2 Examining Relationships

Exploratory data analysis consists of the following steps
→ Examine each variable individually, then study relationships between variables.
→ Use graphs and then use numerical summaries.
We have already studied how to analyze one variable in detail. In the case of relationships between variables, we are going to consider only numerical varibles and examine relationships between them. The graphical technique that we are going to use to examine relationships between two numerical variables is known as **scatterplot analysis**. The numerical summary of relationship between two numerical analysis is **correlation**. We will also look at **regression analysis** as a technique of explaining how one variable affects another.

The study of relationships between two variables looks at possible **cause and effect** relationships. We might be interested in testing the relationship between amount of rainfall and yield, or in particular we might be interested in testing to see if the amount of rainfall has an effect on the yield, and if so how much does it affect it.
Hence we can divide variables into one of two types.

Response Variable: measures the outcome of a study. (effect)
Explanatory Variable: explains/influences change in the response variable. (cause)

Note that a cause and effect relationship need not be necessary when considering relationships between two variables.
Is there a cause and effect relationship between the following variables ? If so which is the response variable and which is the explanatory variable?

1. The number of hours studied and grade.

2. Student grade in economics and statistics.

3. Exericise and heart rate.

4. Incomes of husband and wife.

5. Temperatures in Atlanta and Raleigh.

2.1 Scatterplots

Scatterplots: show the relationship between two quantitative variables.. To draw the scatterplot, just plot the point on its corresponding X,Y axis.

2.1.1 Interpreting Scatterplots

What to look for in a scatterplot:
→ Overall Pattern
→ Form, Direction and Strength
→ Outliers

Associations can be of two types.

Positive: When above average values of one variable are associated with above average values with above average values of the other and the same for below average values.

Negative: When above average values of one variable are associated with below average values of the other and vice versa.

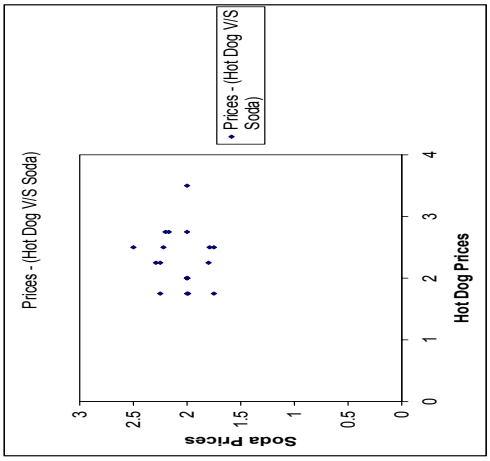

Figure 1: Scatter Plot of prices of Hot Dogs and Sodas at MLB stadiums

2.2 Correlation

A scatterplot tells us a lot about the form, direction and strength of the relationship between two quantitative variables. One way to describe the form of relationships is a line. A linear relationship is one of the most widely used ways of desribing relationships. A linear relationship is said to be strong if the

points on the scatterplot are close to the line and is said to be weak if the points are away from the line. Correlation is a measure of how far the points are away from the line that describes the relationship.

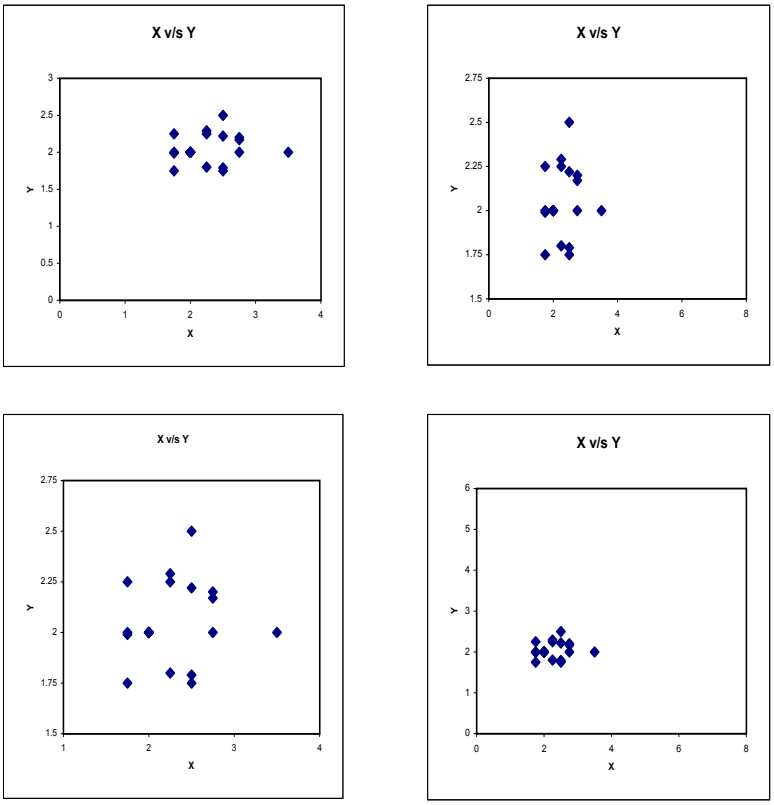

Figure 2: Scatter Plot and Correlation: What are the correlations?

2.2.1 The correlation r

The **correlation** measures the direction and strength of the **linear relationship** between two quantitative variables. It is usually denoted by r. The correlation r between two quantitaive variables x and y

is

$$r = \frac{1}{n-1}\sum(\frac{x_i - \bar{x}}{s_x})(\frac{y_i - \bar{y}}{s_y})$$

where $\bar{x}$, $\bar{y}$, s_x and s_y are the means and standard deviations of x and y respectively.
The following data relate the number of suitors for 7 bees with their degree of symmetry (a sign of beauty)

Degree of Symmetry	Number of Suitors
58	1
67	2
70	2
75	3
81	3
90	4
98	5

Table 1: Beauty and Attraction

Some summary statistics
$n = 7 \quad \bar{x} = 77 \quad s_x = 13.784 \quad \bar{y} = 2.857 \quad s_y = 1.345$

$(x - \bar{x})$	$(\frac{x-\bar{x}}{s_x})$	$(y - \bar{y})$	$(\frac{y-\bar{y}}{s_y})$	$(\frac{x-\bar{x}}{s_x})(\frac{y-\bar{y}}{s_y})$
-19	-1.378	-1.857	-1.381	1.903
-10	-0.725	-0.857	-0.637	0.462
-7	-0.508	-0.857	-0.637	0.324
-2	-0.145	0.143	0.106	-0.015
4	0.290	0.143	0.106	0.031
13	0.943	1.143	0.850	0.801
21	1.5235	2.143	1.593	2.427
$\sum(\frac{x-\bar{x}}{s_x})(\frac{y-\bar{y}}{s_y})$				5.932
$r = \frac{1}{n-1}\sum(\frac{x-\bar{x}}{s_x})(\frac{y-\bar{y}}{s_y})$				0.989

2.2.2 Facts about correlation

Some facts about correlation

- Correlation makes no distinction between explanatory and response variables. Hence the correlation between height and weight is the same as the correlation between weight and height.

- We require quantitative variables to calculate correlation. We cannot find the correlation between incomes of people and the city they live in.
- r does not change with changes in units of measurment. In fact the correlation r itself has no units. It is just a number between -1 and 1.
- Positive r $(0 < r < 1)$ indicates positive association between the variables and negative r $(-1 < r < 0)$ indicates negative association between the variables.
- Values of r very close to 0 indicate a very weak linear relationship. The strength of the linear relationship increases as r moves away from 0 towards 1 or -1. A correlation of 1 or -1 indicates a perfect linear relationship.
- Since r measures only the linear relationship between the variables, nothing can be said about curved or quadratic relationships.
- The correlation is affected by outliers just like the mean and standard deviation. Use r with caution when dealing with datasets with outliers.

→ Calculate the correlation for the following data.

Tobec Estimate	0.1	1.9	4	4.6	4.2	4.2	7.1
Weight of Fat Pads	216	233	251	272	282	295	298
Tobec Estimate	5.4	6.2	4.6	3.6	7.1	5.4	
Weight of Fat Pads	303	311	318	318	327	335	

Table 2: Tobec Estimates and Actual Weight of Fat Pads

2.3 Least Square Regression

Suppose a scatterplot suggests a linear relationship between two variables. Correlation would give us the strength and direction of the relationship, but does not give us the actual line. However a **regression line** summarizes the relationship between the two variables in a specific setting where you have an explanatory variable and a response variable. To be specific the regression line describes how a response variable y changes as the explanantory variable x changes.

2.3.1 The least-squares regression line

Based on the scatterplot one can draw many lines through the points, based on our best guess. However, the line that is closest to all points in terms of average distance is also called the **Least Squares Regression Line** .
The **Least Squares Regression Line** is the line that makes the sum of squares of the vertical distances of the data points from the line as small as possible.
The equation of the least squares is given by $\hat{y} = a + bx$, where b is the slope, given by $b = r(\frac{s_y}{s_x})$ and a is the intercept, given by $a = \bar{y} - (b)(\bar{x})$. The least squares regression line calculations for the data on beauty and attraction are given below.

$$\bar{x} = 77 \quad \bar{y} = 2.857 \quad s_x = 13.784 \quad s_y = 1.345 \quad r = 0.989$$

$$b = r(\tfrac{s_y}{s_x}) = 0.989(\tfrac{1.345}{13.784}) = 0.096 \quad a = \bar{y} - (b)(\bar{x}) = 2.857 - (0.096)(77) = -4.573$$

The equation of the regression line is $\hat{y} = -4.573 + 0.096.x$

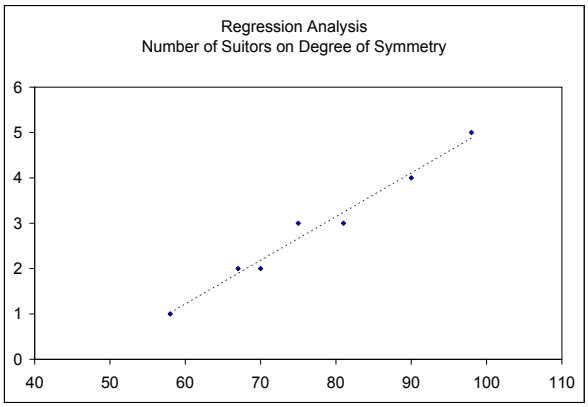

Figure 3: The least squares regression line

Interpretation of the regression line

Slope: The rate of change in y. The amount of change in y when x increases by 1 unit.

Intercept: The value of $\hat{y}$ when x=0. Note that sometimes x wont take values at 0, so it may not make sense to talk about the intercept.

Prediction: The value of $\hat{y}$ for given values of x. We should be careful when we make predictions for values of x outside the data range.

→ Calculate the least squares regression line for data from table 2 on the relationship between tobec estimates and actual weights.

2.3.2 Facts about least-squares regression

- There is a distinction between the explanatory variable and the response variable. If you switch them you will have a different line.
- There is a close relationship between correlation and slope.

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

  A change of 1 standard deviation in x corresponds to a change of r standard deviations in y. As $r \uparrow 1$ or $\downarrow -1$ the change in $\hat{y}$ becomes the same as change in x. A change of 13.784 degrees of symmetry for a bee results in a $1.345 * 0.989 = 1.33$ change in the number of suitors.

- The least square regression line passes through $(\bar{x}, \bar{y})$.
- The square of the correlation is r^2 and is described as the variation in y that is explained by the variation in x. The percentage of variation in the number of suitors for the bees that can be explained by the degree of symmetry is $(0.989)^2 = o.978$

2.3.3 Residuals

Residuals are a good way to analyze the fit of least squares regression. Residual = Observed - Predicted $= y - \hat{y}$. The mean of the residuals is 0.

Residual Plot: is a scatter plot of residuals against the explanatory variable, i.e. a plot of $y - \hat{y}$ v/s x.
How to analyze residual plots?

- curved patterns $\Rightarrow$ no linear relationship
- increasing spread $\Rightarrow$ prediction of $\hat{y}$ will be less accurate for large x and vice versa.
- individual plots with large residuals are outliers. For data from table 2, leave out the last observation and see the difference in the regression line. (The last observation is an outlier)
- individual points that are extreme in the x direction (influential observations) have strong influence on the regression line. Again for table 2, leave out the first observation and see how the regression line changes. (The first observation is an influential one)

2.4 Cautions with Correlation and Regression

Correlation and Regression analysis are useful techniques to analyze relationship between two quantitative variables. However they have their limitations. They describe only linear relationships. They are also easily affected by outliers. Here are a few more things to remember.

2.4.1 Extrapolation

Extrapolation is the act of using the regression line to predict values for x far outside the range of values on which the regression line is based. For example, for the data on beauty and attraction from Table 1, predicting values of y when x is 30 not only gives us a value which does not make sense (-1.678, a negative number for the number of suitors), it is beyond the range of values for which the regression line is based (only for values of x between 58 and 98).

2.4.2 Lurking Variables

A **lurking variable** is a variable that has an important effect on the relationship between the variables in a study but is not included among the variables studied. An example would be the relationship between x - the number of coffees before an exam and y - the grade obtained on the exam, the lurking variable being the number of hours studied.

2.4.3 Association is not causation

Association does not imply causation. If you are using x to predict y, it does not mean that changes in x cause changes in y. An example would be to use temperatures in Raleigh to predict the temperatures of Durham. There might be a strong association, but changes in temperature in Raleigh dont change temperatures in Durham. There might be a common phenomenon affecting temperatures in both cities!