions*** and to recognize the limitations of those predictions
- To understand some concepts associated with regression such as ***fitted values, residuals,*** and ***proportion of variability explained***
- To use the calculator to apply regression techniques with judgment and thoughtfulness to genuine data
- To understand the distinction and importance of ***outliers*** and ***influential observations*** in the context of the regression analysis
- To learn to use ***residual plots*** to indicate when the linear relationship is not a satisfactory model for describing the relationship between two variables
- To discover how to ***transform*** variables to create a linear relationship between variables

Equation of a line *y = a + bx (Same concept as y = mx + b statisticians need to be different!)*
x represents the variable that will be used to predict (dependent)
y represents the variable to be predicted (independent)
a represents what happens when *x = 0* (*y-intercept*)
b represents the slope of the line (coefficient of relationship)

slope may be calculated by dividing the amount the line rises by how far it goes to the right to obtain that rise (rise/run)

Correlation coefficient A simpler way to calculate Correlation coefficient does exist – This looks messier, but notice it only involves x's and y's, no means or standard deviations. Well here it is:

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\left(\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\right)\left(\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}\right)}$$

n = number of data pairs

$\sum x$ = sum of all x's

$\sum xy$ = sum of all pairs of x and y's products

$\sum x^2$ = sum of all squares of x's

$\left(\sum x\right)^2$ = the square of the sum of all the x's

Least Squares The least squares method uses the *y = a + bx* form of the equation using the following for finding *a* and *b*

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

s_x = standard deviation of x values

$$a = \bar{y} - b\bar{x}$$

$\bar{x} = mean(x)$

Prediction the primary use of regression is for the prediction of y when x is known. If the correlation is strong enough, we then calculate *a* and *b* and form the line. If we plot *x* and calculate for the corresponding *y* we have a good idea what *y* will be in most situations (most being relative depending on the value of *r*.

Extrapolation and interpolation If we make a prediction based on data that surrounds the we are looking at, it is called ***interpolation.*** If we make a prediction and all the previous data is on one side of the data point we are predicting it is called ***extrapolation.*** Extrapolation can be dangerous since we onle know one side of the information, we don't know for sure if the tendency carries to the other side of the point we are trying to predict. Furthermore – there is no way to know if it does in fact carry through beyond the know points. A relationship may appear linear through the known points but may shoot up exponentially beyond the data. (Population data often does this)

Fit/residual Each data point is to be thought of as containing two parts. The first part is the ***fit,*** what we predict to occur (The prediction). The second part is the leftover part; the amount the prediction deviates from the actual value. This is called the ***residual.***

residual = actual – fit

The residual is the measure of the vertical distance from the regression line to the actual data points.

Proportion of variability the square of the correlation coefficient (r^2). a measure of confidence in the prediction we make based on the regression line. We may think of this as the percentage of data that is explained by the regression model. If r^2 =1 the 100% of the data is explained by the regression. If r^2 = .9 then 90% of the data is explained by the regression. We have 90% confidence in the validity of the prediction.

Outliers In the context of regression lines, ***outliers*** are observations with large (in absolute value) residuals. Observations that are unusually far from the regression line – not following the typical pattern.

Influential observation: an observation that has a great influence on the regression line by virtue of an extreme x-value. Note that it does not have to be an outlier to be an influential observation, i.e. if you remove an extreme x-value observation, which may closely follow the regression line, the regression line can be changed considerably. But, if you remove a "typical" x-value observation the line will not change very much.

Non-linear relationships: Often the plot of relationships do not form a line but still may have a stable, predictable relationship, e.g. the path of a ball thrown in the air, a planets position and distance from the sun. In these cases, we have a curvilinear relationship. In these instances a nice pattern can be seen in a plot of the residuals. If we can see the pattern we can create a ***transformation*** to the data to make a linear relationship that can be studied.

Examples of transformations that may be used:

Take the log of the planets distance from sun: *log(x)*

Take the square root of the height of a ball in flight: *sqrt(x)*

Take the square root of the period of a planets revolution to relate it to the area swept by the planets path: $sqrt(T^3)$ (Keplers law)