0.1 Course Overview

1. Exploratory Data Analysis. Revision of measures of centrality and dispersion.
2. Inference Procedures. Confidence Intervals. Hypothesis tests. Significance and the probability value.
3. Regression Models. revision of simple linear regression. Introduction to multiple linear regression. Correlation matrices. Multicollinearity.
4. Analysis of Variance. Variable Selection procedures; Backward elimination, forward selection, stepwise regression.
5. The Data Deluge. What is Data. Sources of Data. Storage Issues.
6. Data Quality. Data scrubbing
7. Introduction to Research . Qualitative Research. Quantitative Research. The purpose of Research.
8. Introduction to Databases. Structures Query Language (SQL).

Contents

Chapter 1

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to

- maximize insight into a data set;
- uncover underlying structure;
- extract important variables;
- detect outliers and anomalies;
- test underlying assumptions;
- develop parsimonious models; and
- determine optimal factor settings.

The seminal work in EDA is Exploratory Data Analysis, Tukey, (1977).

Most EDA techniques are graphical in nature with a few quantitative techniques. The reason for the heavy reliance on graphics is that by its very nature the main role of EDA is to open-mindedly explore, and graphics gives the analysts unparalleled power to do so, enticing the data to reveal its structural secrets, and being always ready to gain some new, often unsuspected, insight into the data. In combination with the natural pattern-recognition capabilities that we all possess, graphics provides, of course, unparalleled power to carry this out.

EDA techniques are subjective and depend on interpretation which may differ from analyst to analyst, although experienced analysts commonly arrive at identical conclusions. The particular graphical techniques employed in EDA are often quite simple, consisting of various techniques of:

- Plotting the raw data (such as data traces, histograms, bihistograms, probability plots, lag plots, block plots, and Youden plots.
- Plotting simple statistics such as mean plots, standard deviation plots, box plots, and main effects plots of the raw data.
- Positioning such plots so as to maximize our natural pattern-recognition abilities, such as using multiple plots per page.

1.0.1 Summary Analysis

A summary analysis is simply a numeric reduction of a historical data set. It is quite passive. Its focus is in the past. Quite commonly, its purpose is to simply arrive at a few key statistics (for example, mean and standard deviation) which may then either replace the data set or be added to the data set in the form of a summary table.

1.1 Revision of basic measures

1.1.1 Measures of Centrality

The most common measures of centrality are the mean and median. **Median** Another measure of location just like the mean. The value that divides the frequency distribution in half when all data values are listed in order. It is insensitive to small numbers of extreme scores in a distribution. Therefore, it is the preferred measure of central tendency for a skewed distribution (in which the mean would be biased) and is usually paired with the **interquartile range** (IQR) as the accompanying measure of dispersion.

The trimmed mean

The trimmed mean looks to reduce the effects of outliers on the calculated average. This method is best suited for data with large, erratic deviations or extremely skewed distributions. A trimmed mean is stated as a mean trimmed by X%, where X is the sum of the percentage of observations removed from both the upper and lower bounds.

For example, a figure skating competition produces the following scores: 6.0, 8.1, 8.3, 9.1, 9.9. A mean trimmed 40% would equal 8.5 ((8.1+8.3+9.1)/3), which is larger than the arithmetic mean of 8.28. To trim the mean by 40%, we remove the lowest 20% and highest 20% of values, eliminating the scores of 6.0 and 9.1. As shown by this example, trimming the mean can reduce the effects of outlier bias in a sample.

The winsorized mean

The winsorized mean is less sensitive to outliers because it replaces them with less influential values. This method of averaging is similar to the trimmed mean; however, instead of eliminating data, observations are altered, allowing for a degree of influence.

Let's calculate the first winsorized mean for the following data set: 1, 5, 7, 8, 9, 10, 14. Because the winsorized mean is in the first order, we replace the smallest and largest values with their nearest observations. The data set now appears as follows: 5, 5, 7, 8, 9, 10, 10. Taking an arithmetic average of the new set produces a winsorized mean of 7.71 ((5+5+7+8+9+10+10) / 7).

1.1.2 Measures of Dispersion

The most common measures of dispersion are the variance and standard deviation. The range is also quite useful.

Variance is the major measure of variability for a data set. To calculate the variance, all data values, their mean, and the number of data values are required. It is expressed in the squared unit of measurement. Its square root is the **standard deviation**. It is symbolized by σ^2 for a population and s^2 for a sample

1.2 Other measures

1.2.1 Skewness and Kurtosis

A fundamental task in many statistical analyses is to characterize the location and variability of a data set. A further characterization of the data includes skewness and kurtosis. **Skewness** is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Data sets with low kurtosis tend to have a flat top near the mean rather than a sharp peak. A uniform distribution would be the extreme case.

The histogram is an effective graphical technique for showing both the skewness and kurtosis of data set.

1.3 Simpson's paradox

1.3.1 Example

- Say a company tests two treatments for an illness. In trial No. 1, treatment A cures 20% of its cases (40 out of 200) and treatment B cures 15% of its cases (30 out of 200). In trial No. 2, treatment A cures 85% of its cases (85 out of 100) and treatment B cures 75% of its cases (300 out of 400)....
- So, in two trials, treatment A scored 20% and 85%. Also in two trials, treatment B scored only 15% and 75%. No matter how many people were in those trials, treatment A (at 20% and 85%) is surely better than treatment B (at 15% and 75%)?
- No, Treatment B performed better. It cured 330 (300+30) out of the 600 cases.
- (200+400) in which it was tried–a success rate of 55%. By contrast, treatment A cured 125 (40+85) out of the 300 cases (200+100) in which it was tried, a success rate of only about 42%.

1.4 The Ecological Fallacy

Ecological fallacy: The aggregation bias, which is the unfortunate consequence of making inferences for individuals from aggregate data. It results from thinking that relationships observed for groups necessarily hold for individuals. The problem is that it is not valid to apply group statistics to an individual member of the same group.

Chapter 2

Inference Procedures

2.1 Inference Procedures

Type I error: If the null hypothesis is true but we reject it this is an error of first kind or type I error (also called a error). This results in a false positive finding.
Type II error: If the null hypothesis is accepted when it is in fact wrong, this is an error of the second kind or type II error (also called b error). This results in a false negative result.

2.1.1 Types of Error

Type I error: If the null hypothesis is true but we reject it this is an error of first kind or type I error (also called a error). This results in a false positive finding.

Type II error: If the null hypothesis is accepted when it is in fact wrong, this is an error of the second kind or type II error (also called b error). This results in a false negative result.

2.1.2 Hypotheses

Null model: A model in which all parameters except the intercept are 0. It is also called the intercept-only model. The null model in linear regression is that the slope is 0, so that the predicted value of Y is the mean of Y for all values of X. The F test for the linear regression tests whether the slope is significantly different from 0, which is equivalent to testing whether the fit using non-zero slope is significantly better than the null model with 0 slope.

Alternative hypothesis: In practice, this is the hypothesis that is being tested in an experiment. It is the conclusion that is reached when a null hypothesis is rejected. It is the opposite of null hypothesis, which states that there is a difference between the groups or something to that effect.

- a single proportion,
- a single mean,
- the difference between two proportions
- the difference between two means;

Some commonly used tests Hypothesis test for the mean of a single sample Hypothesis test for the mean of two independent samples Hypothesis test for the proportion of a single group Hypothesis test for the proportions of two independent samples

2.1.3 Hypothesis test for the mean of a single sample

This procedure is used to assess whether the population mean has a specified value, based on the sample mean. The hypotheses are conventionally written in a form similar to below (here the hypothesized population mean is zero).

There are two hypothesis test for the mean of a single sample.

1) The sample is of a normally-distributed variable for which the population standard deviation (s) is known. 2) The sample is of a normally-distributed variable where s is estimated by the sample standard deviation (s).

In practice, the population standard deviation is rarely known. For this reason, we will consider the second case only in this course.In most statistical packages, this analysis is performed in the summary statistics functions.

2.1.4 Hypothesis test for the means of two independent samples.

The procedure associated with testing a hypothesis concerning the difference between two population means is similar to that for testing a hypothesis concerning the value of one population mean. The procedure differs only in that the standard error of the difference between the means is used to determine the test statistic associated with the sample result. For two tailed tests, the null hypothesis states that the population means are the same, with the alternative stating that the population means are not equal.

2.1.5 Hypothesis test of proportion

This procedure is used to assess whether an assumed proportion is supported by evidence. For two tailed tests, the null hypothesis states that the population proportion p has a specified value, with the alternative stating that p has a different value.

The hypotheses are typically as follows:

Example

A manufacturer is interested in whether people can tell the difference between a new formulation of a soft drink and the original formulation. The new formulation is cheaper to produce so if people cannot tell the difference, the new formulation will be manufactured. A sample of 100 people is taken. Each person is given a taste of both formulations and asked to identify the original. Sixty-two percent of the subjects correctly identified the new formulation. Is this proportion significantly different from 50

The first step in hypothesis testing is to specify the null hypothesis and an alternative hypothesis. In testing proportions, the null hypothesis is that p, the proportion in the population, is equal to 0.5. The alternate hypothesis is p not equal to 0.5.

The computed p-values is compared to the pre-specified significance level of 5%. Since the p-value (0.0214) is less than the significance level of 0.05, the effect is statistically significant.

Since the effect is significant, the null hypothesis is rejected. It is concluded that the proportion of people choosing the original formulation is greater than 0.50.

This result might be described in a report as follows:

The proportion of subjects choosing the original formulation (0.62) was significantly greater than 0.50, with p-value = 0.021. Apparently at least some people are able to distinguish between the original formulation and the new formulation. Tests of Differences between Proportions This procedure is used to compare two proportions from two different populations. For two tailed tests, the null hypothesis states that the population proportion has a specified value, with the alternative stating that .

2.1.6 Example

An experiment is conducted investigating the long-term effects of early childhood intervention programs (such as head start). In one (hypothetical) experiment, the high-school drop out rate of the experimental group (which attended the early childhood program) and the control group (which did not) were compared. In the experimental group, 73 of 85 students graduated from high school. In the control group, only 43 of 82 students graduated. Is this difference statistically significant?

The computed p-values is compared to the pre-specified significance level of 5%. Since the p-value (¡0.0001) is less than the significance level of 0.05, the effect is statistically significant.

Since the effect is significant, the null hypothesis is rejected. The conclusion is that the probability of graduating from high school is greater for students who have participated in the early childhood intervention program than for students who have not.

The results could be described in a report as:

> The proportion of students from the early-intervention group who graduated from high school was 0.86 whereas the proportion from the control group who graduated was only 0.52. The difference in proportions is significant, with p ¡ 0.0001.

2.2 One Way ANOVA

A One-Way Analysis of Variance is a way to test the equality of three or more means at one time by using variances.

2.2.1 Assumptions

- The populations from which the samples were obtained must be normally or approximately normally distributed.
- The samples must be independent.
- The variances of the populations must be equal.

2.2.2 Hypotheses

The null hypothesis will be that all population means are equal, the alternative hypothesis is that at least one mean is different.

Commonly lower case letters apply to the individual samples and capital letters apply to the entire set collectively. That is, n is one of many sample sizes, but N is the total sample size.

2.2.3 Decision Rule

The decision will be to reject the null hypothesis if the test statistic from the table is greater than the F critical value with $k - 1$ numerator and $N - k$ denominator degrees of freedom.

If the decision is to reject the null, then at least one of the means is different. However, the ANOVA does not tell you where the difference lies.

2.3 Confidence Intervals

2.4 Using the p-value for Hypothesis tests

Significance

The P value commonly, but mistakenly, understood to be the probability that the null hypothesis is correct. If it is a below a certain threshold value (like ¡0.05), the null hypothesis is rejected.

The P-value is the probability of having observed our data (or more extreme data) when the null hypothesis is true. The smaller the p-value, the less likely it is that the sample results come from a situation where the null hypothesis is true.

$P > 0.05$: no evidence against H_0 in favour of H_a.
$P < 0.05$: evidence against H_0 in favour of H_a.

Examples

With a p-value of zero to three decimal places, the model is statistically very significant.

2.5 Chi Square test for goodness of fit

The chi-squared test applied to contingency tables.

The Chi-squared test is the most commonly used test for frequency data and goodness-of-fit. In theory, it is nonparametric but because it has no parametric equivalent, it is not classified as such. It is not an exact test and with the current level of computing facilities, there is not much excuse not to use Fishers exact test for 2x2 contingency table analysis instead of Chi-squared test. Also for larger contingency tables, the G-test (log-likelihood ratio test) may be a better choice. The Chi-square value is obtained by summing up the values (residual2/fit) for each cell in a contingency. In this formula, residual is the difference between the observed value and its expected counterpart and fit is the expected value.

2.5.1 Yates's correction

The approximation of the Chi-square statistic in small 2×2 tables can be improved by reducing the absolute value of differences between expected and observed frequencies by 0.5 before squaring. This correction, which makes the estimation more conservative, is usually applied when the table contains only small observed frequencies (¡20).

The effect of this correction is to bring the distribution based on discontinuous frequencies nearer to the continuous Chi-squared distribution. This correction is best suited to the contingency tables with fixed marginal totals. Its use in other types of contingency tables (for independence and homogeneity) results in very conservative significance probabilities. This correction is no longer needed since exact tests are available.

Chapter 3

Regression and Correlation

3.1 Revision of Simple Linear Regression

The intecept estimate is denoted a, while the slope estimate is denoted b.

3.2 Scatterplots and Anscombe's quartet

Anscombes quartet is a fine example of this. The quartet is four sets of data that have the same sample statistics (mean, variance, correlation coefficient and regression equation), but when graphed, they are clearly very different.

3.3 Correlation

Pearson's correlation coefficient (r) is a measure of the strength of the 'linear' relationship between two quantitative variables. A major assumption is the normal distribution of variables. If this assumption is invalid (for example, due to outliers), the non-parametric equivalent Spearman's rank correlation should be used.

3.3.1 Formal test of Correlation

3.3.2 Lurking variables and Spurious Correlation

Spurious Correlations. Although you cannot prove causal relations based on correlation coefficients, you can still identify so-called spurious correlations; that is, correlations that are due mostly to the influences of "other" variables. For example, there is a correlation between the total amount of losses in a fire and the number of firemen that were putting out the fire; however, what this correlation does not indicate is that if you call fewer firemen then you would lower the losses. There is a third variable (the initial size of the fire) that influences both the amount of losses and the number of firemen. If you "control" for this variable (e.g., consider only fires of a fixed size), then the correlation will either disappear or perhaps even change its sign. The main problem with spurious correlations is that we typically do not know what the "hidden" agent is. However, in cases when we know where to look, we can use partial correlations that control for (partial out) the influence of specified variables.

3.3.3 Simpson's Paradox

3.3.4 Rank correlation

Spearman's Rank correlation coefficient

3.4 Multiple Linear Regression

Multiple regression: To quantify the relationship between several independent (predictor) variables and a dependent (response) variable. The coefficients $(a, b_1 tob_i)$ are estimated by the least squares method, which is equivalent to maximum likelihood estimation. A multiple regression model is built upon three major assumptions:

1. The response variable is normally distributed,
2. The residual variance does not vary for small and large fitted values (constant variance),
3. The observations (explanatory variables) are independent.

3.4.1 Estimates

3.5 Model building

The traditional approach to statistical model building is to find the most parsimonious model that still explains the data. The more variables included in a model (overfitting), the more likely it becomes mathematically unstable, the greater the estimated standard errors become, and the more dependent the model becomes on the observed data. Choosing the most adequate and minimal number of explanatory variables helps to find out the main sources of influence on the response variable, and increases the predictive ability of the model. Ideally, there should be more than 10 observations for each variable in the model.

The usual procedures used in variable selection in regression analysis are: univariate analysis of each variable (using C2 test), stepwise method (backward or forward elimination of variables; using the deviance difference), and best subsets selection. Once the essential main effects are chosen, interactions should be considered next. As in all model building situations in biostatistics, biological considerations should play a role in variable selection.

3.6 Overfitting

Overfitting occurs when a statistical model does not adequately describe of the underlying relationship between variables in a regression model. Overfitting generally occurs when the model is excessively complex, such as having too many parameters (i.e. predictor variables) relative to the number of observations. A model which has been overfit will generally have poor predictive performance, as it can exaggerate minor fluctuations in the data.

3.7 Multicollinearity

In multiple regression, two or more predictor variables are colinear if they show strong linear relationships. This makes estimation of regression coefficients impossible. It can also produce unexpectedly large estimated standard errors for the coefficients of the X variables involved.

This is why an exploratory analysis of the data should be first done to see if any collinearity among explanatory variables exists. Multicolinearity is suggested by non-significant results in individual tests on the regression coefficients for important explanatory (predictor) variables. Multicolinearity may make the determination of the main predictor variable having an effect on the outcome difficult.

3.7.1 How to Identify Multicollinearity

You can assess multicollinearity by examining **tolerance** and the **Variance Inflation Factor** (VIF) are two collinearity diagnostic factors that can help you identify multicollinearity. Tolerance is a measure of collinearity reported by most statistical programs such as SPSS; the variables tolerance is 1-R2. A small tolerance value indicates that the variable under consideration is almost a perfect linear combination of the independent variables already in the equation and that it should not be added to the regression equation. All variables involved in the linear relationship will have a small tolerance. Some suggest that a tolerance value less than 0.1 should be investigated further. If a low tolerance value is accompanied by large standard errors and nonsignificance, multicollinearity may be an issue.

3.7.2 The Variance Inflation Factor (VIF)

The Variance Inflation Factor (VIF) measures the impact of collinearity among the variables in a regression model. The Variance Inflation Factor (VIF) is 1/Tolerance, it is always greater than or equal to 1. There is no formal VIF value for determining presence of multicollinearity. Values of VIF that exceed 10 are often regarded as indicating multicollinearity, but in weaker models values above 2.5 may be a cause for concern. In many statistics programs, the results are shown both as an individual R2 value (distinct from the overall R2 of the model) and a Variance Inflation Factor (VIF). When those R2 and VIF values are high for any of the variables in your model, multicollinearity is probably an issue. When VIF is high there is high multicollinearity and instability of the b and beta coefficients. It is often difficult to sort this out.

You can also assess multicollinearity in regression in the following ways:

(1) Examine the correlations and associations (nominal variables) between independent variables to detect a high level of association. High bivariate correlations are easy to spot by running correlations among your variables. If high bivariate correlations are present, you can delete one of the two variables. However, this may not always be sufficient.

(2) Regression coefficients will change dramatically according to whether other variables are included or excluded from the model. Play around with this by adding and then removing variables from your regression model.

(3) The standard errors of the regression coefficients will be large if multicollinearity is an issue.

(4) Predictor variables with known, strong relationships to the outcome variable will not achieve statistical significance. In this case, neither may contribute significantly to the model after the other one is included. But together they contribute a lot. If you remove both variables from

the model, the fit would be much worse. So the overall model fits the data well, but neither X variable makes a significant contribution when it is added to your model last. When this happens, multicollinearity may be present.

3.7.3 Variance Inflation Factor

The variance inflation factor (VIF) quantifies the severity of multicollinearity in a regression analysis.

The VIF provides an index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of collinearity.

A common rule of thumb is that if the VIF is greater than 5 then multicollinearity is high. Also a VIF level of 10 has been proposed as a cut off value.

3.8 Law of Parsimony

Parsimonious: The simplest plausible model with the fewest possible number of variables.

3.9 Akaike Information Criterion

The Akaike information criterion is a measure of the relative goodness of fit of a statistical model. It was developed by Hirotsugu Akaike, under the name of "an information criterion" (AIC), and was first published by Akaike in 1974. AIC provides a means for comparison among modelsa tool for model

selection. AIC is good for prediction.

p is the number of free model parameters.
For AIC to be valid, n must be large compared to p.

3.10 The coefficient of determination

The coefficient of determination R^2 is used in the context of statistical models whose main purpose is the prediction of future outcomes on the basis of other related information. It is the proportion of variability in a data set that is accounted for by the statistical model.It provides a measure of how well future outcomes are likely to be predicted by the model.

In the case of simple linear regression, the coefficient of determination is equivalent to the squared value of the Pearson correlation coefficient.

R^2 is a statistic that will give some information about the goodness of fit of a model. In regression, the R2 coefficient of determination is a statistical measure of how well the regression line approximates the real data points. An R^2 of 1.0 indicates that the regression line perfectly fits the data.

3.10.1 The adjusted coefficient of determination

Adjusted R^2 (often written as and pronounced "R bar squared") is a modification of R2 that adjusts for the number of explanatory terms in a model. Unlike R^2, the adjusted R^2 increases only if the new term improves the model more than would be expected by chance. The adjusted R^2 can be negative, and will always be less than or equal to R2.

3.11 ANOVA

3.11.1 The F Distribution

F distribution: A continuous probability distribution of the ratio of two independent random variables, each having a Chi-squared distribution, divided by their respective degrees of freedom. The commonest use is to assign P values to mean square ratios (variance ratios) in ANOVA. In regression analysis, the F-test can be used to test the joint significance of all variables of a model.

For regression analyses, the degrees of freedom are as follows.

- $n - k$
- $k + 1$

3.12 Variance Selection Procedures

- Forward Selection
- Backward Elimination
- Stepwise Regression

A method in multiple regression studies aimed to find the best model. This method seeks a model that balances a relatively small number of variables with a good fit to the data by seeking a model with high R2a (the most parsimonious model with the highest percentage accounted for).

The stepwise regression can be started from a null or a full model and can go forward or backward, respectively. At any step in the procedure, the statistically most important variable will be the one that produces the greatest change in the log-likelihood relative to a model lacking the variable. This would be the variable, which would result in the largest likelihood ratio statistics, G (a high percentage accounted for gives an indication that the model fits well).

3.13 Leverage and Influence

- Outliers
- Leverage
- Influence

In our last chapter, we learned how to do ordinary linear regression with SPSS, concluding with methods for examining the distribution of variables to check for non-normally distributed variables as a first look at checking assumptions in regression. Without verifying that your data have met the regression assumptions, your results may be misleading. This chapter will explore how you can use SPSS to test whether your data meet the assumptions of linear regression. In particular, we will consider the following assumptions.

(a) Linearity - the relationships between the predictors and the outcome variable should be linear

(b) Normality - the errors should be normally distributed - technically normality is necessary only for the t-tests to be valid, estimation of the coefficients only requires that the errors be identically and independently distributed

(c) Homogeneity of variance (homoscedasticity) - the error variance should be constant

(d) Independence - the errors associated with one observation are not correlated with the errors of any other observation

(e) Model specification - the model should be properly specified (including all relevant variables, and excluding irrelevant variables)

Additionally, there are issues that can arise during the analysis that, while strictly speaking are not assumptions of regression, are none the less, of great concern to regression analysts.

(f) Influence - individual observations that exert undue influence on the coefficients

(g) Collinearity - predictors that are highly collinear, i.e. linearly related, can cause problems in estimating the regression coefficients.

Many graphical methods and numerical tests have been developed over the years for regression diagnostics and SPSS makes many of these methods easy to access and use. In this chapter, we will explore these methods and show how to verify regression assumptions and detect potential problems using SPSS.

Regression diagnostics Tests to identify the main problem areas in regression analysis: normality, common variance and independence of the error terms; outliers, influential data points, collinearity, independent variables being subject to error, and inadequate specification of the functional form of the model. The purpose of the diagnostic techniques is to identify weaknesses in the regression model or the data. Remedial measures, correction of errors in the data, elimination of **true** outliers, collection of better data, or improvement of the model, will allow greater confidence in the final product.

Outlier An extreme observations that is well separated from the remainder of the data. In regression analysis, not all outlying values will have an influence on the fitted function. Those outlying with regard to their X values (high leverage), and those with Y values that are not consistent with the regression relation for the other values (high residual) are expected to be influential. The test the influence of such values, the Cook statistics is used.

Outlier or Unusual observation??

Influential points: Observations that actually dominate a regression analysis (due to high leverage, high residuals or their combination). The method of ordinary least squares gives equal weight to every observation. However, every observation does not have equal impact on the least squares results. The slope, for example, is influenced most by the observations having values of the independent variable farthest from the mean. An observation is influential if deleting it from the dataset would lead to a substantial change in the fit of the generalized linear model.

High-leverage points have the potential to dominate a regression analysis but not necessarily exert an influence (i.e., a point may have high leverage but low influence as measured by Cook statistics).

Cook statistics is used to determine the influence of a data point on the model.

Leverage points In regression analysis, these are the observations that have an extreme value on one or more explanatory variable. The leverage values indicate whether or not X values for a given observation are outlying (far from the main body of the data). A high leverage value indicates that the particular observation is distant from the centre of the X observations.

High-leverage points have the potential to dominate a regression analysis but not necessarily influential. If the residual of the same data point and Cook's distance are also high, then it is an influential point.

Cook statistics: A diagnostic influence statistics in regression analysis designed to show the influential observations. **Cook's distance** considers the influence of the ith value on all n fitted values and not on the fitted value of the ith observation. It yields the shift in the estimated parameter from fitting a regression model when a particular observation is omitted. All distances should be roughly equal; if not, then there is reason to believe that the respective case(s) biased the estimation of the regression coefficients.

Relatively large Cook statistics (or Cook's distance) indicates influential observations. This may be due to a high leverage, a large residual or their combination. An index plot of residuals may reveal the reason for it. The leverages depend only on the values of the explanatory variables (and the model parameters). Cook statistics depends on the residuals as well. Cook statistics may not be very satisfactory in binary regression models. Its formula uses the standardized residuals but the modified Cook statistics uses the deletion residuals.

Half-normal plot A diagnostic test for model inadequacy or revealing the presence of outliers. It compares the ordered residuals from the data to the expected values of ordered observations from a normal distribution. While the full-normal plots use the signed residuals, half-normal plots use the

absolute values of the residuals. Outliers appear at the top right of the plot as distinct points, and departures from a straight line mean that the model is not satisfactory. It is appropriate to use a half-normal plot only when the distribution is symmetrical about zero because any information on symmetry will be lost.

Normal probability plot of the residuals: A diagnostic test for the assumption of normal distribution of residuals in linear regression models. Each residual is plotted against its expected value under normality. A plot that is nearly linear suggests normal distribution of the residuals. A plot that obviously departs from linearity suggests that the error distribution is not normal.

3.13.1 Heteroskedasticity

Another assumption of ordinary least squares regression is that the variance of the residuals is homogeneous across levels of the predicted values, also known as homoskedasticity. If the model is well-fitted, there should be no pattern to the residuals plotted against the fitted values. If the variance of the residuals is non-constant then the residual variance is said to be "heteroskedastic." Below we illustrate graphical methods for detecting heteroskedasticity. A commonly used graphical method is to use the residual versus fitted plot to show the residuals versus fitted (predicted) values.

Below we use the /scatterplot subcommand to plot *zresid (standardized residuals) by *pred (the predicted values). We see that the pattern of the data points is getting a little narrower towards the right end, an indication of mild heteroscedasticity.

```
regression
  /dependent api00
  /method=enter meals ell emer
  /scatterplot(*zresid *pred).
```

Chapter 4

Data Quality

4.1 Data Scrubbing

Data scrubbing, sometimes called data cleansing, is the process of detecting and removing or correcting any information in a database that has some sort of error. This error can be because the data is wrong, incomplete, formatted incorrectly, or is a duplicate copy of another entry. Many data-intensive fields of business such as banking, insurance, retail, transportation, and telecommunications may use these sophisticated software applications to clean up a database's information.

Errors are in databases can be the result of human error in entering the data, the merging of two databases, a lack of company wide or industry wide data coding standards, or due to old systems that contain inaccurate or outdated data. Before computers had the capabilities to sort through and clean data, most data scrubbing was done by hand. Not only was this time consuming and expensive, but it oftentimes led to even more human error.

The need for data scrubbing is made clear when considering how easily errors can be made. For example, consider a database of names and addresses. One name is Bobby Johnson of Needham, MA. Another name is Bob Johnson of Needham, MA. This variation of names is most likely an error, and is referring to one person. However, a computer would normally deal with the information as though it were two different people. Specialized data scrubbing software is able to distinguish the discrepancy and fix it.

While these small errors may seem like a trivial problem, when merging corrupt or erroneous data into multiple databases, the problem may be multiplied by the millions. This so-called "dirty data" has been a problem as long as there have been computers, but the problem is becoming more critical as businesses are becoming more complex and data warehouses are merging data from multiple sources. There is no point in having a comprehensive database if that database is filled with errors and disputed information.

Companies using specialized data scrubbing software can either develop it in-house or buy it from a variety of vendors. The software is not cheap and can range anywhere from a price of $20,000 to 300,000$. It oftentimes also requires some customization so that the software will work to the business' specific needs. The software goes through a process of using algorithms to standardize, correct, match, and consolidate data and is able to work with single or multiple sets of data.

Data scrubbing is sometimes skipped as part of a Data Warehouse implementation but it is one of the most critical steps to having a good, accurate end product. Because mistakes will always be made in data entry, the need for data scrubbing will always be present.

4.2 Censored Data

4.3 Missing Data

4.3.1 Missing completely at random

There are several reasons why the data may be missing. They may be missing because equipment malfunctioned, the weather was terrible, or people got sick, or the data were not entered correctly. Here the data are missing completely at random (MCAR). When we say that data are missing completely at random, we mean that the probability that an observation (Xi) is missing is unrelated to the value of Xi or to the value of any other variables. Thus data on family income would not be considered MCAR if people with low incomes were less likely to report their family income than people with higher incomes. Similarly, if Whites were more likely to omit reporting income than African Americans, we again would not have data that were MCAR because "missingness" would be correlated with ethnicity. However if a participant's data were missing because he was stopped for a traffic violation and missed the data collection session, his data would presumably be missing completely at random. Another way to think of MCAR is to note that in that case any piece of data is just as likely to be missing as any other piece of data.

Notice that it is the value of the observation, and not its "missingness," that is important. If people who refused to report personal income were also likely to refuse to report family income, the data could still be considered MCAR, so long as neither of these had any relation to the income value itself. This is an important consideration, because when a data set consists of responses to several survey instruments, someone who did not complete the Beck Depression Inventory would be missing all BDI subscores, but that would not affect whether the data can be classed as MCAR.

This nice feature of data that are MCAR is that the analysis remains unbiased. We may lose power for our design, but the estimated parameters are not biased by the absence of data.

4.3.2 Missing at random

Often data are not missing completely at random, but they may be classifiable as **missing at random** (MAR). For data to be missing completely at random, the probability that Xi is missing is unrelated to the value of Xi or other variables in the analysis. But the data can be considered as missing at random if the data meet the requirement that "missingness" does not depend on the value of X_i after controlling for another variable.

For example, people who are depressed might be less inclined to report their income, and thus reported income will be related to depression. Depressed people might also have a lower income in general, and thus when we have a high rate of missing data among depressed individuals, the existing mean income might be lower than it would be without missing data. However, if, within depressed patients the probability of reported income was unrelated to income level, then the data would be considered MAR, though not MCAR.

The phraseology is a bit awkward here because we tend to think of randomness as not producing bias, and thus might well think that Missing at Random is not a problem. Unfortunately is is a problem, although in this case we have ways of dealing with the issue so as to produce meaningful and relatively unbiased estimates. But just because a variable is MAR does not mean that you can just forget about the problem.

4.3.3 Missing Not at random

If data are not missing at random or completely at random then they are classed as **Missing Not at Random** (MNAR). For example, if we are studying mental health and people who have been diagnosed as depressed are less likely than others to report their mental status, the data are not missing at random. Clearly the mean mental status score for the available data will not be an unbiased estimate of the mean that we would have obtained with complete data. The same thing happens when people with low income are less likely to report their income on a data collection form.

When we have data that are MNAR we have a problem. The only way to obtain an unbiased estimate of parameters is to model missingness. In other words we would need to write a model that accounts for the missing data. That model could then be incorporated into a more complex model for estimating missing values. This is not a task anyone would take on lightly. See Dunning and Freedman (2008) for an example.

Chapter 5

Introduction to databases and data analytics

5.1 Knowledge Discovery in Databases

Knowledge Discovery and Data Mining (KDD) is an interdisciplinary area focusing upon methodologies for extracting useful knowledge from data. The ongoing rapid growth of online data due to the Internet and the widespread use of databases have created an immense need for KDD methodologies. The challenge of extracting knowledge from data draws upon research in statistics, databases, pattern recognition, machine learning, data visualization, optimization, and high-performance computing, to deliver advanced business intelligence and web discovery solutions.

5.1.1 Data Rich, Information Poor

The amount of raw data stored in corporate databases is exploding. From trillions of point-of-sale transactions and credit card purchases to pixel-by-pixel images of galaxies, databases are now measured in gigabytes and terabytes. (One terabyte = one trillion bytes. A terabyte is equivalent to about 2 million books).

For instance, every day, Wal-Mart uploads 20 million point-of-sale transactions to an A&T massively parallel system with 483 processors running a centralized database. Raw data by itself, however, does not provide much information. In today's fiercely competitive business environment, companies need to rapidly turn these terabytes of raw data into significant insights into their customers and markets to guide their marketing, investment, and management strategies.

5.1.2 Data Warehouses

The drop in price of data storage has given companies willing to make the investment a tremendous resource: Data about their customers and potential customers stored in "Data Warehouses." Data warehouses are becoming part of the technology. Data warehouses are used to consolidate data located in disparate databases. A data warehouse stores large quantities of data by specific categories so it can be more easily retrieved, interpreted, and sorted by users. Warehouses enable executives and managers to work with vast stores of transactional or other data to respond faster to markets and make more informed business decisions. It has been predicted that every business will have a data warehouse within ten years. But merely storing data in a data warehouse does a company little good.

Companies will want to learn more about that data to improve knowledge of customers and markets. The company benefits when meaningful trends and patterns are extracted from the data.

5.1.3 What is Data Mining?

Data mining, or knowledge discovery, is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Data mining derives its name from the similarities between searching for valuable information in a large database and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find where the value resides.

5.1.4 What Can Data Mining Do?

Although data mining is still in its infancy, companies in a wide range of industries - including retail, finance, heath care, manufacturing transportation, and aerospace - are already using data mining tools and techniques to take advantage of historical data. By using pattern recognition technologies and statistical and mathematical techniques to sift through warehoused information, data mining helps analysts recognize significant facts, relationships, trends, patterns, exceptions and anomalies that might otherwise go unnoticed.

For businesses, data mining is used to discover patterns and relationships in the data in order to help make better business decisions. Data mining can help spot sales trends, develop smarter marketing campaigns, and accurately predict customer loyalty. Specific uses of data mining include:

- Market segmentation - Identify the common characteristics of customers who buy the same products from your company.
- Customer churn - Predict which customers are likely to leave your company and go to a competitor.
- Fraud detection - Identify which transactions are most likely to be fraudulent.
- Direct marketing - Identify which prospects should be included in a mailing list to obtain the highest response rate.
- Interactive marketing - Predict what each individual accessing a Web site is most likely interested in seeing.
- Market basket analysis - Understand what products or services are commonly purchased together; e.g., beer and diapers.
- Trend analysis - Reveal the difference between a typical customer this month and last.

Data mining technology can generate new business opportunities by:

Automated prediction of trends and behaviors: Data mining automates the process of finding predictive information in a large database. Questions that traditionally required extensive hands-on analysis can now be directly answered from the data. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets

most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

Automated discovery of previously unknown patterns: Data mining tools sweep through databases and identify previously hidden patterns. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

Using massively parallel computers, companies dig through volumes of data to discover patterns about their customers and products. For example, grocery chains have found that when men go to a supermarket to buy diapers, they sometimes walk out with a six-pack of beer as well. Using that information, it's possible to lay out a store so that these items are closer.

AT&T, A.C. Nielson, and American Express are among the growing ranks of companies implementing data mining techniques for sales and marketing. These systems are crunching through terabytes of point-of-sale data to aid analysts in understanding consumer behavior and promotional strategies. Why? To gain a competitive advantage and increase profitability!

Similarly, financial analysts are plowing through vast sets of financial records, data feeds, and other information sources in order to make investment decisions. Health-care organizations are examining medical records to understand trends of the past so they can reduce costs in the future.

5.1.5 The Evolution of Data Mining

Data mining is a natural development of the increased use of computerized databases to store data and provide answers to business analysts.

Evolutionary Step Business Question Enabling Technology

Data Collection (1960s) "What was my total revenue in the last five years?" computers, tapes, disks

Data Access (1980s) "What were unit sales in New England last March?" faster and cheaper computers with more storage, relational databases

Data Warehousing and Decision Support "What were unit sales in New England last March? Drill down to Boston." faster and cheaper computers with more storage, On-line analytical processing (OLAP), multidimensional databases, data warehouses

Data Mining "What's likely to happen to Boston unit sales next month? Why?" faster and cheaper computers with more storage, advanced computer algorithms

Traditional query and report tools have been used to describe and extract what is in a database. The user forms a hypothesis about a relationship and verifies it or discounts it with a series of queries against the data. For example, an analyst might hypothesize that people with low income and high debt are bad credit risks and query the database to verify or disprove this assumption. Data mining can be used to generate an hypothesis. For example, an analyst might use a neural net to discover a pattern that analysts did not think to try - for example, that people over 30 years old with low incomes and high debt but who own their own homes and have children are good credit risks.

5.1.6 How Data Mining Works

How is data mining able to tell you important things that you didn't know or what is going to happen next? That technique that is used to perform these feats is called modeling. Modeling is simply the act of building a model (a set of examples or a mathematical relationship) based on data from situations where the answer is known and then applying the model to other situations where the answers aren't

known. Modeling techniques have been around for centuries, of course, but it is only recently that data storage and communication capabilities required to collect and store huge amounts of data, and the computational power to automate modeling techniques to work directly on the data, have been available.

As a simple example of building a model, consider the director of marketing for a telecommunications company. He would like to focus his marketing and sales efforts on segments of the population most likely to become big users of long distance services. He knows a lot about his customers, but it is impossible to discern the common characteristics of his best customers because there are so many variables. From his existing database of customers, which contains information such as age, sex, credit history, income, zip code, occupation, etc., he can use data mining tools, such as neural networks, to identify the characteristics of those customers who make lots of long distance calls. For instance, he might learn that his best customers are unmarried females between the age of 34 and 42 who make in excess of 60,000 per year. This, then, is his model for high value customers, and he would budget his marketing efforts to accordingly.

5.1.7 Data Mining Technologies

The analytical techniques used in data mining are often well-known mathematical algorithms and techniques. What is new is the application of those techniques to general business problems made possible by the increased availability of data and inexpensive storage and processing power. Also, the use of graphical interfaces has led to tools becoming available that business experts can easily use.

Some of the tools used for data mining are:

- Artificial neural networks - Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- Decision trees - Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset.
- Rule induction - The extraction of useful if-then rules from data based on statistical significance.
- Genetic algorithms - Optimization techniques based on the concepts of genetic combination, mutation, and natural selection.
- Nearest neighbor - A classification technique that classifies each record based on the records most similar to it in an historical database.

5.1.8 Real-World Examples

Details about who calls whom, how long they are on the phone, and whether a line is used for fax as well as voice can be invaluable in targeting sales of services and equipment to specific customers. But these tidbits are buried in masses of numbers in the database. By delving into its extensive customer-call database to manage its communications network, a regional telephone company identified new types of unmet customer needs. Using its data mining system, it discovered how to pinpoint prospects for additional services by measuring daily household usage for selected periods. For example, households that make many lengthy calls between 3 p.m. and 6 p.m. are likely to include teenagers who are prime candidates for their own phones and lines. When the company used target marketing that emphasized convenience and value for adults - "Is the phone always tied up?" - hidden demand surfaced. Extensive telephone use between 9 a.m. and 5 p.m. characterized by patterns related to voice, fax, and modem usage suggests a customer has business activity. Target marketing offering those customers "business

communications capabilities for small budgets" resulted in sales of additional lines, functions, and equipment.

The ability to accurately gauge customer response to changes in business rules is a powerful competitive advantage. A bank searching for new ways to increase revenues from its credit card operations tested a nonintuitive possibility: Would credit card usage and interest earned increase significantly if the bank halved its minimum required payment? With hundreds of gigabytes of data representing two years of average credit card balances, payment amounts, payment timeliness, credit limit usage, and other key parameters, the bank used a powerful data mining system to model the impact of the proposed policy change on specific customer categories, such as customers consistently near or at their credit limits who make timely minimum or small payments. The bank discovered that cutting minimum payment requirements for small, targeted customer categories could increase average balances and extend indebtedness periods, generating more than $25 million in additional interest earned,

Merck-Medco Managed Care is a mail-order business which sells drugs to the country's largest health care providers: Blue Cross and Blue Shield state organizations, large HMOs, U.S. corporations, state governments, etc. Merck-Medco is mining its one terabyte data warehouse to uncover hidden links between illnesses and known drug treatments, and spot trends that help pinpoint which drugs are the most effective for what types of patients. The results are more effective treatments that are also less costly. Merck-Medco's data mining project has helped customers save an average of 10-15% on prescription costs.

5.1.9 The Future of Data Mining

In the short-term, the results of data mining will be in profitable, if mundane, business related areas. Micro-marketing campaigns will explore new niches. Advertising will target potential customers with new precision.

In the medium term, data mining may be as common and easy to use as e-mail. We may use these tools to find the best airfare to New York, root out a phone number of a long-lost classmate, or find the best prices on lawn mowers.

The long-term prospects are truly exciting. Imagine intelligent agents turned loose on medical research data or on sub-atomic particle data. Computers may reveal new treatments for diseases or new insights into the nature of the universe. There are potential dangers, though, as discussed below.

5.1.10 Privacy Concerns

What if every telephone call you make, every credit card purchase you make, every flight you take, every visit to the doctor you make, every warranty card you send in, every employment application you fill out, every school record you have, your credit record, every web page you visit ... was all collected together? A lot would be known about you! This is an all-too-real possibility. Much of this kind of information is already stored in a database. Remember that phone interview you gave to a marketing company last week? Your replies went into a database. Remember that loan application you filled out? In a database. Too much information about too many people for anybody to make sense of? Not with data mining tools running on massively parallel processing computers! Would you feel comfortable about someone (or lots of someones) having access to all this data about you? And remember, all this data does not have to reside in one physical location; as the net grows, information of this type becomes more available to more people.

5.2 Cluster Anlysis

5.3 Data dredging

Data dredging (data fishing, data snooping) is the inappropriate (sometimes deliberately so) use of data mining to uncover misleading relationships in data. Data-snooping bias is a form of statistical bias that arises from this misuse of statistics. Any relationships found might appear to be valid within the test set but they would have no statistical significance in the wider population.

Data dredging and data-snooping bias can occur when researchers either do not form a hypothesis in advance or narrow the data used to reduce the probability of the sample refuting a specific hypothesis. Although data-snooping bias can occur in any field that uses data mining, it is of particular concern in finance and medical research, both of which make heavy use of data mining techniques.

5.4 Web-mining

More than ever, entities and individuals alike are using the World Wide Web to conduct a host of business and personal transactions. As a result, companies are increasingly employing Web data mining tools and techniques in order to find ways to improve their bottom lines and grow their customer base. Web data mining involves the process of collecting and summarizing data from a Web sites hyperlink structure, page content, or usage log in order to identify patterns. Using Web data mining, a company can identify a potential competitor, improve customer service, or target customer needs and expectations. A government agency may also seek to uncover terrorist threats or other criminal activities through the use of a Web data mining application.

Some common Web data mining techniques include Web content mining, Web usage mining, and Web structure mining. Web content mining examines the subject matter of a Web site. For example, Web content miners may analyze a site's audio, text, images, and video features. Web content miners typically focus on a sites textual information more than other site features. Natural language processing and information retrieval are two data mining techniques often used by Web content miners.

Web usage mining is usually an automated process whereby Web servers collect and report user access patterns in server access logs. A company may, for example, use a Web usage data mining tool to report on server access logs and user registration information in order to create a more effective Web site structure. Web structure mining studies the node and connection structure of Web sites. It can be useful in identifying similarities and relationships that exist among different Web sites. Web structure mining often involves uncovering patterns from hyperlinks or pulling out document structures on a Web page.

Two general data mining techniques that can be employed by Web data miners are data mining association analysis and data mining regression. Data mining association analysis helps uncover noteworthy relationships buried in large data sets. **Data mining regression** is a statistical technique whereby mathematical formulas are used to predict future results, such as profit margins, house values, or sales figures.

Data mining software vendors offer Web data mining tools that can pull out predictive information from large quantities of data. Businesses often use these software mining tools to analyze specific data sets regarding consumer behavior. Using the results of the data analysis, companies are able to forecast future business trends.

Data mining (DMM), also called Knowledge-Discovery in Databases (KDD) or Knowledge-Discovery and Data Mining, is the process of automatically searching large volumes of data for patterns using tools such as classification, association rule mining, clustering, etc. Data mining is a complex topic and has links with multiple core fields such as computer science and adds value to rich seminal computational techniques from statistics, information retrieval, machine learning and pattern recognition.

Data mining has been defined as *"the nontrivial extraction of implicit, previously unknown, and potentially useful information from data"* and *"the science of extracting useful information from large data sets or databases"* . It involves sorting through large amounts of data and picking out relevant information. It is usually used by businesses, intelligence organizations, and financial analysts, but is increasingly used in the sciences to extract information from the enormous data sets generated by modern experimental and observational methods. Metadata, or data about a given data set, are often expressed in a condensed data mine-able format, or one that facilitates the practice of data mining. Common examples include executive summaries and scientific abstracts.

Although data mining is a relatively new term, the technology is not. Companies for a long time have used powerful computers to sift through volumes of data such as supermarket scanner data, and produce market research reports. Continuous innovations in computer processing power, disk storage,

and statistical software are dramatically increasing the accuracy and usefulness of analysis. Data mining identifies trends within data that go beyond simple analysis. Through the use of sophisticated algorithms, users have the ability to identify key attributes of business processes and target opportunities. The term data mining is often used to apply to the two separate processes of knowledge discovery and prediction. Knowledge discovery provides explicit information that has a readable form and can be understood by a user.

Forecasting, or predictive modeling provides predictions of future events and may be transparent and readable in some approaches (e.g. rule based systems) and opaque in others such as neural networks. Moreover, some data mining systems such as neural networks are inherently geared towards prediction and pattern recognition, rather than knowledge discovery.

The term "data mining" is often used incorrectly to apply to a variety of other processes besides data mining. In many cases, applications may claim to perform "data mining" by automating the creation of charts or graphs with historic trends and analysis. Although this information may be useful and timesaving, it does not fit the traditional definition of data mining, as the application performs no analysis itself and has no understanding of the underlying data. Instead, it relies on templates or pre-defined macros (created either by programmers or users) to identify trends, patterns and differences. A key defining factor for true data mining is that the application itself is performing some real analysis. In almost all cases, this analysis is guided by some degree of user interaction, but it must provide the user some insights that are not readily apparent through simple slicing and dicing. Applications that are not to some degree self-guiding are performing data analysis, not data mining.

5.5 Predictive analytics

Predictive analytics encompasses a variety of techniques from statistics, data mining and game theory that analyze current and historical facts to make predictions about future events.

In business, predictive models exploit patterns found in historical and transactional data to identify risks and opportunities. Models capture relationships among many factors to allow assessment of risk or potential associated with a particular set of conditions, guiding decision making for candidate transactions.

Predictive analytics is used in financial services, insurance, telecommunications, retail, travel, healthcare, pharmaceuticals and other fields.

One of the most well-known applications is credit scoring, which is used throughout financial services. Scoring models process a customer's credit history, loan application, customer data, etc., in order to rank-order individuals by their likelihood of making future credit payments on time.

Chapter 6

Research

Very simply, the aim of research is to add something of value to an existing body of knowledge.

6.1 Theory of Research

Superficially the research process can appear to be relatively simple - if you carry out the basic steps methodically and carefully, then you should arrive at useful conclusions. However, the nature of research can be very complex and when you are reading textbooks on research methodology you will come across many unfamiliar words and terms. We first look at types of research and explain some of the terms.

The main different types of research can be classified by its purpose, its process and its outcome. These can in turn be broken down further:

The purpose of the research can be classified as:

- exploratory
- descriptive
- analytical
- predictive.

The process of the research can be classified as:

- Empirical
- Theoretical.

The outcome of the research can be classified as:

- applied
- basic or pure
- action.

6.1.1 Exploratory research

This is conducted when there are few or no earlier studies to which references can be made for information. The aim is to look for patterns, ideas or hypotheses rather than testing or confirming a hypothesis. In exploratory research the focus is on gaining insights and familiarity with the subject area for more rigorous investigation later. Descriptive research This describes phenomena as they exist. It is used to identify and obtain information on the characteristics of a particular issue. It may answer such questions as:

What is the absentee rate amongst a particular group of workers? What are the feelings of workers faced with redundancy?

The data collected are often quantitative, and statistical techniques are usually used to summarise the information. Descriptive research goes further than exploratory research in examining a problem since it is undertaken to ascertain and describe the characteristics of the issue. Analytical or explanatory research This is a continuation of descriptive research. The researcher goes beyond merely describing the characteristics, to analyse and explain why or how something is happening. Thus, analytical research aims to understand phenomena by discovering and measuring causal relations among them. It may answer questions such as:

How can the number of complaints made by customers be reduced? How can the absentee rate among employees be reduced? Why is the introduction of empowerment seen as a threat by departmental managers? Predictive research Predictive research goes further by forecasting the likelihood of a similar situation occurring elsewhere. It aims to generalise from the analysis by predicting certain phenomena on the basis of hypothesised, general relationships. It may attempt to answer questions such as:

Will the introduction of an employee bonus scheme lead to higher levels of productivity? What type of packaging will improve our products?

Predictive research provides how, why, and where answers to current events as well as to similar events in the future. It is also helpful in situations where What if? questions are being asked.

6.1.2 Process of research

There is no consensus about how to conceptualise the actual undertaking of research. There are, however, two main traditions of approaching a research topic Empirical (quantitative in nature) and Theoretical (qualitative in nature). Each approach demands different research methods.

6.1.3 Empirical research

The quantitative approach usually starts with a theory or a general statement proposing a general relationship between variables. With this approach it is likely that the researchers will take an objective position and their approach will be to treat phenomena as hard and real. They will favour methods such as surveys and experiments, and will attempt to test hypotheses or statements with a view to generalising from the particular. This approach typically concentrates on measuring or counting and involves collecting and analysing numerical data and applying statistical tests.

6.1.4 Theorerical research

The alternative tradition is the theoretical approach. Here the investigator views the phenomena to be investigated as more personal and softer. He or she will use methods such as personal accounts, unstructured interviews and participant observation to gain an understanding of the underlying reasons

and motivations for peoples attitudes, preferences or behaviours. With this approach, the emphasis is more on generating hypotheses from the data collection rather than testing a hypothesis.

In reading around the subject you will find many alternative names for qualitative and quantitative research. It is good to have an understanding of these and to recognise them when you see them in research methods textbooks.

The features and differences between the two research processes are detailed below.

You should note the following points:

Qualitative and quantitative research methods are not clear-cut nor mutually exclusive most research draws on both methods. Both approaches can generate quantitative and qualitative data. The difference between the two methods is in the overall form and in the emphasis and objectives of the study.

6.2 Outcome of research

6.2.1 Applied research

Applied research is problem-oriented as the research is carried out to solve a specific problem that requires a decision, for example, the improvement of safety in the workplace, or market research. For your dissertation it is not usually acceptable to carry out applied research as it is very much limited to one establishment or company and you are required to look at issues of wider significance, perhaps to your industry as a whole or to a sector of it. You may have already carried out a problem-based piece of research related to your placement. It is important to understand that the dissertation requires you to carry out some form of basic research.

Basic research is also called fundamental or pure research, and is conducted primarily to improve our understanding of general issues, without any emphasis on its immediate application. It is regarded as the most academic form of research since the principal aim is to make a contribution to knowledge, usually for the general good, rather than to solve a specific problem for one organisation. This may take the form of the following:

Discovery where a totally new idea or explanation emerges from empirical research which may revolutionise thinking on that particular topic. An example of this would be the Hawthorne experiments. (Gillespie 1991, note on Hawthorne experiments at end of paper.)

Invention where a new technique or method is created. An example of this would be the invention of TQM (total quality management).

Reflection where an existing theory, technique or group of ideas is re-examined possibly in a different organisational or social context.

6.2.2 Action research

This is a form of research where action is both an outcome and a part of the research. The researcher interferes with or changes deliberately what is being researched. The critics of action research argue that since the researcher is changing what is being researched during the process of research, the work cannot be replicated. If it cannot be replicated its findings cannot be tested in other situations. This prevents general knowledge being developed and thus it cannot contribute to theory. Also, as the researcher is involved in the change process there is a loss of critical, detached objectivity. There are two approaches to action research:

Classical action research begins with the idea that if you want to understand something you should try changing it.

New paradigm research is based on a new model or framework for research. It claims that research can never be neutral and that even the most static and conventional research exposes the need for change in what is being researched. It involves inquiry into persons and relations between persons, and is based on a close relationship between researcher and those being researched. The research is a mutual activity of a co-ownership involving shared power with respect to the process and the outcomes of the research. Those being researched can, for example, decide how the research will be undertaken, in what form and with what questions being asked. The researcher is a member of a community and brings to it special skills and expertise. The researcher does not dictate what will happen. This type of research is most easily carried out when working with individuals or small groups. It means that the researcher must be highly skilled not only in research methods but also in the interpersonal skills of facilitating others. It is not, therefore, usually appropriate for an undergraduate student who is carrying out a major piece of research for the first time. Action research is often used by educationalists who are trying to improve their own practice by making changes to the delivery of their classes and by observing and asking students which actions work best.

6.3 The Research Question

A Research Question is a statement that identifies the phenomenon to be studied.

For example, What resources are helpful to new data analysis researchers?

To develop a strong research question from your ideas, you should ask yourself these things:

Do I know the field and its literature well? What are the important research questions in my field? What areas need further exploration? Could my study fill a gap? Lead to greater understanding? Has a great deal of research already been conducted in this topic area? Has this study been done before? If so, is there room for improvement? Is the timing right for this question to be answered? Is it a hot topic, or is it becoming obsolete? Would funding sources be interested? If you are proposing a service program, is the target community interested? Most importantly, will my study have a significant impact on the field?

6.3.1 Business research

In general, business research refers to any type of researching done when starting or running any kind of business. For example, starting any type of business requires research into the target customer and the competition to create a business plan. Conducting business market research in existing businesses is helpful in keeping in touch with consumer demand. Small business research begins with researching an idea and a name and continues with research based on customer demand and other businesses offering similar products or services. All business research is done to learn information that could make the company more successful.

Business research methods vary depending on the size of the company and the type of information needed. For instance, customer research may involve finding out both a customers feelings about and experiences using a product or service.

The methods used to gauge customer satisfaction may be questionnaires, interviews or seminars. Researching public data can provide businesses with statistics on financial and educational information in regards to customer demographics and product usage, such as the hours of television viewed per week by people in a certain geographic area.

Business research used for advertising purposes is common because marketing dollars must be carefully spent to increase sales and brand recognition from ads.

Other than business market research and advertising research, researching is done to provide information for investors.

Business people aren't likely to invest in a company or organization without adequate research and statistics to show them that their investment is likely to pay off. Large or small business research can also help a company analyze its strengths and weaknesses by learning what customers are looking for in terms of products or services the business is offering. Then a company can use the business research information to adjust itself to better serve customers, gain over the competition and have a better chance of staying in business.

Most industries have trade journals that include research reports and statistics that relate to a certain type of business.

International information is especially important to businesses that have ties with other countries and need to understand more about the cultures and demographics of other nations. For example, International Business Research is a publication of the Canadian Center of Science and Education and includes business essays and academic editorials from businesspeople from different parts of the world such as Australia, India and Malaysia.

6.3.2 Applied Research

Every organizational entity engages in applied research. The basic definition for applied research is any fact gathering project that is conducted with an eye to acquiring and applying knowledge that will address a specific problem or meet a specific need within the scope of the entity. Just about any business entity or community organization can benefit from engaging in applied research. Here are a couple of examples of how applied research can help an organization grow.

When most people think of applied research, there is a tendency to link the term to the function of research and development (R and D) efforts. For business entities, R and D usually is involved with developing products that will appeal to a particular market sector and generate revenue for the company. The research portion of the R and D effort will focus on uncovering what needs are not being met within a targeted market and use that information to begin formulating products or services that will be attractive and desirable. This simplistic though systematized approach may also be applied to existing products as well, leading to the development of new and improved versions of currently popular offerings. Thus, applied research can open up new opportunities within an existing client base, as well as allow the cultivation of an entirely new sector of consumers.

Non-profit organizations also can utilize the principles of applied research. Most of these types of organizations have a specific goal in mind. This may be to attract more people to the organization, or to raise public awareness on a given issue, such as a disease. The concept of applied research in this scenario involves finding out what attracts people to a cause, and then developing strategies that will allow the non-profit entity to increase the public profile of the organization, and entice people to listen to what they have to say and offer.

Applied research can be very simplistic within a given application or it can become quite complicated. While the principle of applied research is easily grasped, not every organization contains persons who are competent in the process of engaging in applied research. Fortunately, there are a number of professionals who are able to step in and help any entity create a working model for applied research.

In some cases, this may be the most productive approach, since an outsider often notices information that may be easily overlooked by those who are part of the organization. Whether implemented as an internal effort or outsourced to professionals who routinely engage in applied research, the result is often a higher public profile for the organization, and improved opportunities for meeting the goals of the entity.

6.3.3 Research Techniques

There are two different types of research techniques: scientific and historical. The purpose of both techniques are to use a logical approach to obtain information about a specific subject. Research techniques can be applied to a broad range of issues or areas of research.

Basic research techniques are based on a formal process. The exact order of the steps depend on the subject and the reason for the research. The eight steps are the same for both basic and applied research.

The first four phases are: formation of a topic, hypothesis, conceptual definition and operational definition.

6.3.4 The formation phase

The formation of a topic is usually phrased as a question. The question is generally within the researchers field of expertise. The hypothesis is a theory proposed by the researcher, which is often phrased as a question. The conceptual and operational definitions provide the scope and focus for the research.

6.3.5 The Research process

The next four steps are: gathering data, analysis, testing and conclusion. The gathering of data, analysis and testing steps are the heart of all research. It is very important to use reliable sources, perform experiments, and test the hypothesis thoroughly. If the testing results do not support the hypothesis, the research is not a failure. On the contrary, these results provide an opportunity to revisit the hypothesis and new knowledge is gained.

Historical research techniques, or methods, are most commonly used to review data from the past and draw conclusions that impact on the present or future. Although commonly used by historians, these techniques are also used by scientific researchers. Using these techniques, they attempt to identify trends, and theorize on the causes of disease outbreaks and epidemics.

There are six steps in historical research. The first three are: define the starting date, locate independent verification of basic background information and investigate the author. These steps are necessary to confirm the evidence used is factual, reporting on by multiple sources and that the bias of the author.

The next three steps are to analyze the information, validate against other sources and measure the creditability of the information. These steps require the use of multiple sources and a process of questioning all aspects of the information. This includes using generally accepted knowledge about the time period in question, historical facts and physical evidence.

The process of historical research requires a significant amount of reading, translating, researching and discussion. The volume of information required to support a historical theory is quite substantial. This method is often used by professionals with an extensive background in a specific subject.