CORRELATION and REGRESSION ANALYSIS

1. What is Correlation

Correlation techniques are used to:

- Explore the association between pairs of variables (correlation)
- Predict scores on one variable from scores on another variable (bivariate regression)
- Predict scores on a dependent variable from scores on a number of independent variables (multiple regression)

2. Correlation vs. Causality

- Correlation provides an indication of a relationship between variables
- It does not indicate that one variable causes the other
- Strong correlation between variables A & B
    - A causes B?
    - B causes A?
    - C causes both A & B ?
- Ice cream sales and homicides in New York, Smoking and lung cancer

3. Correlation Statistics

- Correlation is used to describe the strength and direction of the relationship between two variables (usually continuous – but can be used when one of the variables is dichotomous i.e. has only two values)

- The most common statistic obtained is
  Pearson's product-moment correlation (r)

- Partial correlation is used when you wish to explore the relationship between two variables while statistically controlling for a third variable

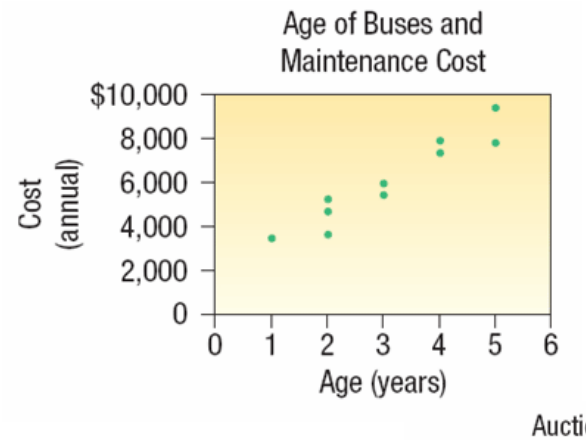
4. What is Scatter Diagram

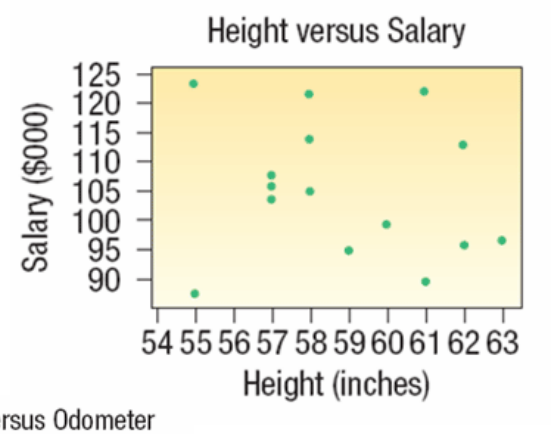
A scatterplot will provide information on both the direction of the relationship (positive or negative) and the strength of the relationship

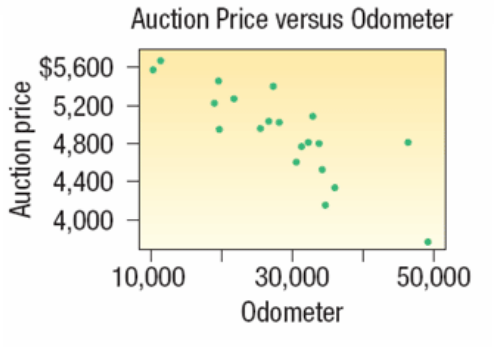
a. Perfect correlation – a straight line

b. No correlation (r = 0) – blob of points/no pattern

.







5. Correlation coefficient

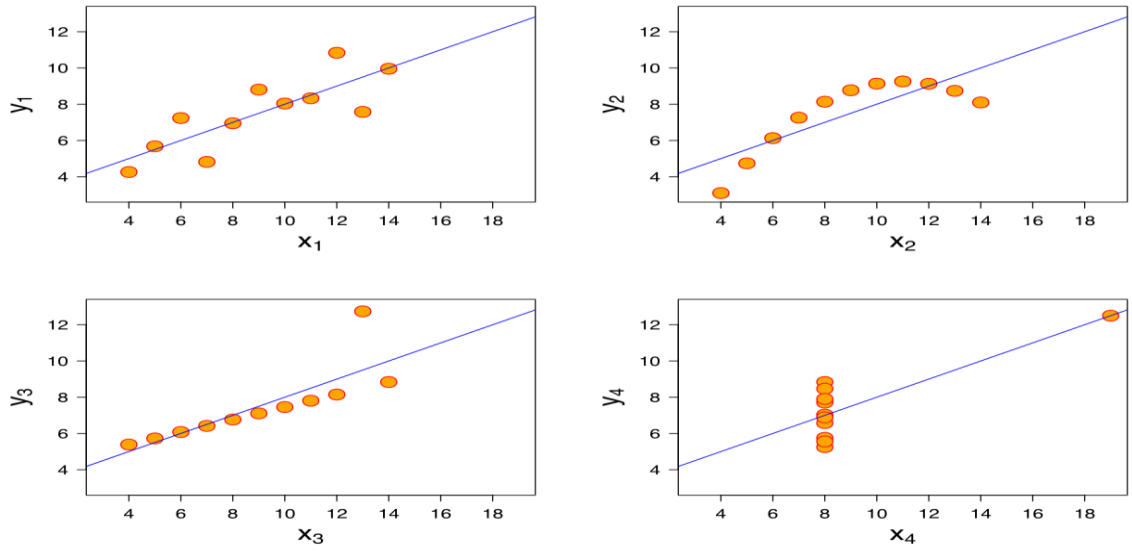
- The correlation coefficient (r) provides an indication of the linear (straight line) relationship between variables.
    - Pearson's r will underestimate the the strength of relationship when the variables are related in non-linear form
- Outliers can have a dramatic effect on the correlation coefficient (especially with small samples) – check scatterplot
- Be careful of a restricted range of scores – there should be as wide a range of scores on each of the two variables as possible.

6. Anscombe's Quartet:

- They are four set of data with different scatter plot.
- Correlation = 0.816 in all cases!
- It is always advisable to view your data

Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



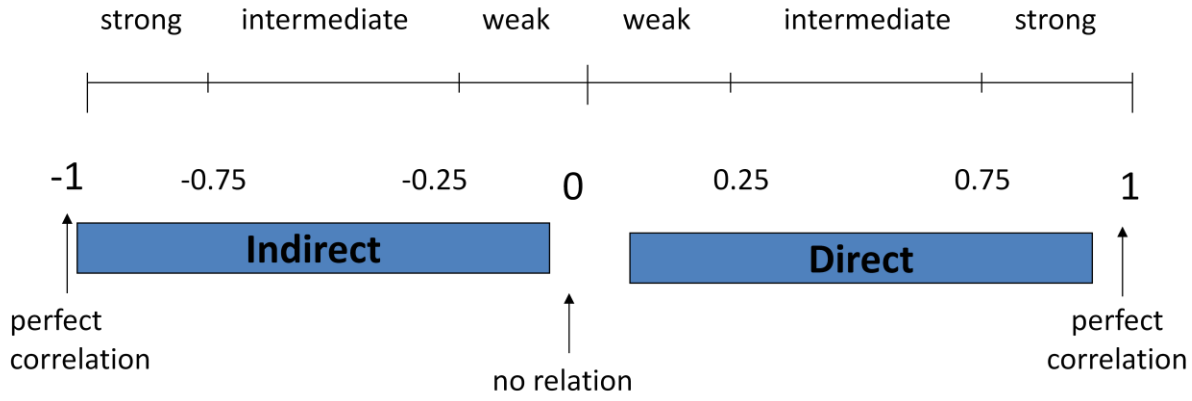
7. Dependent vs. Independent Variable

- The **dependent variable** is the variable being predicted or estimated.
- The **independent variable** provides the basis for estimation. It is the predictor variable.
- In the questions below, which are the dependent and independent variables?

1. Is the number of square feet in a home related to the cost to heat the home in January?
2. In a study of fuel efficiency, is there a relationship between miles per gallon and the weight of a car?
3. Does the number of hours that students studied for an exam influence the exam score?

8. More on Coefficient of Correlation, *r*

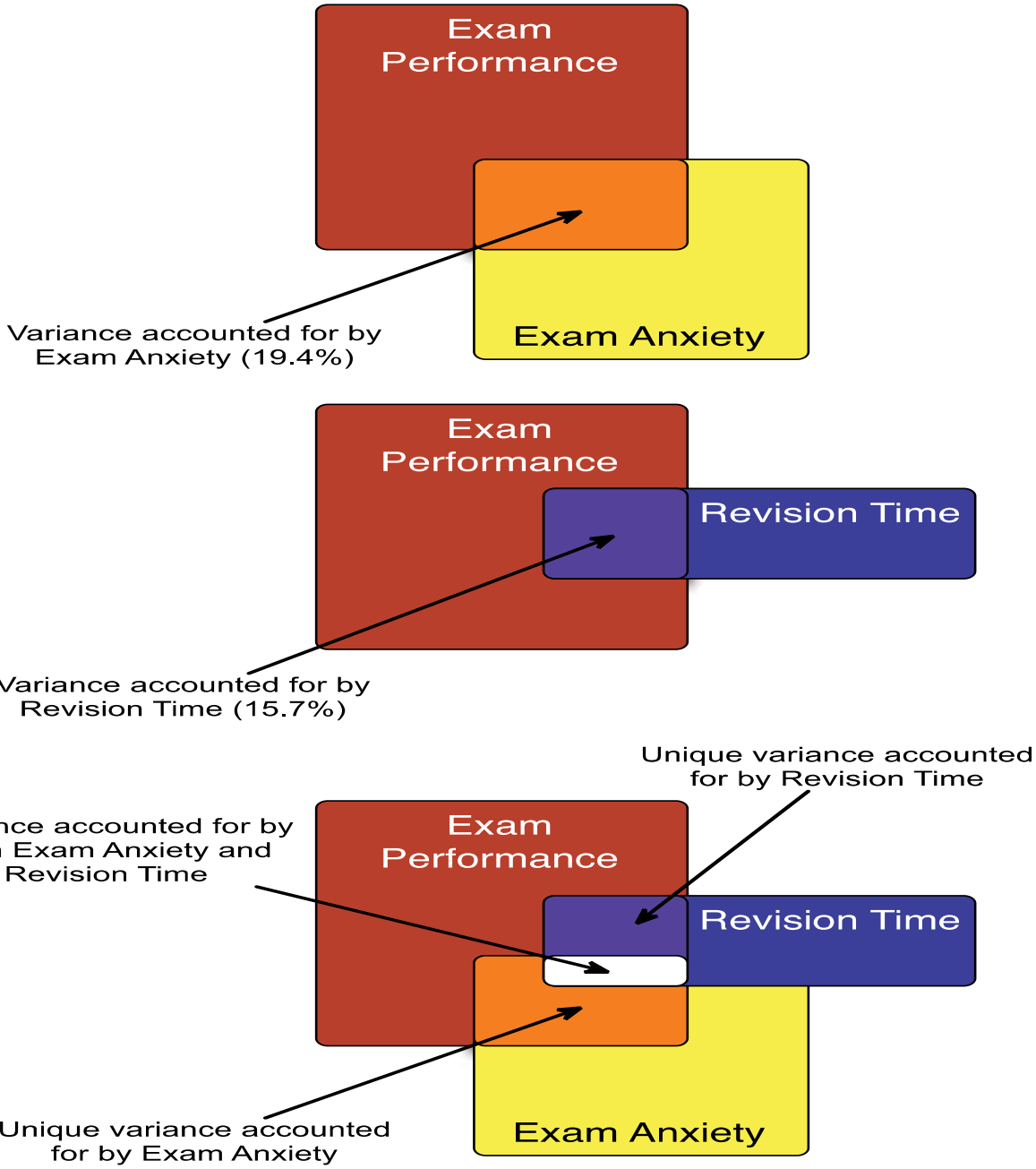
- Shows the direction and strength of the linear relationship between two interval or ratio-scale variables.
- Ranges from –1.00 to +1.00.
- Values of –1.00 or +1.00 indicate perfect and strong correlation.
- Values close to 0.0 indicate weak correlation.
- Negative values indicate an **inverse** relationship, and positive values indicate a **direct** relationship.



But in real world it really depends on the field in which you are using the correlation. Social sciences may accept lower correlations as being intermediate and strong as very high correlations are extremely rare in that field.

9. Partial Correlation

- Measures the relationship between two variables, controlling for the effect that a third variable has on them both.
- Allows you to control for an additional variable
- Statistically remove the influence of a 'confounding' variable
- Relationship between A and B is influenced to some extent by a third variable C


Exam Performance
Exam Anxiety
Variance accounted for by Exam Anxiety (19.4%)
Exam Performance
Revision Time
Variance accounted for by Revision Time (15.7%)
Unique variance accounted for by Revision Time
Variance accounted for by both Exam Anxiety and Revision Time
Exam Performance
Revision Time
Exam Anxiety
Unique variance accounted for by Exam Anxiety

10. REGRESSION ANALYSIS

Regression analysis is an attractive extension to correlation analysis because it postulates a model that can be used not only to measure the direction and the strength of a relationship between the response and predictor variables, but also to numerically describe that relationship

A regression is used to understand the cause - effect relationships.

Regression analysis is a very widely used statistical tool to establish a relationship model between two or more variables.

Example:

1. Predicting weight of a person when his height is known.
2. Understand the factors that may influence the birth weight of a baby What factors would you think of?

11. Regression Equation

In regression analysis, we use the independent variable (*X*) to estimate the dependent variable (*Y*). It is done with the help of an equation that expresses linear relationship between two variables as follows. Y=mx + c

- The relationship between the variables is linear.
- Both variables must be at least interval scale.
- The least squares criterion is used to determine the equation.

12. Prerequisites for Regression

Two major conditions must be met before you apply a simple linear regression model to a data set:

- The y's must have an approximately normal distribution for each value of x.
- The y's must have a constant amount of spread (standard deviation) for each value of x.

13. Slope

It is the rate of change of y when x changes

SLOPE OF THE REGRESSION LINE $$b = r\frac{S_y}{S_x}$$

Where

r is the correlation coefficient.
s_y is the standard deviation of Y (the dependent variable).
s_x is the standard deviation of X (the independent variable).

14. Intercept

The y-intercept is the point where the line crosses the y-axis; in other words, it's the value of y when x equals zero.

Y-INTERCEPT $a = \overline{Y} - b\overline{X}$

where
$\overline{Y}$ is the mean of Y (the dependent variable).
$\overline{X}$ is the mean of X (the independent variable).

- The y-intercept of a regression line may or may not have a practical meaning depending on the situation.
- Does the y-intercept fall within the actual values in the data set? If yes, it has practical meaning.
- Does the y-intercept fall into negative territory where negative y-values aren't possible?
- Does the value x = 0 have practical meaning? For example, if x is temperature then x = 0 may be a value that's relevant to examine. If x = 0 has practical meaning, then the y-intercept does too, because it represents the value of y when x = 0. If the value of x = 0 doesn't have practical meaning in its own right then the y-intercept doesn't either.

15. Coefficient of Determination

The **coefficient of determination** (r^2) is the proportion of the total variation in the dependent variable (Y) that is explained or accounted for by the variation in the independent variable (X). It is the square of the coefficient of correlation.

- Value ranges from 0 to 1, or 0 to 100%.
- Does not give indication on the direction of the relationship between the variables.

16. Residuals

- A residual is the difference between the predicted value (from the best-fitting line) and the observed value of y, also known as y (from the data set).
- Its notation is $(y - \hat{y})$. Specifically, for any data point, you take its observed y-value (from the data) and subtract its expected y-value (from the line).
- If the residual is large, the line doesn't fit well in that spot. If the residual is small, the line fits well in that spot.

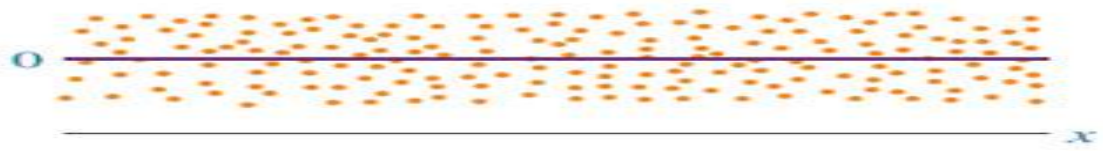
17. Using Residuals to Test the Assumptions of the Regression Model

- the model is linear
- the error terms have constant variances (homoscedasticity)
- the error terms are independent
- the error terms are normally distributed

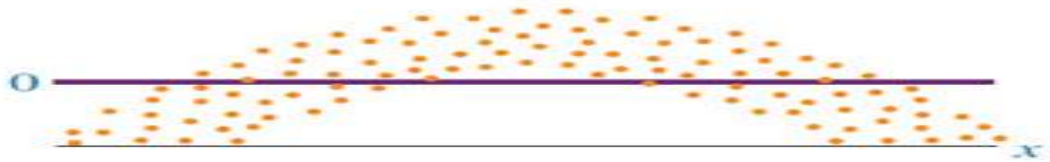
18. Residual plot

A graph in which the residuals for a particular regression model are plotted along with their associated value of *x.*

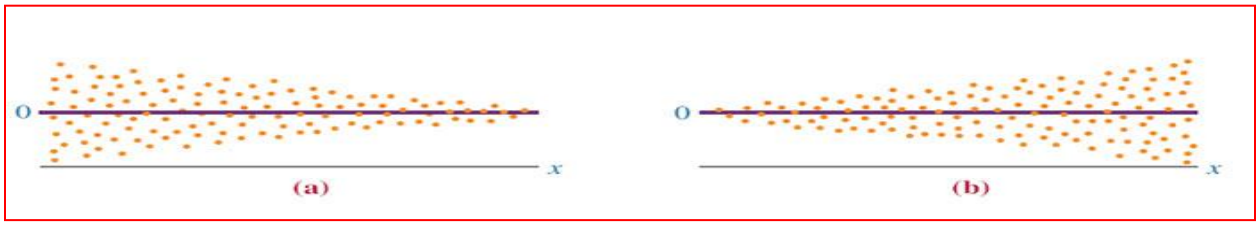
A healthy residual plot will look like this. Otherwise it will fall into any of below mentioned categories which shows that linear regression is not able to explain the relationship.



A non linear residual plot will look like this



A non constant error variance will look like this.



A graph of non independent error terms will look like this