Financial Time Series I

Topic 3: Analysis of Variance

Hung Chen

Department of Mathematics

National Taiwan University
11/25/2002

OUTLINE

1. One-Way Analysis of Variance

   – Introduction

   – Examples

   – Basic Set-up

   – Data Example

   – Algebra

   – Validality Conditions on F-test

2. Analysis of Variance of Ranks

3. Logistic Regression

4. Two-Way Analysis of Variance

   – Interaction

   – Data Example

   – Related R-command

   – Design

Comparison of Several Populations

- When some measurement, such as height or aptitude for a particular job, is made on several individuals, the values vary from person to person.

  - The variability of a quantitative scale is measured by its variance.
  - If the set of individuals is stratified into more homogeneous groups, the variance of the measurements within the more homogeneous group will be less than that of the measurements in the entire group; that is what "more homogeneous" means.

- As an example, consider the heights of pupils on an elementary school.

  - Fact 1: The variance of the heights of pupils on an elementary school is usually greater than the variance of heights of pupils in just the first grade, the variance in the second grade, and the variance in each of the other grades.
    The within-grade variances are less than the overall variance

- – Fact 2: The average height of pupils also varies from grade to grade.
  The averages vary between grades.
- – The total variability (of heights) is made up of two components: the variability of individuals within groups (grades) and the variability of means between groups (grades).
  Recall

  $$Var(Y) = EVar(Y|X) + Var(E(Y|X)).$$

  Y refers to the height of students and X refers to the grades.
- – At the extreme, all of the variability of a measured variable may be within the groups and none of it between groups, that is, the means of the subgroups are equal. Mathematically, it refers to the case that $E(Y|X) = E(Y)$.

- The analysis of variance is a set of statistical techniques for studying variability from different sources and comparing them to understand the relative importance of each of the sources.

– It is also used to make inferences about the population through tests of significance, including the very important comparison of the means of two or more separate populations.

– The analysis of variance is the most straightforward way to examine the association between a categorical variable ("groups") and a numerical variable (the measure on which the means are based)?

– In the above example, we try to find the association between height of student and grade of student.

One-Way Analysis of Variance

Suppose that we wish to compare the means of three populations on some measured dependent variable.

- The population means are represented as μ_1, μ_2, and μ_3.

- A hypothesis that may be tested is

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

  that is, that the means of the three populations are equal.

- The hypothesis is rejected if the population means are different in any way, for example, if one of the means differs from the other two or if all three means are different.

- The procedure for testing H_0 through the analysis of variance parallels that for other tests of significance.

- Samples of data are obtained and a test statistics is computed. The test statistic is proposed based on the following idea.

- The total variability of the data is allocated to two sources: variability among the group means and variability of the observations within the groups.

- Two measures of variability are computed, Mean Square between Groups and Mean Square within Groups, respectively.

- The hypothesis that the 3 population means are equal is tested by comparing the magnitude of these two mean squares.
It is the ratio of Mean Square Between Groups to Mean Square Within Groups.

- We need a reference, the F-distribution, to provide significance points for deciding if the ratio is large enough to reject the hypothesis of equal means.

- In this application the test statistic reflects the extent to which variation among the sample means is greater than variation among observations within the groups.

- If the test statistic is large enough (that is, if it exceeds the corresponding significance point) then H_0 is rejected.

- If the sample means are close together and hence the test statistic is small, then H_0 is accepted.

- Can we apply this procedure for comparing the means of two populations?

- When only two groups are being compared, it is obvious from knowledge of the sample means which group has a significantly higher mean than the other.

  - When there are three or more groups, however, rejection of H_0 only means that at least one population mean is different from the others.

  - It is necessary to follow the test of significance by additional analysis-of-variance procedures to determine which if the population means are different from which other means.

  - What are these follow-up procedures?

Set-Up

- The "groups" compared through analysis of variance may be created in a number of different ways.

  - In many studies the data are already classified into groups, such as states, countries, gender groups, medical diagnostic categories, or religious affiliations.

  - In other instances, the statistician may define the groups based on one or more measured characteristics.
    For example, socioeconomic classifications may be created by combining individuals' educational levels, incomes, and employment statuses according to some set of rules. The result might be, say, three or four or five separate socioeconomic categories.

  - A medical research might classify patients as "high risk" or "low risk" for some disease based on a number of indicators, for example, age, family history, and health behaviors.

– In conducting surveys, groups are frequently formed from the respondents' answers, for example, all people who support a particular political issue, those who are opposed, and those who have no opinion.

– In an experiment, the groups are defined by the various experimental conditions or "treatments," such as instructional methods, psychological or social interventions, or dosages or different forms of a drug.
In these situations, individuals are assigned to the conditions by the researcher. When possible, the assigned should be done randomly so that the only difference between the groups is the explained by differences in the treatments received and not by other factors that are irrelevant to the study.

EXAMPLE 1

A psychologist is interested in comparing the effects of three different informational "sets" in children's ability to memorize words.

- Eighteen 7-year-old children were randomly assigned to one of three groups.

  - In condition 1, the children were shown a list of 12 words and were asked to study them in preparation for a recall test.

  - In condition 2, the children were told that the words comprised three global categories, *flowers*, *animals*, and *foods*, and were asked to study the list.

  - In condition 3, the children were told that the words comprised 6 more detailed categories.

- After studying the word list for 10 minutes, each youngster was asked to list all of the words that he/she could remember.

- The number of correctly recalled words was the measured outcome variable of the study.

EXAMPLE 2

A study was performed to see if school grades are related to television viewing habits among high-school juniors.

- Respondents were monitored for 15 weekdays during the year and classified into 4 groups according to the average amount of television they viewed on those days ($0 - 0.5$ hour; $0.5 - 1.5$ hours; $1.5 - 3.0$ hours; more than three hours).

- Each student's grade-average (GPA) was recorded for all courses taken during the year.

- Mean GPAs were compared among the four television viewing groups.

EXAMPLE 3

In the study The Academic Mind by Lazarsfeld and Thielens, a total of 2451 social science faculty members from 165 of the larger American colleges and universities were interviewed in order to assess the impact of the McCarthy era on social science faculties.

- At each college, the number of "academic freedom incidents" was counted.

- These were incidents mentioned by more than one respondent as an attack in the academic freedom of the faculty.

- They ranged from small-scale matters, such as a verbal denunciation of a professor by a student group, to large-scale matters, including a Congressional investigation.

- It was of interest to examine whether and how the institutional basis of a school's support and control affected the number of "incidents" occurring there.

- Hence, each college was classified as publicly controlled, privately controlled, or controlled by some other institution. (Teachers' colleges and schools controlled by a religious institution were included in the "other" category.)

- The distributions of numbers of "incidents" in the different types of institutions were studied.

EXAMPLE 4

A large manufacturing firm employs high-school

dropouts, high-school graduates, and individuals who attended college as production-line workers.

- The company management speculated that job proficiency was related to educational attainment.

- If so, only high-school graduates, or perhaps only individuals who had attended college, would be hired in the future.

- To test their idea, the job performance of a sample of employees was rated by their supervisors on an extensive rating scale that yield possible proficiency scores ranging from 0 to 200.

- The mean ratings of the three education groups (high-school dropouts, high-school graduates, college attendees) were compared by the analysis of variance.

A Complete Example with equal sample size

The analysis of variance indicates whether population means are different by comparing the variability among sample means with variability among individual observations within groups.
The following data gives the numbers of words memorized by 18 children given three different information sets as described in Example 1.

	No information	3 categories	6 categories
	2	9	4
	4	10	5
	3	10	6
	4	7	3
	5	8	7
	6	10	5
Sums	24	54	30
Sample means	4	9	5

- The scores for three samples and pooled sample can be graphed by box plot.

- The means of the three samples are 4, 9, and 5, respectively.

- There is clear variability from group to group.

- The variability in the entire pooled sample of 18 is much larger than the variability of these three groups.

- Variability within Groups: The measure of variability among observations within the groups is called the Mean Square Groups and is denoted MS_w.

  - MS_w is the variance of all the scores, but computed separately for each of the three samples and then combined.

  - Like the sample variance, MS_w has a sum of squared deviations in the numerator and the corresponding number of degrees of freedom in the denominator.

  - The numerator is called the Sum of Squares within Groups and the denominator is called the degrees of freedom within groups.

  - The Sum of Squares within Groups (SS_w) is the sum of squared deviations of individual scores from their subgroup means. In this example, $SS_w = 10 + 8 + 10 = 28$.

  - The degrees of freedom within groups (df_w) is the total number of degrees of freedom

of the deviations within the groups.
In this example, $df_w = (6 - 1) + 5 + 5 = 15$.

- Variability between Groups

    - The measure of variability among the group means is called the Mean Square between Groups and is denoted MS_B.

    - MS_B is similar to a variance but is computed from the three subgroup means, 4, 9, and 5.

    - MS_B also has a sum of squared deviations in the numerator and a corresponding number of degrees of freedom in the denominator is called the Sum of Squares between Groups and the denominator is called the degrees of freedom between groups.

    - The sum of Squares between Groups (SS_B) is the sum of squared deviations of the subgroup means from the overall or "pooled" mean.

        * The pooled mean of all 18 observations is 6.

        * The squared deviations of the group

means from the pooled mean are $(4 - 6)^2 + (9 - 6)^2 + (5 - 6)^2 = 14$.

* Before dividing by the degrees of freedom between groups, one additional adjustment is necessary.

  To test the null hypothesis of equal means, SS_B is going to be compared to SS_w by converting each to a mean square and then dividing one mean square by the other.

  SS_B cannot be meaningfully compared to SS_w because SS_B reflects variability among means while SS_W reflects variability among individual measurements.

  Put SS_B in a comparable scale to SS_w, it is multiplied by the number of observations in each subgroup, that is, 6 in this example.

  * $SS_B = 6 \times 14 = 84$.

- The degrees of freedom between groups (df_B) is the number of independent deviations summarized in SS_B.

  In this example, $df_B = 3 - 1 = 2$.

- In this example, $MS_B = 84/2 = 42$.

- The Test Statistic:

  - The test statistic to test $H_0 : \mu_1 = \mu_2 = \mu_3$ is the ratio of MS_B/MS_w, or $42/1.867 = 22.50$.
    How do you connect them with $Var(Y)$ and $Var(E(Y|X))$?

  - This ratio seems to indicate that the variability among groups is much greater than that within groups.
    However, we know that different samples would give us different values of the ratio even of the three population means are equal.

  - The calculated ratio of 22.50 is compared to values from the F-distribution to see if the test statistic is large enough for 3 samples and 18 observations to reject H_0.

- Conclusion: If we are using a 1% significance level, the significance point from the table of the F-distribution with 2 and 15 degrees of freedom is 6.36.
  Thus the calculated ratio of 22.50 is very sig-

nificance and we conclude that there are real differences in the average number of words memorized depending on the amount of organizing information that is provided.

- The follow-up question remains, specifically: Which of the three means are significantly different from which others?

The Algebra of ANOVA

- Suppose the number of groups being compared is k.

- The null hypothesis is

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k,$$

that is, the k population means are equal.

- <u>Notation and Data:</u>

		Group			
		1	2	$\cdots$	k
Population	Mean	μ_1	μ_2	$\cdots$	μ_k
	Variance	σ_1^2	σ_2^2	$\cdots$	σ_k^2
Sample	Observations	x_{11}	x_{21}	$\cdots$	x_{k1}
		x_{12}	x_{22}	$\cdots$	x_{k2}
		$\vdots$	$\vdots$		$\vdots$
	Sample size	n_1	n_2	$\cdots$	n_k
	Mean	$\bar{x}_1$	$\bar{x}_2$	$\cdots$	$\bar{x}_k$
	Variance	s_1^2	s_2^2	$\cdots$	s_k^2

- In R, try help.search("anova") and look at the help files of aov and manov.

- The number of sampled observations in a "typical" group, group g, is denoted by n_g

and the total number of observations is $n = n_1 + n_2 + \cdots + n_k$.

- Each observation has two subscripts, the first indicating the group to which the observation belongs and the second indicating the observation number within that group. Thus, x_{gi} represents the ith observation in group g.

  - The sample mean of the ith subgroup is denoted by $\bar{x}_g = n_g^{-1} \, \Sigma_i \, x_{gi}$.
  - The overall mean is denoted by $\bar{x} = n^{-1} \, \Sigma_g \, \Sigma_i \, x_{gi}$.

- The ANOVA Decomposition: The "decomposition" of a single typical score x_{gi} as

$$x_{gi} - \bar{x} = (\bar{x}_g - \bar{x}) + (x_{gi} - \bar{x}_g).$$

  - This expression shows that the deviation of a score from the overall mean $\bar{x}$ $(Var(Y))$ can be decomposed as the sum of two parts, the deviation of the group mean from the overall mean $(Var(E(Y|X)))$ and the deviation of an observation from its group mean $(Var(Y|X))$.
  - The terms on the right-hand side of the

above expression are the fundamental elements of "between-group" and "within-group" variability, respectively.

 – By squaring and summing the terms in the expression we obtain exactly the sum-of-squares partition essential for the analysis of variance.

- Total Sum of Squares:

$$SS_T = \sum_g \sum_i (x_{gi} - \bar{x})^2.$$

- By the property of least squares method, we have

$$SS_T = SS_B + SS_w.$$

- Between-group Sum of Squares and Mean Square:

$$SS_B = \sum_g n_g (\bar{x}_g - \bar{x})^2$$

and $MS_B = SS_B/df_B$ where $df_B = k - 1$.

- Within-group Sum of Squares and Mean Square:

 – SS_w can be obtained by subtracting SS_B from SS_T. (Recall the geometric interptation of least squares method.)

– df_w is the number of independent deviations of observed values from their subgroup means. There are n_g observations in subgroup g, then the number of independent deviations in that group is $n_g - 1$.

Summing these across the k groups, we have $df_w = n - k$.

– The Mean Square within Groups is $MS_w = SS_w/df_w$.

• The F-ratio, $F = MS_B/MS_w$.

Validity Conditions for the F-test

The analysis of variance rests on a set of assumptions that should be considered before the procedure is applied.

- The most inflexible assumption is that the n observations must be independent.

  - Sampling observations at random is an important step in assuring that this condition is met.

  - Steps should be taken to assure that the response on the measured outcome variable of no respondent is in any way affected by other respondents.
    If observations are humans, they should not have the opportunity to hear, see, or otherwise be influenced by other subjects' answers of behavior.

  - If the assumption of independence is not met, then the analysis of variance tests of significance are not generally valid.

- The sampling distribution of the subgroup means should be nearly normal.

– This condition is usually met in analysis of variance, especially if the subgroup n's are moderate to large, since the Central Limit Theorem assures us that means based on large sample sizes are nearly normally distributed.

– The normality condition will also be met if the distribution of the underlying measured variable is normal, whether or not the sample sizes are large.

– It is always advisable to make histograms of the data to see whether there are any gross irregularities in the distributions, however.

– If the sample sizes are small and the distribution of the measured variable is highly non-normal, then the analysis of variance of ranks should be used in place of the ANOVA methods presented so far.

• The F-test is based on an assumption that the population variances are all equal, that is $\sigma_1^2 = \cdots = \sigma_k^2$.

– This condition is especially important if

the sample sizes $(n_1, n_2, \ldots, n_k)$ are not equal.

– Sample variances should be computed prior to conducting the analysis of variance to see if they are in the same general range as one another.
If they appear to be very different, a formal test of equality of the σ^2's may be conducted.

– If the test indicates that the variances are not homogeneous, several options may be available.
The data may be transformed to a scale on which the variances are more equal.
For example, this might involve analyzing the logarithms of the original observed values, the square roots of the observed values, or some other function of the data.

Which Groups Differ from Which, and by How Much?

The F-test gives information about all the means $\mu_1, \mu_2, \ldots, \mu_k$ simultaneously.

- If the hypothesis of equal means is accepted, the conclusion is that the data do not indicate differences among the population means.

- If the null hypothesis is rejected, the conclusion is that there are some differences; then the researcher may want to know which specific means are significantly different from which other and the direction and the magnitudes of the differences.

What is the methods for comparing two means?

- Construct a confidence interval for the difference of the two means as follows:

$$(\bar{x}_1 - \bar{x}_2 - t_{n-2}(\alpha/2)s\sqrt{n_1^{-1} + n_2^{-1}}, \bar{x}_1 - \bar{x}_2 + t_{n-2}(\alpha/2)s\sqrt{n_1^{-1} +$$

Extend the above idea to the comparison of three or more groups.

- When three or more groups are compared by the analysis of variance, several specific

comparisons may be made after the overall hypothesis of equality is rejected.

- For example, if there are three groups and H_0 is rejected, μ_1 may be compared with μ_2, μ_1 may be compared with μ_3, and μ_2 may be compared with μ_3.

- It is up to the researcher to decide which comparisons to make. The decision rests partially on the design of the research. For example, if group 1 is a control group and groups 2 and 3 are two different experimental conditions, then it would be sensible to compare μ_2 with μ_1 and μ_3 with μ_1, that is, both experimental conditions with the control.

- If students from four different universities are being compared on mean scores on the Law School Admissions Test, every school's mean might be routinely compared to every other school's mean. In the latter case, the number of pairwise comparisons among k means is $k \times (k-1)/2$.

– The procedure for any one of these comparisons is very much like comparing two groups as described before.

– But one additional factor needs to be considered: the probability of making a Type I error when performing several tests of significance from the same data set.

– If, indeed, all the μ's were equal, so that there were no real differences, the probability that any particular one of the pairwise differences would exceed the corresponding t-value is α.
However, the probability of making at least one Type I error out of two or more pairwise comparisons is greater than this. That is, when many differences are tested, the probability that some will appear to be "significant" when the corresponding population means are equal is greater than the nominal significance level α. The more comparisons that are made, the greater the probability of making at least one Type I error.

• How can a researcher protect against too

high a Type I error rate?

– One widely used approach is based on the Bonferroni inequality.

– The Bonferroni inequality states that the probability of making at least one Type I error out of a given set of comparisons is less than or equal to the sum of the α's used for the separate comparisons.

– For example, if we make 2 comparisons among 3 group means and use an α level of 0.05 for each comparison, the probability of making at least one Type I error is no greater than 0.10.

– This overall α level for the pair or "family" of tests is called the familywise (or experimentwise) Type I error rate.

– The Bonferroni inequality can be put to use to keep the familywise error rate acceptably small.

– Suppose that we wish to make m specific pairwise comparisons. We can then decide on a reasonable familywise error rate (α) and divide this value by m to

obtain a significance level to be used for each comparison separately; call this result α^*.

- If α^* is used for each of m comparisons, the probability of making at least one Type I error out of the set is not greater than $m \times \alpha^* = \alpha$.

Analysis of Variance of Ranks

- In ANOVA, its basic idea is to compare the distributions of a variable in several populations by focusing on the means of the samples.

  - The ANOVA is based on the conditions that the sample means have a normal distribution and that the populations being compared have the same variance σ^2.

  - When these conditions are not met, or when the raw data are ordinal, can we still use ANOVA technique to compare several populations?

  - Consider the test procedure developed by Kruskal and Wallis (1952).

- Another approach to compare the distributions of a variable in several populations:

  - The *locations* of distribution can also be compared through an analysis of ranks.

  - This approach can be applied even when the data are ordinal, and does not require the assumption of normality or equal variances.

– The null hypothesis is that the locations of the populations are the same, and the alternative hypothesis is that they are not.

• Kruskal and Wallis (1952) considered the daily outputs of three bottle-cap machines.

– Use the following 12 values to test whether the three machines produce equal numbers of bottle caps.

Machine A	$340, 345, 330, 342, 338$
Machine B	$339, 333, 344$
Machine C	$347, 343, 349, 355$

– Instead of using the above data, we use the following ranks instead.

Ranks

Machine A	$5, 9, 1, 6, 3$
Machine B	$4, 2, 8$
Machine C	$10, 7, 11, 12$

– The test applies ANOVA formulas to these ranks. We have

$$SS_T = \frac{n(n+1)(2n+1)}{6} - n\left(\frac{n+1}{2}\right)^2$$

$$= \frac{n(n+1)(n-1)}{12}$$

and $MS_T = n(n+1)/12$.

- The F-ratio in ANOVA is $F = MS_B/MS_w$.
- Unlike ANOVA, we consider $H = MS_B/MS_T$.
- Observe that

$$SS_B = \sum_g \frac{(\Sigma_i\, r_{gi})^2}{n_g} - n\left(\frac{n+1}{2}\right)^2.$$

- In this example,

$$H = \frac{12}{12 \times 13}\left(\frac{24^2}{5} + \frac{14^2}{3} + \frac{40^2}{4}\right) - 3 \times (12+1) = 5.66.$$

• The null hypothesis of no difference between the locations of the three populations is equivalent to tandom sampling, that is, that the ranks have been allocated at random to the k groups.

- When this hypothesis is true, the sampling distribution of H is approximately χ^2 with $k-1$ degrees of freedom if the sample sizes are large.
- Why? What is F-distribution?

Introduction to Logistic Regression

There are many situations in which we want to forecast events which cannot be modeled as a continuous variable. Examples include

- Whether a consumer will purchase a product from a menu of products, click on a web button, respond to a direct mail offer.

- Whether a firm will decide to repurchase stock, change accounting procedures, or write-off an asset.

- In these cases, the Y variable is not only discrete but takes only two values (1 "Event" or 0 "No Event").
  This is the simplest type of random variable, a Bernoulli random variable. The classic example of this variable is the outcome of a coin toss.

- We want to make predictions about this discrete Y on the basis of some other explanatory variables.

  – For example, we want to study how product attributes and price influence consumer choice from an array of products

or how industry and firm financial conditions affect the decision to write-off assets.

  – To make the conditional predictions, we need to incorporate the X variables into a prediction about Y.
  – Unfortunately, the standard linear regression of Y on the X variables is inappropriate.
  Our forecasts should be probabilities!

• If we think back to the introduction of regression, we viewed the regression model as a model for the conditional mean of Y given the X variables.

• For a Bernoulli random variable, the conditional mean will be a probability.

$$E(Y|X) = P(Y = 1|X)$$

• The logit model is one very convenient and useful way of forming probabilties from x variables.

$$P(Y = 1|x) = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}$$

– Think of $V = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$ as a "score."

– As V gets large, the probability $Y = 1$ should increase (but never exceed 1).

– For example, if a firm is considering write-off an asset, then V might be the "desirability" of the write-off.
If V is negative, the write-off is not desired.

– The logit model is a particular form for how V gets mapped into a probability.
$$P(Y = 1) = \frac{\exp(V)}{1 + \exp(V)}.$$
What does this probability curve or locus look like?

– Use R to plot it by yourself.
There are several keys aspects to notice about this curve:
i) As V increases to levels very much above or below 0 the probability of purchase goes very close to to 1 or 0.
ii) The sensitivity to changes in V varies depending on the level of utility and associated probability. Around prob $= 0.5$,

the slope is at its maximum of 0.25. but at low or high probs the slope declines to very small numbers. This is because of the "Logistic" or S-shaped curve.

How do we estimate unknown parameters in such a model?

Use the method called Maximum Likelihood Estimation ("MLE!L).

- Form a "likelihood function!L for each observation.

$$\prod_i \left(\frac{\exp(V_i)}{1 + \exp(V_i)} \right)^{Y_i} \left(\frac{1}{1 + \exp(V_i)} \right)^{1-Y_i}$$

where $V_i = b_0 + b_1 X_{i1} + \cdots + b_k X_{ik}$.

- Consider the logarithm of the likelihoods for all the observations.

- Find the values of $b_0, b_1, b_2, \ldots$ that make this sum (log likelihood function) as large as possible.

- In R, try help(glm) and help(family) to find the MLE estimate.

  - Logistic regression is a special case of generalized linear model.

– We need to specify the error distribution and the link function.

– For logistic regression, the error distribution is binomial distribution and logit is being used to link the linear predictor and the mean of the error distribution.

– In glm, the "binomial" family admits the links "logit", "probit", "log, and "cloglog" (complementary log-log).

An Example of Logistic Regression

In "Causes and Effects of Discretionary Asset Write-Offs" by Francis, Hanna and Vincent (JAR 1996),

- The authors investigate the decision of firms to write-off assets by gathering a large sample of firms that made and did not make write-offs in the 1989-92 period.

- The authors propose a set of independent variables which might be expected to influence the write-off decision.
  The set of explanatory variables are as follows:

  1. ryear1 - cumulative abnormal return over the preceding year
  2. ryear5 - cumulative abnormal return
  3. mtbdif - firm's industry-adjusted Book to Market ratio
  4. mtbchg - change in firms B-t-M ratio
  5. indmtb - change in industry's B-t-M ratio
  6. roa - change in roa
  7. indroa - change in industry roa

8. history - number of yrs in which firm reported special negative items

9. indhis - mean value of history for all other firms in industry

10. dmgmt - 1 if management changed in previsous year

11. poor - Unexpected earnings if $UE < 0$, 0 otherwise

12. good - UE- $amt of write-off, if > 0, 0 otherwise

13. lnsale - log of sales in yr preceding write-off

14. unless otherwise stated, all variables are averages of preceding 5 years

- They are interested in determining whether write-offs are the result of managers attempts to manipulate accounting performance or simply the result of declines in the value of assets.

- The article relates the predicted probability of write-off to changes in stock price, but we will focus on the problem of predicting write-offs given on the basis of the financial

performance of the firm and the industry in which the firm operates.

Two-Way ANOVA; General Model

Two-way ANOVA has two factors of interest that occur in all combinations.

- Sometimes there is only one observation on each combination.

- Sometimes there are more (called "replications").

- The general model with replications has observations labeled as Y_{ijk}.
  The index i $(i = 1, \ldots, I)$ labels the value of the first factor, j $(j = 1, \ldots, J)$ labels the value of the second factor, and k labels the replicated observations on the combination ij.

- When we always have the same number of replications for each ij combination, so that $k = 1, \ldots, K$. Such a model is called "balanced".

- There are two types of models - the "additive model" and the "additive model with interactions".
  The second model needs $K \geq 2$.

Additive Model

- The additive model decomposes the population means $\mu_{ij} = E(Y_{ij})$ as a sum of effects of the corresponding i and j factors.

$$\mu_{ij} = \mu + \alpha_i + \beta_j$$

  where $\Sigma_i\, \alpha_i = 0$ and $\Sigma_j\, \beta_j = 0$..

- Without proper constraint on α_i and β_j, those unknown parameters are not identifiable.

- For the balanced model, μ, αs and βs can be estimated in the natural way (like in the one-way model) as $\hat{\mu} = \bar{Y}_{...}$, $\hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{...}$, $\hat{\beta}_j = \bar{Y}_{.j.} - \bar{Y}_{...}$. These formulas correspond to the natural estimates of the β_j. Thus

$$\hat{\mu}_{ij} = \bar{Y}_{i..} + \bar{Y}_{.j.} - \bar{Y}_{...}.$$

  (For non-balanced models the corresponding formulas are more complicated, which is why we treat only balanced models.)

- There is an overall null hypothesis to be tested:

  $H_0 : \alpha_1 = \cdots = \alpha_I = 0$ and $\beta_1 = \cdots = \beta_J = 0$
  versus $H_a : H_0$ is not true.

It includes two sub-hypotheses of interest:

$H_{0A} : \alpha_1 = \cdots = \alpha_I = 0$ versus $H_{aA} : H_{0A}$ is not true.
$H_{0B} : \beta_1 = \cdots = \beta_J = 0$ versus $H_{0B} : H_{0B}$ is not true

The Data:
Hourly arrivals of telephone calls to a
telephone call center

- The data in this example involve telephone calls to a relatively small Israeli Bank telephone call center in 1999.

- The caller desires to speak to a telephone service agent.

- The call center managers want to be able to predict the number of calls that will arrive in any given hour.

  - The working day at this center runs from 7am to 11:59pm.
  - Look at data for all the full work-weeks in November and December 1999.
  - Divide each day up into hourly intervals; from $7 - 8$, $8 - 9$, etc.
    Label these intervals as $i = 1, 2, \ldots, 17$.

Thus interval $i = 2$ corresponds to 8am V 9am and $I = 17$.

- There are 5 regular working days each week (Sunday through Thursday in Israel.).
  Label these as $j = 1, \ldots, 5$.
  Thus $j = 2$ corresponds to a Monday and $J = 5$.

- Let N_{ijk} denote the number of calls arriving during hourly interval i, day-of-week j, and week k. Note that $K = 8$.

- Use ANOVA to summarize this data set.

- How could we model N_{ijk}?

  - It is reasonable to conjecture that these arrival times are well modeled by an inhomogeneous Poisson process.

  - The arrival rate for this process should depend only on time of day, and perhaps other calendar related covariates such as month or day of the week.

  - Theory suggests that they may have a Poisson distribution with mean λ_{ij}.

– If the arrival process for a given call category is as above then the number of arrivals each day within any given interval of time should be independent Poisson variables with a parameter that depends only on the given time interval.
  If other covariates are involved (such as day of the week) then the Poisson parameter may also depend on these.

– Poisson distribution an be approximated by a normal distribution when λ_{ij} is large. If so they would not be homoscedastic (since their variance would equal their mean).

– Anscombe's (1948, *Biometrika*) variance stabilizing transformation suggests that the variables $\sqrt{N_{ijk}}$ might be nearly homoscedastic with variance $1/4$.

– Consider $Y_{ijk} = \sqrt{N_{ijk} + 1/4}$.

• Goal of the Manager:

– The manager of the call center would like to be able to predict the number of customers in any particular hour that will

call the center desiring to speak to an agent.

- Plan how many agents are needed at that time of day.

- (Other considerations also enter into this decision, such as the length of time that it takes an agent to serve a customer.)

- Next page gives a plot that tells the manager how to predict N for each hour on any day 2 (Monday).

- This plot also tells the manager what the 95% prediction intervals are for that prediction.

- Note that the prediction limits are pretty wide.
  THATS UNFORTUNATE,
  but it cant be helped if we only know about time-of-day and day-of-week.

- The 95% confidence intervals are also shown, but probably aren't as important to the manager.

ANOVA Table

The tests of these hypotheses are summarized in an ANOVA table.

Summary of Fit

Rsquare	0.8127
Root Mean Square Error	0.7645
Mean of Response	7.889
Observations	680

Analysis of Variance

Source	DF	SS	MS	F Ratio
Model	$(I-1)+(J-1)=20$	1671.63	83.581	143.00
Error	$n-I-J+1=659$	385.17	0.585	Prob $> F$
C. Total	$n-1=679$	2056.80		$< .0001$

Effect Tests

Source	Npar.	DF	SS	F Ratio	Prob $> F$
Hour	16	$I-1=16$	1658.40	177.34	$< .0001$
day of week	4	$J-1=4$	13.23	5.66	0.0002

- Sum of Squares for "C Total" is $SST = \Sigma_i \Sigma_j \Sigma_k (Y_{ijk} - \bar{Y}_{...})^2$.

- Sum of Squares for "Error" is $SSE = \Sigma_i \Sigma_j \Sigma_k (Y_{ijk} - \hat{\mu}_{ij})^2$.

- In the model, we consider two factors: hour and day of the week.

$$SSModel = SSA + SSB.$$

Residual by Predicted Plot
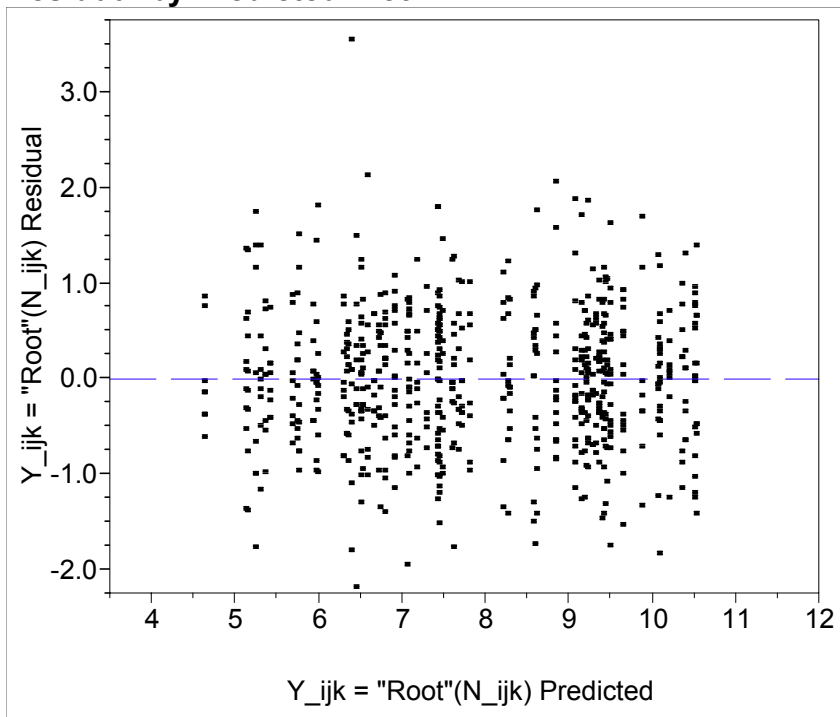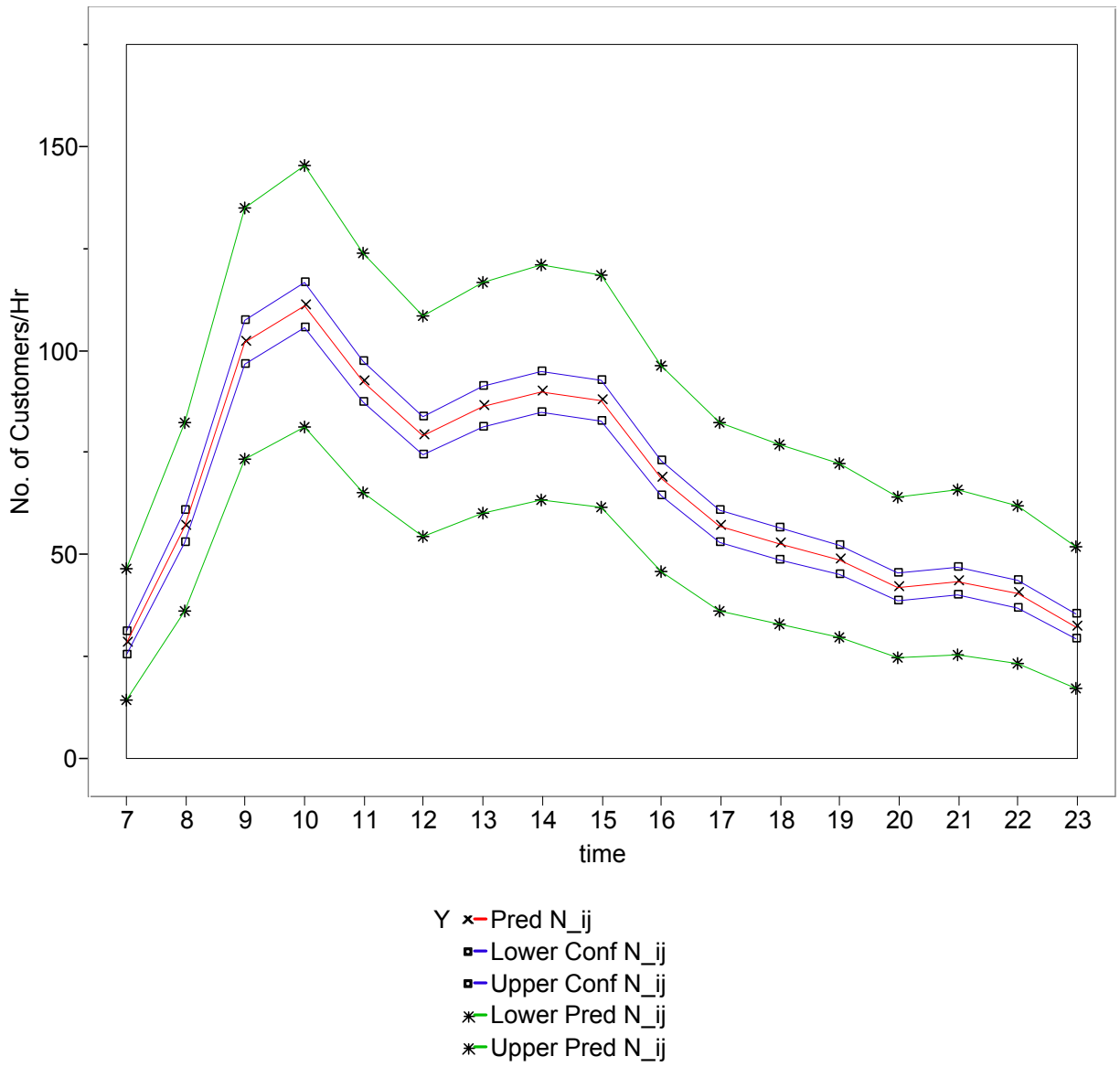
Plot of Predicted Values of N for Mondays, as a function of Hour–of–Day

- $SSModel = SST - SSE$.
- $SS(Hour)$ is $SSA = \Sigma_j \Sigma_k \Sigma_i (\bar{Y}_{i..} - \bar{Y}_{...})^2$.
- SS(day of week) is $SSB = \Sigma_i \Sigma_k \Sigma_j (\bar{Y}_{.j.} - \bar{Y}_{...})^2$.

- F-ratios:

  - The first F-ratio $(= 143 = MSM/MSE)$ tests H_0 and here has 20 and 659 df.

  - The second F-ratio $(= 177.34 = MSA/MSE)$ tests H_{0A} and here has 16 and 659 df.

  - The third F-ratio $(= 5.66 = MSB/MSE)$ tests H_{0B} and here has 4 and 659 df.

- The analysis tells us that knowing the day of the week makes a statistically significant difference, but not a very important one; one could do almost as well just knowing the time of day. This means that there is not much point from the service point of view in the "manager" making different work schedules for different weekdays.

- CONCLUSION: All three null hypotheses are rejected, but the differences among the

factors B are much less striking than those among the factors A.

Model Checking: Checking the residuals

- Each RESIDUAL is the value of the OB-
  SERVATION V ITS PREDICTION. Sym-
  bolically, $r_{ijk} = Y_{ijk} - \hat{\mu}_{ij}$.

- When the basic ANOVA analysis is com-
  plete the residuals should then be examined
  for adherence to the basic assumptions V

  - homoscedasticity
  - normality

- Standard graphical procedures: residual plot
  and quantile plot

- Residual plot:

  - Look at a plot of the residuals against
    the values of Y_{ijk} or against the predicted
    values $\hat{\mu}_{ij}$.
    This provides a check for homoscedastic-
    ity. Refer to next page.
  - What you are looking for is to see that
    the vertical spreads of the data are ap-
    proximately the same at each value of
    $\hat{\mu}_{ij}$. (You need to allow for random vari-
    ation when evaluating such a plot V Our

basic assumption is that such an equality
of spreads holds in the population.)

– CONCLUSION: So far as we can tell it
appears that the populations are homoscedastic (apart from that one outlier).

• Normal probability plot

– Once it has been decided that the data
are acceptably homoscedastic it becomes
of interest to check whether the within ij
group residuals are also acceptably near
normality.

– R can give you the values of the residuals
and then form a Normal Quantile Plot to
check normality.

– This residual plot shows startlingly good
agreement with normality V except for
the one annoying outlier.

– Usually we're satisfied with considerably
less convincing agreement to normality.

– What we want to avoid are heavily skewed
residual distributions.
These would suggest that the analysis
is invalid, and that some corrective ac-

tion (such as a transformation of the Y-variables) needs to be taken before re-analyzing the data.

Model Checking: Compare it to a model with interactions

- The ij population group means are modeled as
$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij},$$
where $\Sigma_i\, \alpha_i = 0$, $\Sigma_j\, \beta_j = 0$, $\Sigma_i\, \gamma_{ij} = 0$, and $\Sigma_j\, \gamma_{ij} = 0$.

- The γ_{ij} are the new part of this model. They describe the difference between the μ_{ij} here and their value in the additive model – $\mu + \alpha_i + \beta_j$.

- This model imposes no restrictions on the μ_{ij}; they can take any value. Correspondingly, it turns out that
$$\hat{\mu}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{ij} = \bar{Y}_{ij\cdot}.$$

Data Analysis

The tests of these hypotheses are summarized in an ANOVA table.

Summary of Fit
Root Mean Square Error 0.7617
Mean of Response   7.889
Observations   680

Analysis of Variance

Source	DF	SS	MS	F Ratio
Model	84	1711.59	20.3760	35.1193
Error	595	345.21	0.5802	Prob $> F$
C. Total	679	2056.80		$< .0001$

Effect Tests

Source	Npar.	DF	SS	F Ratio	Prob $> F$
Hour	16	16	1658.40	177.34	$< .0001$
day of week	4	4	13.23	5.66	0.0002
Hour $\times$ day of week	64	64	39.96	1.08	0.3272

- The "new" part of this table is the line for Hour $\times$ day of week.

- This corresponds to a test of H_{0AB}: all $\gamma_{ij} = 0$ versus H_{aAB}: not H_{0AB}.

- It has $df = (I-1)(J-1) = 64$. (And, note that the "Model" now has $df = IJ - 1 = 84$, so that the Error $df = n - IJ = 595$, which is also not the same as in the additive model.)

- CONCLUSION: We DO NOT REJECT H_{0AB}. We conclude that there is no statistically significant evidence of interaction effects. So far as we can determine the additive model fits this situation as well as the model with interactions.

- Profile Plots:

  - Graphical method for the detection of interactions

  - Plots of $\hat{\mu}_{ij}$ as a function of i for each j are a good way to see the interaction estimates. (Or plots for each j as a function of i.)
    This produces a Profile Plot.

  - If there were no interactions in the data then we would have

  $$\hat{\mu}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j$$

  so that the functions displayed on the profile plot would be exactly parallel.

  - The degree to which they are not parallel indicates how much interaction there is in the data.

Related R-commands

- interaction.plot

  - Two-way Interaction Plot

  - Plots the mean (or other summary) of the response for two-way combinations of factors, thereby illustrating possible interactions.

- lm

  - Fitting Linear Models

  - "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).

  - Models for *lm* are specified symbolically. A typical model has the form "response $\sim$ 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* to-
gether 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.*

- $g = lm(calltransf \sim hour * day, data)$
$anova(g)$

– anova: It gives Anova Tables

– Compute analysis of variance (or deviance)
tables for one or more fitted model ob-
jects.

- Diagnostics checking on the residual

– qqnorm(g$res)

– plot(g$fitted,g$res,xlab="Fitted",ylab="Residuals",
main="Square root response")

Blocking Designs and ANOVA

Suppose we want to compare 4 treatments and have 20 patients available.

- In completely randomized designs (CRD), the treatments are assigned to the experimental units at random.
  The one-way ANOVA can be used to analyze the resulting data to check on treatment effects.

  - This design is appropriate when the units are homogenous.
  - We may suspect that the units are heterogenous, but we can't describe the form it takes - for example, we may know a group of patients are not identical but we may have no further information about them.
  - In this case, it is still appropriate to use a CRD.
  - The randomization will tend to spread the heterogeneity around to reduce bias.
  - Under the null hypothesis, there is no link between a factor and the response.

In other words, the responses have been assigned to the units in a way that is unlinked to the factor. This corresponds to the randomization used in assigning the levels of the factor to the units. This is why the randomization is crucial because it allows us to make this argument.

If the difference in the response between levels of the factor seems too unlikely to have occurred by chance, we can reject the null hypothesis.

- When the experimental units are heterogenous in a known way and can be arranged into blocks where the intrablock variation is ideally small but the interblock variation is large, a block design can be more efficient than a CRD.

  - We might be able divide the patients in 5 blocks of 4 patients each where the patients in each block have some relevant similarity.

  - We would then randomly assign the treatments within each block.

– Suppose we want to test 3 crop varieties on 5 fields.
Divide each field into 3 strips and randomly assign the crop variety.

– We prefer to have block size equal to the number of treatments.
If this is not done or possible, an incomplete block design must be used.

– We just motivate *randomized block design.*
For the example we just described, we have one factor (or treatment) at 3 levels and one blocking variable at 5 levels.
We can use two-way ANOVA to check for interaction and check for a treatment effect.
We can also check the block effect but this is only useful for future reference.

• Notice that under the randomized block design the randomization used in assigning the treatments to the units is restricted relative to the full randomization used in the CRD.

– Blocking is a feature of the experimen-

tal units and restricts the randomized assignment of the treatments.

– This means that we cannot regain the degrees of freedom devoted to blocking even if the blocking effect turns out not to be significant.

– Only use blocking where there is some heterogeneity in the experimental units. The decision to block is a matter of judgment prior to the experiment.
There is no guarantee that it will increase precision.

– How do we compare the efficiency of two designs?

• Latin Squares: These are useful when there are two blocking variables.

– Suppose, in a field used for agricultural experiments, the level of moisture may vary across the field in one direction and the fertility in another.

– In an industrial experiment, suppose we wish to compare 4 production methods (the treatment) - A, B, C, and D.

We have available 4 machines 1, 2, 3, and 4, and 4 operators, I, II, III, IV.

- A Latin square design is

	1	2	3	4
I	A	B	C	D
II	B	D	A	C
III	C	A	D	B
IV	D	C	B	A

- Each treatment is assigned to each block once and only once.

  The design and assignment of treatments and blocks should be random.

- Three-way ANOVA