Financial Time Series I

Topic 3: Regression Analysis and Correlation

Hung Chen
Department of Mathematics
National Taiwan University

11/16/2002

OUTLINE

1. Association
   - Scatter Plot
   - Correlation Coefficient: linear association
   - Interpreting Association
   - Nonlinear Association
   - Causation
   - Distribution of Correlation Coefficient
2. Simple Linear Regression
   - Statistical Relationship
   - Model
   - Method of Least Squares
   - Properties of Least Squares Solution
   - correlation
3. Association of Categorical Variables
   - Frequency Tables
   - Bivariate Categorical Data
   - Test of Independence
   - Measure of Association

Association Between Numerical Scales

- Scientific question: Does studying more help raise scores on the SAT?
  - Data Collection: Record the number of hours spent studying for the SAT and the SAT scores for a sample of students.
  - Does this data lead to the conclusion that "individuals with higher *hours* also have higher *scores*?"
- Scientific question: Is sodium intake related to systolic and diastolic blood pressure?
  - Data Collection: Record the monthly sodium intakes for each individual in a sample and his/her blood pressure.
  - Do individuals with higher sodium consumption also have higher blood pressure readings?
- Question: How do we assess the association of two numerical variables in statistics?
  - scatter plot: a graphical technique Scatter plots frequently depict information about the relationship between vari-

ables that is not indicated by a single summary statistic.

- correlation coefficient: a formal numerical index of **linear** relationship.
  - How do we measure a *curvelinear* relationship?
- How reliable are these two tools?

Scatter plots

- Data: We have the language and nonlanguage mental maturity scores (two kinds of "IQs") of 23 school children.
  How do we explore it?
  - Let x denote language IQ and y denote non-language IQ where

    $$\begin{aligned} x \; <- \; & c(86, 104, 86, 105, 118, 96, 90, 95, 105, \\ & 84, 94, 119, 82, 80, 109, 111, 89, 99, \\ & 94, 99, 95, 102, 102) \\ y \; <- \; & c(44, 53, 42, 50, 65, 52, 37, 50, 46, 30, \\ & 37, 66, 41, 43, 74, 69, 44, 67, 43, 60, \\ & 47, 54, 43) \end{aligned}$$

  - These data are represented visually by making a graph on two axes, the horizontal x axis representing language IQ and the vertical y axis representing non-language IQ.
  - Such a graph is called a scatter plot (or scatter diagram).
  - Use R command, plot(x,y,xlab="language IQ",ylab ="non-languae IQ",main="Scatter

Plot")
Each point in such a plot represents one individual.

  - When all of the observations are plotted, the diagram conveys information about direction and magnitude of the association of x and y.
  - The swarm of points goes in a southwest-northeast direction.
    This indicates a positive or direct association of x and y.
    Namely, individuals who have the lower y values are the same people who have the lower values on x; they form a cluster of points in the lower-left portion of the diagram.
  - If the swarm of points lines in a northwest-southeast direction (i.e., upper left to lower right), there is a negative or inverse association of x with y.

- The strength or magnitude of the association is indicated by the degree to which the points are clustered together around a single line.

- If all of the points fall exactly on the line, there is a "perfect" association of the two variables. In this case, if we knew an individual's value on variable x, we would be able to compute his/her value on y exactly.
- To the extent that the points in the diagram diverge from a straight line, the association is less than perfect.
- Since the scatter plot is a nonnumerical way of assessing association, adjectives are used to describe the strength of association.
- We may say a "strong" (moderate or even weak) association of x with y.
- Are they objective?

The Correlation Coefficient: A Measure of Linear Relationship

- correlation coefficient: $r = \frac{s_{xy}}{s_x s_y}$
  - It is a statistic based on n pairs of measurements (x_i, y_i) on two variables x and y.
  - r is a measure of association between two quantitative variables.
  - $-1 \leq r \leq +1$.
  - The absolute value of r is exactly 1 for a perfect linear relationship, but lower if the points in the scatter plot diverge from a straight line.
- A descriptive statistic that indicates the degree of *linear* association of two numerical variates is the correlation coefficient, usually represented by the letter r.
  Do you see why?
- The sign of r indicates the direction of association- it is positive for a direct association of x and y and negative for an inverse association.
  We might ask

  - Whether the selling prices of various models of automobiles are related to the numbers of each that are sold in a given period of time?
  - Whether the amount of rainfall in a given agricultural area is related to the size of the crop yield?
  - The observational units may not be people, but institutions, objects, or events.

- For the above examples, it involves two numerical variables but questions about categorical variables are equally common which will be discussed later on.

- A correlation close to zero-either positive or negative-indicates little or no linear association between x and y.

- The question “what is a strong correlation?” has several answers.
  Statisticians would generally refer to a correlation close to zero as indicating “no correlation”; a correlation between 0 and 0.3 as “week”; a correlation between 0.3 and 0.6 as “moderate”; a correlation between 0.6 and

1.0 as “strong”; and a correlation of 1.0 as “perfect.”

- Two laboratory technicians counting impurities in the same water samples should have very high agreement-the correlation between their counts may be 0.95 or better.
- Two different human characteristics rarely have such a high correlation.

  The correlation of height and weight is generally in the neighborhood of 0.8; of scores on the Scholastic Assessment Test with college freshman grade average about 0.6; of measured intelligence with socioeconomic status about 0.4; of heart rate with blood pressure about 0.2.
- Example: consider the data collected by Nanji and French (1985) to examine the relationship of alcohol consumption with mortality due to cirrhosis of the liver in the 10 Canadian provinces. In this case, variable x is an index of the amount of alcohol consumed in the province in one year (1978) and variable y is the number

of individuals who died from cirrhosis of the liver per $100,000$ residents.
r is equal to 0.51.

Calculating Correlation Coefficient

- help.search("correlation")
- The software responds with "Help files with alias or title matching 'correlation', type 'help(FOO, package = PKG)' to inspect entry 'FOO(PKG) TITLE':
- We try the first two:
  - cor(base) Correlation, Variance and Covariance (Matrices)
  - corr(boot) Correlation Coefficient
  - acf(ts) Auto- and Cross- Covariance and -Correlation Function Estimation
  - plot.acf(ts) Plotting Autocovariance and Autocorrelation Functions
- Consider the example on language and non-language IQ scores.
- Use R base package, $cor(x, y)$ leads to 0.7689431.

- Use R boot package, $corr(cbind(x, y))$ leads to 0.7689431.
  - From Packages, load the package *boot*.
  - Look at help manual by command help(corr).
  - It only deals with matrix. We need to form a matrix with x and y.

    Test and Confidence Interval on r

- Efron (1982) analyzes data on law school admission, with the object being to examine the correlation between the LSAT (Law School Admission Test) score and the first-year GPA.
  - For each of 15 law schools, we have the pair of data points (acerage LSAT, average GPA):
    (576,3.39), (635, 3.30), (558, 2.81), (578,3.03), (666,3.44), (580,3.07), (555,3.00), (661,3.43), (651,3.36), (605,3.13), (653,3.12), (575, 2.74), (545,2.76), (572,2.88), (594,2.96)
  - Let (X, Y) be a bivariate normal with correlation coefficient ρ and sample cor-

relation r, it can be shown that

$$\sqrt{n}(r-\rho) \to N(0,(1-\rho^2)^2).$$

- The above result can be used to derive confidence interval and carry out hypothesis testing when the data is normally distributed.
- Fisher suggests to consider a z-transformation $\log((1+x)/(1-x))$. Then we have

$$\sqrt{n}\frac{1}{2}\left[\log\left(\frac{1+r}{1-r}\right) - \log\left(\frac{1+\rho}{1-\rho}\right)\right]) \to N(0,1).$$

Interpreting Association

- The correlation coefficient is useful for summarizing the direction and magnitude of association between two variables. It is a widely-used statistic, cited frequently in both scientific and popular reports.

- Limitations to the meaning of any particular correlation:

  - The correlation coefficient reveals only the straight line (linear) association between x and y.

  - r would be close to 0 while the scatter plot reveals important curvilinear patterns.

    In this case, the scatter plots reveal that a straight line is not the whole story of the relationship of x to y.

    This suggests that a scatter plot should always be examined when a correlation coefficient is to be computed or interpreted.

  - Nonlinear association: Pairs of variables are often associated in a clear pattern, but not conforming to a straight line.

Nonlinear Association

Quite common, we will see scatter plots with following characteristics.

- (asymptote) Consider the following two cases.
  - Suppose that patients who experience pain, the observations are variable x might be the amount of aspirin taken orally, in milligrams, and variable y the amount that is absorbed into the bloodstream.

    Beyond a certain point, additional amounts of ingested aspirin are no longer absorbed, and the amount found in the bloodstream reaches a plateau.
  - In research on human memory, it has been found that initial practice trials are very helpful in increasing the amount of material memorized; after a certain number of trials, however, each additional attempt to memorize has only a small added benefit.

    Consider students who are learning a foreign language, what is the relationship between the number of 25-minute periods devoted to studying vocabulary and the number of words memorized.

- (N shape) Examine the effects of advertising on sales.
  - If observations are branches of a large chain of stores with independent control over expenditures, variable x might be advertising outlays and y the total sales volume in a given period of time.
  - What will the plot look like?
  - If initial advertising outlays have a substantial effect on sales, additional advertising outlays have little additional impact on sales, but expensive "saturation" advertising again gives a significant boost to sales.
- (concave upward) Consider the amount of water provided to agricultural plots and the proportion of plants on the plot that do not grow to a given size.

  This pattern would suggest that too much as well as too little water is harmful, while moderate amounts of water minimize the plant loss.

- (concave downward) In examining the responses of humans and animals to various kinds of physiological stimulation, little or no stimulation produces little or no response, while moderate amounts of stimulation produce maximal response.
  Levels of stimulation that exceed the individual's ability to process the input, however, can result in a partial or complete suppression of t he response.

Causation

- Even a strong correlation between two variables does not imply that one causes the other.

  When a statistical analysis reveals association between two variables it is generally desirable to know more about the association.

  - Does it persist under different conditions?
  - Does one factor "cause" the other?
  - Is there a third factor that causes both?
  - Is there another link in the chain, a factor influenced by one variable and in turn influencing the other?

- The correlation indicates only that certain pairings of values on x and y occur more frequently than other combinations.

- Example: If the x and y variables were "years on the job" and "job satisfaction ratings" for a sample of employees, the positive correlation between them might indicate that

- if employees hold their jobs for more years, the work seems to become more satisfying, or
- if employees are more satisfied, they keep their jobs longer.

That is, the casual connection may go in either direction and may also be affected by other intermediary mechanisms.
It may be that

- more senior employees are given subtle or overt rewards that in turn enhance their satisfaction. Thus, the distribution of rewards-referred to as an intervening variable-explains the association of years with satisfaction; the correlation itself does not imply a direct cause-and-effect relationship.

- Association between two variables may also occur because x and y are both consequences of some third variable that has not been observed. This is seen in the following illustration:
- Example: Do Storks Bring Babies?

In Scandinavian countries a positive association between the number of storks living in the area and the number of babies born in the area was noticed.

Do storks bring the babies? Without shattering the illusions of the incurably romantic, we may suggest the following:

Districts with large populations have a large number of births and also have many buildings, on the chimneys of which storks can nest.

Consider the diagram representing the idea that the population factor explains both the number of births and the frequency with which storks are sighted.

$$\text{Large population}\begin{cases}\text{Many babies born}\\ \text{Many buildings} \rightarrow \text{Many storks}\end{cases}$$

The three variables to study are populations of districts, numbers of births in districts, and numbers of storks seen in the districts.

Simple Regression Analysis

- Find the statistical relationship between two quantitative variables.
- Statistical data are often used to answer questions about relationships between variables.
  - How do we summarize the relationship or association between 2 or among 3 or more variables?
  - How do we find the association among variables measured on numerical scales? How about categorical variables?
- Functional Relationship: A variable y is said to be a function of a variable x if to any value of x there corresponds one and only one value of y.
  - We symbolize a functional relationship by writing $y = f(x)$, where f represents the function.
  - The variable x is called the independent variable; the variable y is called the dependent variable because it is considered to depend on x.

  - If x is the height from which a ball is dropped and y is the time the ball takes to fall to the ground, then y is functionally related to x because the law of gravity determines y in terms of x.

- Statistical relationship: When one variable is used to predict or "explain" values of the second variable, we allow some imperfection in the prediction.

  - How do we express statistical relationships?
  - What kind of methods can be used for estimating those relationships from a sample of data?
  - How do we measure and interpret variability around the predicted or explained values?

Statistical Relationship

- The relationship between x and y is not an exact, mathematical relationship, but rather several y values corresponding to a given x value scatter about a value that depends on the x value.
- For example, although not all persons of the same height have exactly the same weight, their weights bear some relation to that height.
  - On the average, people who are 6 feet tall are heavier than those who are 5 feet tall; the mean weight in the population of 6-footers exceeds the mean weight in the population of 5-footers.
- The relationship between height and weight is modeled statistically as follows:
  - For every value of x there is a corresponding population of y values.
    Denote it by $F_Y(y|x)$ where $F(\cdot)$ is a distribution function.
  - The population mean of y for a particular value of x is denoted by $\mu(x)$. Note that $\mu(x) = E(Y|X = x)$.

  - As a function of x, $\mu(x)$ is called the regression function.
  - The population of y values at a particular x value also has a variance, denoted σ^2; the usual assumption is that the variance is the same for all values of x.
    homoscedastic: $Var(Y|X = x) = \sigma^2$
    heteroscedastic: $Var(Y|X = x)$ depends on x

- For many variables encountered in statistical research, the regression function is a linear function of x, and thus may be written as $\mu(x) = \alpha + \beta x$.
  - The quantities α and β are parameters that define the relationship between x and $\mu(x)$.
  - Write $Y = \alpha + \beta x + \epsilon$ with $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$.
  - In conducting a regression analysis, we use a sample of data, (x_i, y_i), to estimate the values of these parameters so that we can understand this relationship.

- The focus of regression analysis is on making inferences about α, β, and σ^2.

  - Estimate the magnitude of these parameters and test hypotheses about them.
  - Consider the hypothesis $H_0 : \beta = 0$.
    If this null hypothesis is true then $\mu(x) = \alpha + 0 \times x = \alpha$, the same number for all values of x.
    This means that the values of y do not depend on x, that is, there is no statistical relationship between x and y.
    If H_0 is rejected, then the existence of a statistical relationship between x and y is confirmed.

- The data required for regression analysis are observations on the pair of variables (x, y).
  - Variable x may be uncontrolled of "naturally occurring" as in the case of observing a sample of n individuals with their heights x (random design) and their weights y, or it may be controlled, as in an experiment in which persons are trained as data processors for different lengths of time x (fixed design), and one

measures the accuracy of their work y.

- Examples:
  - "Does studying more help raise scores on the Scholastic Assessment Tests (SAT)?" This question could also be worded as follows: If we recorded the number of hours spent studying for the SAT and the SAT scores for a sample of students, do individuals with higher "hours" also have higher SAT scores and individuals with lower hours have lower SAT scores?
  - We might ask if high sodium intake in one's diet is associated with elevated blood pressure.

    The question could be worded as follows: If we recorded the monthly sodium intake for each individual in a sample and his/her blood pressure, do individuals with higher sodium consumption also have higher blood pressure readings while those with lower sodium intakes have the lower blood pressure readings?
  - The term *regression* stems from the work

of Sir Francis Galton (1822-1911), a famous geneticist, who studied the sizes of seeds and their offspring and the heights of fathers and their sons.

In both cases, he found that

- The offspring of parents of larger than average size tended to be smaller than their parents.
- The offspring of parents of smaller than average size tended to be larger than their parents.
- He called this phenomenon "regression toward mediocrity."

Least-Square Estimates

- The data consist of n pairs of numbers $(x_1, y_1), (x_2, y_2), \ldots, ($
- To explore their association, they can be plotted as a scatter plot.
- How do we find a "best fit" line for two variables x and y?

  $$\begin{aligned} y_i &= \hat{\alpha} + \hat{\beta} x_i + r_i \\ \hat{y}_i &= \hat{\alpha} + \hat{\beta} x_i, \end{aligned}$$

  where $r_i = y_i - \hat{y}_i$ is called the residuals. Is r_i close to unobserved ϵ_i?
- In regression analysis we seek to determine the equation of that line that gives $\hat{y}_i$ values as close as possible to the data values y_i.
  - How do we determine $\hat{\alpha}$ and $\hat{\beta}$, estimates of α and β, respectively?
  - How do we obtain confidence intervals and to test hypotheses about the parameters of interest?
- If we all agree on choosing a line of best fit from all the lines that gives $\hat{y}_i$ values as close as possible to the data values y_i, this

question is equivalent to asking how we can choose an appropriate y intercept and slope, a and b, such that the deviation $y_i - \hat{y}_i$ as small as possible.

One approach is to find a and b to minimize the criterion

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

- The principle of least squares leads to

$$\begin{aligned} \bar{y} &= \hat{\alpha} + \hat{\beta}\bar{x} \\ \hat{\beta} &= \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}. \end{aligned}$$

Example 1. One of the questions about peaceful uses of atomic energy is the possibility that radioactive contamination poses health hazards.

- Since World War II, plutonium has been produced at the Hanford, Washington, facility of the Atomic Energy Commission.

- Over the years, appreciable quantities of radioactive wastes have leaked from their open-pit storage areas into the nearby Columbia River, which flows through parts of Oregon to the Pacific.

- To assess the consequences of this contamination on human health, investigators calculated, for each of the nine Oregon counties having frontage on either the Columbia River or the Pacific Ocean, an "index of exposure."
  - This index of exposure was based on several factors, including distance from Hanford and average distance of the population from water frontage.
  - The cancer mortality rate, cancer mortality per $100,000$ person-years (1959-1964), was also determined for each of these nine counties. They are Clatsop, Columbia, Gilliam, Hood River, Morrow, Portland, Sherman, Umatilla, and Wasco.
  - Data: $(8.34, 210.3)$, $(6.41, 177.9)$, $(3.41, 129.9)$, $(3.83, 1623)$, $(2.57, 130.1)$, $(11.64, 207.5)$, $(1.25, 113.5)$, $(2.49, 147.1)$, $(1.62, 137.5)$ (the index of exposure, cancer mortality rate)

Simple Regression Model

- A simple regression model for a set of pairs (X_i, Y_i), $1 \leq i \leq n$ of points with dependent or response variable $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ and explanatory variable $\mathbf{X} = (X_1, \ldots, X_n)$ is the following:

  $$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

  where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ is a vector of independent, identically distributed random variables with $\epsilon_i \sim N(0, \sigma^2)$, the normal distribution with mean zero and variance σ^2.

- The assumptions on the random variables ϵ_i are relaxed to just uncorrelated, i.e. $E[\epsilon_i \epsilon_j] = 0$ for $i \neq j$ but not necessarily independent.

- We can never separate the random variables ϵ_i from the observations Y_i since we don't know the value of ϵ_i, which means that we can never know the true values of (β_0, β_1).

- The method of least squares leads to choose a "best" $(\hat{\beta}_0, \hat{\beta}_1)$. This is where the sum of squared errors, or SSE of the model comes in. We measure the goodness of linear regression model by its squared error

$$SSE(b_0, b_1) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2.$$

The least squares regression model is the model with smallest least squares.

- The least squares solutions $(\hat{\beta}_0, \hat{\beta}_1)$ aren't equal to the true values of (β_0, β_1) but they do have the property that

$$E[\hat{\beta}_0] = \beta_0, \quad E[\hat{\beta}_1] = \beta_1.$$

That is, if the data do come from some simple linear regression model, the least squares solutions are unbiased estimates of the true values (β_0, β_1).
They also have the property that, among all unbiased estimates of (β_0, β_1) that are linear functions of the response variable, the least squares solutions have the smallest variance, this is what the classic Gauss-Markov Theorem states.

Properties of Least Squares Solution

- To summarize, we have made the following decomposition of each observation using the fitted and residual values
$$Y = \hat{Y} + r,$$
where $\hat{Y}$ is made from X and $corr(r, X) = 0$.

- Under the standard assumption that $Var(Y_i) = \sigma^2$ and $Cov(Y_i, Y_j) = 0$ where $i \neq j$, we have
$$\begin{aligned} Var(\hat{\beta}_0) &= \frac{\sigma^2 \Sigma_{i=1}^n x_i^2}{n \Sigma_{i=1}^n x_i^2 - \left(\Sigma_{i=1}^n x_i\right)^2}, \\ Var(\hat{\beta}_1) &= \frac{n\sigma^2}{n \Sigma_{i=1}^n x_i^2 - \left(\Sigma_{i=1}^n x_i\right)^2}, \\ Cov(\hat{\beta}_0, \hat{\beta}_1) &= \frac{-\sigma^2 \Sigma_{i=1}^n x_i}{n \Sigma_{i=1}^n x_i^2 - \left(\Sigma_{i=1}^n x_i\right)^2}. \end{aligned}$$

- Define the **residual sum of squares** (RSS) to be
$$RSS = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2,$$
where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least squares solutions.

Let $s^2 = RSS/(n-2)$ which is an unbiased estimate of σ^2.

- If the errors, ϵ_i are independent normal random variables, then the estimated slope and intercept, being linear combinations of independent random variables, are normally distributed as well.
  - Under the normality assumption, it can be shown that
    $$\frac{\hat{\beta}_i - \beta}{s_{\hat{\beta}_i}} \sim t_{n-2}.$$
    This result makes possible the construction of confidence intervals and Hypothesis tests.
  - If the ϵ_i are independent and the x_i satisfy certain assumptions, a version of the central limit theorem implies that, for large n, the estimated slope and intercept then the estimated slope and intercept are approximately normally distributed. The above t-distributions can be used.
- Since the absolute value of r cannot exceed

1, the squared correlation has possible values from 0 to 1.

- If $r = 0$, then x is of no help in predicting y.
- If $r = 1$ (and $r^2 = 1$), then x predicts y exactly.

  This can only occur if each point (x_i, y_i) falls exactly on the regression line; all of the variation in y can be explained by variability in x.
- In between these extremes, a weak correlation (e.g., in the range 0 to 0.3) is accompanied by a small proportion of explained variation (0 to 0.09), while a strong correlation (e.g., between 0.6 and 1.0) is accompanied by a substantially larger proportion of explained variation (0.36 to 1.0).
- Both r and r^2 are useful measures of association for regression analysis.

  The correlation tells the direction of association and its square tells the extent to which y is predictable from x.

  For the cancer mortality data, the correlation coefficient is $r = 900.13/... = 0.926$. This value is very high, and the proportion of variation in mortality attributable to radioactive exposure is also high, $0.9262 = 0.875$ (that is, 85%).

Data Analysis

- In *R*, “lm” is used to fit linear models.
  - It can be used to carry out regression, single stratum analysis of variance and analysis of covariance (although ‘aov’ may provide a more convenient interface for these).
  - Usage: lm(formula, data, subset, weights, na.action, method = “qr”, model = TRUE, x = FALSE, y = FALSE, qr = TRUE, singular.ok = TRUE, contrasts = NULL, offset = NULL,. . .)
  - Models for ‘lm’ are specified symbolically. A typical model has the form ‘response terms’ where ‘response’ is the (numeric) response vector and ‘terms’ is a series of terms which specifies a linear predictor for ‘response’.

    A terms specification of the form ‘first + second’ indicates all the terms in ‘first’ together with all the terms in ‘second’ with duplicates removed.

    A specification of the form ‘first:second’

indicates the set of terms obtained by taking the interactions of all terms in ‘first’ with all terms in ‘second’.
The specification ‘first*second’ indicates the cross of ‘first’ and ‘second’. This is the same as ‘first + second + first:second’.

– Additive model and multiplicative model

- Example: Consider Plant Weight Data in Annette Dobson (1990) “An Introduction to Generalized Linear Models.”

  – Data input:

$$
\begin{aligned}
\mathit{ctl} &< - \; c(4.17, 5.58, 5.18, 6.11, 4.50, 4.61, 5.17, 4.53\\
\mathit{trt} &< - \; c(4.81, 4.17, 4.41, 3.59, 5.87, 3.83, 6.03, 4.89\\
\text{group} &< - \; gl(2, 10, 20, labels = c(\text{``}Ctl\text{''}, \text{``}Trt\text{''}))\\
\text{weight} &< - \; c(\mathit{ctl}, \mathit{trt})
\end{aligned}
$$

  – Regress weight to the group.
  lm(weight $\sim$ group) leads to the estimate of β_0 to be 5.032 and the estimate of β_1 to be -0.371. Here β_1 refers to the group effect.

  – Carry out an ANOVA analysis on the proposed model.

anova(lm(weight $\sim$ group)) leads to the following table.

Analysis of Variance Table					
	Df	Sum Sq	Mean Sq	F value	$Pr(> F)$
group	1	0.6882	0.6882	1.4191	0.249
Residuals	18	8.7293	0.4850		

The Association Among Categorical Variables

- "Is inoculation with a polio vaccine related to the occurrence of paralytic polio?"
  - Data may be collected on a sample of observations and we ask whether a particular value on one variable (i.e., "vaccinated") co-occurs with a particular value on the other ("polio absent").
  - Both of these variables are simple yes/no dichotomies.
- How do we measure the association among categorical variables?
  - The idea of association is basically the same as in quantitative variables, that is, do certain values of one variable tend to occur more frequently with certain values of another?
  - The values of categorical variables, however, may not have a range from lower to higher quantities, but may represent qualitatively distinct conditions, such as a condition being "present" or "absent."

  - Summary statistics such as the mean and standard deviation are not applicable. Instead, observations are simply counted; for example, how many observations have both condition A and condition B present?
  - Counts are entered into a *frequency table.*

- As with numerical variables, an association of two categorical variables does not imply that one causes the other.
  - The direction of causation may be reciprocal, with each variable affecting the other.
  - Causation may be mediated by one or more intervening third variable(s).
  - Both variables may be consequences of additional variables not included in the data causing the two outcomes to occur together.
  - To understand these effects more completely, it is often necessary to examine the association of three or more variables. The tabulation of data for three categor-

ical variables results in a three-way for two variables is a two-way table.

Bivariate Categorical Data

- Bivariate categorical data result from the observation of two categorical variables for each individual.
- A variable with two categories, such as male/female or graduate/undergraduate, is called a dichotomous variable.
  A pair of dichotomous variables is called a double dichotomy.
  Because each variable has just two variables, the table that results is a two-by-two (2×2) frequency table.
- Double dichotomies arise, for example, when each of a number of persons is asked a pair of yes-no questions.
  - Company X ask each of 600 men whether or not they use Brand X razors and whether or not they use Brand X blades.
  - A physician may classify patients according to whether of not they have been inoculated against a disease and whether or not they contracted the disease.

  - Businesses of a certain type, for example, might be classified based on whether or not they provide “day-care” facilities and whether or not they provide maternity leave for pregnant employees.

- Association between the pair of variables is seen by examining the patterns of frequencies in the individual cells.

- Many statistical studies involve 2 × 2 tables and because many concepts of statistics can be presented in this form, we shall consider such tables in some detail.
  This discussion is simplified by the notation displayed in the following table.

		Question 2		
		Yes	No	Total
Question 1	Yes	a	b	$a+b$
	No	c	d	$c+d$
	Total	$a+c$	$b+d$	n

  Here

  - a is the number of persons answering “yes” to both questions.
  - b is the number of persons answering “yes”

to Question 1 and "no" to Question 2.

  - c is the number of persons answering "no" to Question 1 and"yes" to Question 2.
  - d is the number of persons answering "no" to both questions.
  - We denote by n the total number of persons included in the table and $n = a + b + c + d$.

- Independence:
  - When the joint frequencies for two variables have no association, they are said to be independent.
  - For example, suppose communities were cross-classified according to average income and crime rate as the following table.

		Average income level			
		Low	Medium	High	Total
Crime	Low	1(10%)	4(10%)	3(10%)	8(10%)
	Medium	7(70%)	28(70%)	21(70%)	56(70%)
rate	High	2(20%)	8(20%)	6(20%)	16(20%)
	All crime levels	10	40	30	80

- The table shows that for each income level most communities (70%) have a medium crime rate.
- The entire distribution of crime rates is the same for each level of average income; that is, crime rate is independent of level of average income.
- In this case, knowledge of a community's average income level provides no information about its crime rate.
- The percentages of districts with low, medium, and high crime rates are 10%, 70%, and 20% regardless of the average income.
- Note that the marginal distribution of crime rate has to be the same as the distribution for each income level.
- When will $\frac{a}{a+b} = \frac{c}{c+d}$?
- When will $\frac{a}{a+c} = \frac{b}{b+d}$?

- Index of association for 2 × 2 tables
  - A numerical indicator of association between two dichotomous variables can be based on the quantity $ad - bc$.

  - a and b are the upper-left and lower-right entries in the above table, and b and c are the upper-right and lower-left entries.
  - If $ad - bc = 0$, it indicates independence.
  - If $ad - bc > 0$, it indicates that A1 occurs more frequently with B1, and A2 with B2, than the other way around.
  - If $ad - bc < 0$, it indicates that A1 occurs more frequently with B2, and A2 with B1, than the other way around.
  - The difference $ad - bc$ is divided by a quantity that keeps the final index of association between -1 and $+1$, like the correlation coefficient for numerical scales.

- Measure of Association: ϕ coefficient

$$\phi = \frac{ad - bc}{\sqrt{(a + b)(c + d)(a + c)(b + d)}}$$

  - If the variables A and B are *ordered* then the magnitude of ϕ indicates the strength of association and the sign indicates the direction of association as well.
  - $(a + b + c + d)\phi$ is identical to the chi-square test we will discuss later on.

Other kinds of 2 × 2 tables

- Dichotomous variables sometimes have categories that are ordered, so that one value reflects more of some characteristic than other.
  - To study the association of income and education level, the income might be classified simply as above or below poverty level, education as having completed fewer than 12 years of schooling or 12 years or more.
- Another type double dichotomy is shown by the two-by-two tabulation of a dichotomous variable for the same individuals at two different times.
  - Suppose in August we asked a number of people which of two presidential candidates they favored and then asked the same people the same question in September.
  - We could tabulate the results as follows.

		September		
		Bush	Clinton	Total
August	Bush	a	b	$a+b$
	Clinton	c	d	$c+d$
	Total	$a+c$	$b+d$	n

Here the number c would be the number of persons who switched from Democratic candidate Clinton Republican candidate Bush between August and September 1992.

– These data are change-in-time data and the table is called a "turnover" table.

– We may ask whether the number of people who kept their original preference (cells a and d) is substantially larger than the number who changed (cell b and c).

Interpretation of frequencies

- Question: Is there a tendency for those who use Brand X razors also to use Brand X blades?

- Company X conducted a market survey, interviewing a sample of 600 men.
  Each man was asked whether he uses the Brand X razor and whether he uses Brand X blades.

	Use Brand X blades	Not use Brand X blades	Total
Use Brand X razor	186(67%)	93(33%)	279(100%)
Not Use	59(18%)	262(82%)	321(100%)
Total	245(41%)	355(59%)	600(100%)

- Association: Most (67%) of the men using Brand X razors also use Brand X blades; only a few (18%) of the men not using Brand X razors who use Brand X blades differs from that percentage among men not using Brand X razors.
  - We say there is an association between

using the Brand X razor and using Brand X blades.

- We can say that men who use the Brand X razor are more likely to use Brand X blades than are men who do not use the razor.
- This difference may suggest that an advertising campaign for blades should be directed to men not using Brand X razors.

Chi-Square Tests of Independence

Consider frequency tables in the case in which individuals were classified simultaneously on two categorical variables.

- Consider testing the null hypothesis that in a population two variables are independent on the basis of a sample drawn from that population.
  - Even though the variables are independent in the in the population, they may not be (and in fact probably will not be) independent in a sample.

Consider the following hypothetical data.

- At each of three different dates, a sample of 1000 registered voters was drawn; at each time each respondent was asked which of two potential candidates he or she would favor.

	10/91	1/92	Total
Bush	523	502	1025
Clinton	477	498	975
Total	1000	1000	2000

- In the underlying population, let the proportion favoring Bush be p_1 at the time of the first poll and p_2 at the time of the second poll.
- The null hypothesis that the proportion favoring one candidate does not depend on the date of polling is $H_0 : p_1 = p_2$.
- In this example the estimates of p_1 and p_2 are $\hat{p}_1 = 0.523$ and $\hat{p}_2 = o.502$, respectively, based on sample sizes $n_1 = n_2 = 1000$
- The estimate of the SD of the difference between two sample proportions when H_0 is true is
$$\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = 0.02235.$$
The test statistic is
$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}(1/n_1 + 1/n_2)}} = 0.940.$$
- What is the distribution of the above test statistic?

Measure of Association Based on Prediction

- Consider another measure of association which is based on the idea of using one variable to predict the other.
- Consider the cross-classification of exercise and health in the following Table.

	-3 Health status		
	Good	Poor	Total
Exerciser	92	14	106
Non-exerciser	25	71	96
Total	117	85	202

  - How well does exercise group predict health status?
  - If one of the 202 persons represented in this table is selected at random, our best guess of the health status-if we don't know anything about the person-is to say that the person is in the good-health group because more of the people are in that group (117, compared with 85 for the poor-health group); the good-health category is the mode.
  - If we make this prediction for each of the

202 persons, we shall be right in 117 cases and wrong in 85 cases.

- If we take the person's exercise level into account, we can improve our prediction.
  - If we know the person exercises regularly, our best guess is still that the person is in good-health group, for 92 of the regular exercisers are in the good-health group, compared with only 14 in the poor-health group.
  - If we know the person is not an exerciser, we should guess that the person is in the poor-health group; in this case we would be correct 71 times and incorrect 25 times.
  - Our total number of errors in predicting all 202 health conditions for both exercises and nonexercisers is $14 + 25 = 39$, compared with 85 errors if we do not use the exercise category in making the prediction.
- For this example, a coefficient of association that measures the improvement in predic-

tion of the column category due to using the row classification is

$$\lambda_{c\dot{r}} = \frac{85-(14+25)}{85} = 0.54.$$

Here $85-(14+25)$ is the *reduction in errors when using the rows to predict columns* and 85 is the *number of errors not using the rows.*

- The subscript $c \cdot r$ refers to predicting the column category using the row category.

Recall the correlation coefficient $r = s_{xy}/s_x s_y$ is also a measure of association between two quantitative variables.

- Although the measure is symmetric in the two variables, it can be interpreted in terms of how well one variable y can be predicted form the other variable x.
- Recall the definition of λ which defines a measure of association for categorical variables.
- With quantitative variables,

– The notion that replaces “number of errors” is “sum of squared deviations” of the predicted values of y from the actual values.

– *Number of errors in prediction of y without using x* is replaced with *sum of squares of deviations in predicting y without using x.* It is

$$\Sigma(y_i - \bar{y})^2.$$

– *Number of errors in prediction of y using x* is replaced with *sum of squares of deviations in predicting y using x.* It is

$$\Sigma(y_i - \hat{y}_i)^2.$$

– With the above interpretation, λ is equal to r^2.